



IEMTRONICS

International Conference

Toronto, Canada

2022 CONFERENCE PROCEEDINGS

Date: 1st - 4th June, 2022

Editors:

Satyajit Chakrabarti, Rajashree Paul, Bob Gill,
Malay Gangopadhyay, Sanghamitra Poddar



About the Conference

IEMTRONICS 2022

Continuing with the outstanding success of IEEE IEMCON, IEEE CCWC, IEEE UEMCON, IEMANTENNA we are proud to present IEMTRONICS 2022 (International IOT, Electronics and Mechatronics Conference) which will be held during **1st- 4th June, 2022 at Toronto, Canada, in online mode**. Keeping in mind the pandemic situation prevalent globally due to Covid 19 and following the legacy of organizing highly successful conferences, we have planned for the online conference. The conference aims to bring together scholars from different backgrounds to emphasize dissemination of ongoing research broadly in the fields of IOT, Electronics and Mechatronics. Research papers are invited describing original works in above mentioned fields and related technologies. The conference will include a peer-reviewed program of technical sessions, special sessions, tutorials and demonstration sessions.

All accepted papers which will be presented during the parallel sessions of the Conference will be submitted for publication in IEEE Xplore Digital Library (Scopus, DBLP, Ei Compendex, Web of Science and Google Scholar).

This conference will also promote an intense dialogue between academia and industry to bridge the gap between academic research, industry initiatives, and governmental policies. This is fostered through panel discussions, keynotes, invited talks and industry exhibits where academia is exposed to state-of-practice and results from trials and interoperability experiments. The industry in turn benefits by exposure to leading-edge research in networking as well as the opportunity to communicate with academic researchers regarding practical problems that require further research.

Our Reviewers

IEMTRONICS 2022 followed a rigorous triple-blind review process in order to identify suitable papers for both presentation and publication. This process helped the organizers to shortlist good quality papers from diverse regional areas and across various domains. A detailed review process was possible due to the excellent and enthusiastic support extended by the strong technical review team of IEMTRONICS 2022. For every stage of submission, IEMTRONICS had a specific template review procedure to analyze the submissions and provide suitable comments for the authors to incorporate. The review team which formed the technical backbone for the selection of submissions for the edited book and the conference presentation was supervised by:

NAME	AFFILIATIONS
Amany Abood	Al-Esraa University College
Qasem Abu Al-Haija	Princess Sumaya University for Technology (PSUT)
Naheem Adesina	Louisiana State University
Md Imtiaz Ahmed	Prime University
Baker Al Smadi	Grambling State University
Md Ali	Rider University
Ali Abdullah S. AlQahtani	North Carolina A&T State University
Nesreen Alsbou	University of Central Oklahoma
Ahmed Ammari	INSAT - Carthage University Tunisia
Mohammad Anees	Xilinx
Vaibhav Anu	Montclair State University
Asif Mohammed Arfi	University of Yamanashi
Haissam Badih	Lawrence Technological University
Anindya Bal	BRAC University
Kuhaneswaran Banujan	Sabaragamuwa University of Sri Lanka
Doina Bein	California State University, Fullerton
Aleksandr Below	National Research University Higher School of Economics
La Verne Certeza	University of Santo Tomas
Pratik Chattopadhyay	Indian Institute of Technology (BHU), Varanasi
Ritu Chaturvedi	University of Guelph
Sangay Chedup	Jigme Namgyel Engineering College
Yuanzhu Chen	Queen's University
Jingyuan Cheng	University of Science and Technology of China
Syantika Chowdhury	Jadavpur University

Ronnie Concepcion II	De La Salle University
José Cornejo	Universidad Tecnológica del Perú
Monica Costa	Polytechnic Institute of Castelo Branco
Omar Darwish	Eastern Michigan University
Arighna Deb	KIIT University
Mohan Dehury	Koneru Lakshmaiah Education Foundation
Sukomal Dey	Indian Institute of Technology, Delhi
S Dhivya	VIT University
Ke-Lin Du	Concordia University
Pallav Dutta	Aliah University
Mahmoud Elkhodr	Central Queensland University
Amin Fadlalla	Mashreq University
Zainab Faisal	Al-Esraa University College
Vilas Gaidhane	Birla Institute of Technology and Science Pilani, Dubai Campus, UAE
Sebastian Garcia	none
Soham Ghosh	University of Kansas
Rajib Kumar Halder	Jagannath University
Maryam Heidari	George Mason University
Deyby Huamanchahua	Universidad de Ingeniería y Tecnología - UTEC
Maysam Hussein	Al-Esraa University College
Qaiser Ijaz	The Islamia University of Bahawalpur
Mohammad Tariq Iqbal	Memorial University of Newfoundland
Ashwini Jadhav	University of the Witwatersrand
Asif Uddin Khan	Silicon Institute of Technology, Bhubaneswar
Shahriar Khan	Independent University
Haruo Kobayashi	Gunma University
Moises Levy	Florida Atlantic University
Pravir Malik	Deep Order Technologies
Olusiji Medaiyese	University of Louisville
Morteza Modarresi Asem	Tehran Medical Sciences University
Nabilt Moggiano	Universidad Continental
Bhabendu kumar Mohanta	GITAM Deemed to Be University
Erkin Navruzov	National University of Uzbekistan
Moses Onibonoje	Afe Babalola University, Ado Ekiti
Madhumita Pal	Institute of Engineering & Management, Kolkata
Rajvardhan Patil	Grand Valley State University
Loreen Powell	Bloomsburg University of Pennsylvania
Kumar Rahul	XILINX
Kuvonchbek Rakhimberdiev	National University of Uzbekistan Named After Mirzo Ulugbek
Biplob Ray	Central Queensland University
K Himaja Reddy	KLEF Deemed to be University
Ashiq Sakib	Florida Polytechnic University
Sowmya Sanagavarapu	Anna University
Daniel Semwayo	University of Witwatersrand
Abhijit Sen	KPU
Ahmed Shafkat	Fareast International University
Lugen Sheet	UNIVERSITY OF MOSUL
Sodessa Shonkora	Arba Minch University

Mayer Silva	none
Rohit Singh	University of Colorado Denver
Kanika Sood	California State University, Fullerton
M. Srilatha	Vardhaman College of Engineering
Elvis Supo	Universidad Nacional de San Agustín de Arequipa
Sourabh Swarnkar	Xilinx
Tri Tran	Gunma University
Yuan Xing	University of Wisconsin-Stout
Wael Yafooz	Taibah University
Lasith Yasakethu	Sri Lanka Technological Campus
Hasan Yasar	Carnegie Mellon University

Sponsors

- Society for Makers, Artists, Researchers and Technologists, Canada
- IEEE VANCOUVER SECTION
- IEEE TORONTO SECTION
- Institute of Engineering & Management, Kolkata
- University of Engineering & Management, Kolkata
- University of Engineering & Management, Jaipur

COPYRIGHT

2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at pubs-permissions@ieee.org. All rights reserved. Copyright ©2021 by IEEE.

CATALOG NUMBERS: CFP22Y72-ART

ISBN: 978-1-6654-8684-2

ORGANIZING COMMITTEE

General Chair:

Rajashree Paul

University of Engineering & Management, Kolkata, India

Technical Co-Chair:

Bob Gill

British Columbia Institute of Technology, Burnaby, Canada

Malay Gangopadhyay

Institute of Engineering & Management, Kolkata, India

Finance Chair:

Sanghamitra Poddar

Institute of Engineering & Management, Kolkata, India

Publicity Chair:

Fatima Hussain

**Professor, Ryerson university, Canada, Editor IEEE
Newsletter, IEEE Toronto section**

ADVISORY COMMITTEE

Name	University
Dr. Chuck Easttom	University of Dallas, USA & Georgetown University, USA
Dr. Phillip Bradford	University of Connecticut-Stamford, USA
Dr. Ronald F. DeMara	University of Central Florida, USA
Dr. Fatima Hussain	Professor, Ryerson university, Canada, Editor IEEE Newsletter, IEEE Toronto section
Dr. Ashutosh Datta	Johns Hopkins University, USA
Dr. Yang Hao	Queen Mary University, London
Dr. Vien Van	University of Alberta, Canada
Dr. Omar Ramahi	University of Waterloo, Canada
Dr. Yahia Antar	Royal Military College, Canada
Dr. Zhizhang (David) Chen	Dalhousie University, Canada
Dr. Detlef Streitferdt	Technische Universitat Ilmenau, Germany
Prof. Shahab Tayeb	California State University, Fresno.

TECHNICAL COMMITTEE

Name	University
Dr. Nabeeh Kandalaft	Grand Valley State University, USA
Dr. Alex "Sandy" Antunes	Capitol Technology University, USA
Dr. Izzat Alsmadi	Texas A&M, San Antonio, USA
Dr. Lo'ai Tawalbeh	Texas A&M University-San Antonio, USA
Dr. Pratik Chattopadhyay	Indian Institute of Technology (BHU), Varanasi
Dr. Doina Bein	California State University, Fullerton, USA
Dr. Hasan Yasar	Carnegie Mellon University, USA
Dr. Moises Levy	West Texas A&M University, USA
Dr. Christian Trefftz	Grand Valley State University, USA
Dr. Mrinal Sen	Indian Institute of Technology(ISM), Dhanbad
Dr. Petros Spachos	University of Guelph, Canada
Dr. Kanika Sood	California State University, Fullerton
Dr. Ke-Lin Du	Concordia University, Canada
Dr. Wenlin Han	California State University, Fullerton
Dr. Ashiq Adnan Sakib	Florida polytechnic University, USA
Dr. Morteza Modarresi Asem	Islamic Azad University, Iran
Dr. Md. Liakat Ali	Rider University, USA
Dr. Tarek El Salti	Sheridan College, Canada
Dr. Sukomal Dey	Indian Institute of Technology, Palakkad
Dr. Maysam Chamanzar	Carnegie Mellon University, USA
Dr. Kean Boon Lee	Sheffield University, UK

Track Topics:

IoT & Data Science:

- IoT and blockchain
- IoT and big data
- Next-generation infrastructure for IoT
- Cloud computing and IoT
- Edge computing and IoT
- IoT platforms, tools, and applications
- IoT systems development methodologies
- IoT applications

Electronics:

- Antenna and wireless communication
- Microwave Engineering
- Photonics
- Nano science & Quantum Technology
- VLSI and Microelectronic Circuit Embedded Systems
- System on Chip (SoC) Design
- FPGA (Field Programmable Gate Array) Design and Applications
- Electronic Instrumentations
- Sensors & Systems
- NEMS & MEMS
- Integrated circuits & power electronics
- Electronic Power Converters and Inverters
- Electric Vehicle Technologies
- Control Theory, Optimization and Applications
- Robotics and Autonomous Systems
- Intelligent, Optimal, Robust, Adaptive Control
- Linear and Nonlinear Control Systems
- Complex Adaptive Systems
- Industrial Automation and Control Systems Technology
- Modern Electronic Devices
- Biomedical devices & Imaging
- Energy Harvesting & Conversions
- Energy Efficient Hardware systems

Mechatronics:

- Sensing and Control Systems
- Mechatronics Systems
- Mechanical Systems
- Artificial Intelligence
- Applications of Robotics

Information Technology:

- Business Intelligence and Applications
- Computer Network
- Evolutionary Computation and Algorithms
- Intelligent Information Processing
- Information System Integration and Decision Support

- Image Processing and Multimedia Technology
- Signal Detection and Processing
- Technique and Application of Database
- Software Engineering
- Mobile Computing
- Distributed Systems
- Artificial Intelligence
- Visualization and Computer Graphic
- Natural Language Processing
- Deep Learning
- Machine Learning
- Internet of Things, Data Mining
- Data Science
- Cloud Computing in E-Commerce Scenarios
- E-Business Systems Integration and Standardization, E-government
- Electronic Business Model and Method
- E-Commerce Risk Management
- Recommender system
- Semantic Web Service Architecture for E-Commerce
- Service Oriented E-Commerce and Business Processes
- Data Analytics and Big Data
- Software defined networking
- Secured distributed systems

Mobile Communication:

- Ad hoc networks
- Body and personal area networks
- Cloud and virtual networks
- Cognitive radio networks
- Cyber security
- Cooperative communications
- Delay tolerant networks
- Future wireless Internet
- Local dependent networks
- Location management
- Mobile and wireless IP, Mobile computing
- Multi-hop networks
- Network architectures
- Network Security, Information Security, Encoding Technology
- Routing, QoS and scheduling
- Satellite communications
- Self-organising networks
- Telecommunication Systems
- Vehicular networks
- Wireless multicasting, Wireless sensor networks

Chief Guest of IEMTRONICS 2022

Nobel Laureate



Prof. Takaaki Kajita

Honourable Nobel Laureate, Distinguished University Professor, Institute for Cosmic Ray Research, The University of Tokyo, Japan

Bio: Kajita Takaaki, (born 1959, Higashimatsuyama, Japan), Japanese physicist who was awarded the 2015 Nobel Prize in Physics for discovering the oscillations of neutrinos from one flavour to another, which proved that those subatomic particles have mass. He shared the prize with Canadian physicist Arthur B. McDonald.

Kajita received a bachelor's degree from Saitama University in 1981 and a doctorate from the University of Tokyo (UT) in 1986. That year he became a research associate at the International Center for Elementary Particle Physics at the UT, where he worked on the Kamiokande-II neutrino experiment, a tank containing 3,000 tons of water located deep underground in the Kamioka mine near Hida. Most neutrinos passed right through the tank, but on rare occasions a neutrino would collide with a water molecule, creating an electron. Those electrons travelled faster than the speed of light in water (which is 75 percent of that in a vacuum) and generated Cherenkov radiation that was observed by photomultiplier tubes on the walls of the tank. In 1987 Kajita was part of the team that used Kamiokande-II to detect neutrinos from Supernova 1987A, which was the first time neutrinos had been observed from a specific object other than the Sun.

Kamiokande-II could also observe neutrinos generated by cosmic rays, high-speed particles (mainly protons) that collide with nuclei in Earth's atmosphere and produce secondary particles. Those secondary particles decay and produce two of the three flavours of neutrinos: electron neutrinos and muon neutrinos. In 1988 Kajita and the other Kamiokande scientists published results showing that the number of muon neutrinos was only 59 percent of the expected value.

Kajita joined the UT's Institute for Cosmic Ray Research in 1988 as a research associate and continued his work at Kamiokande-II. He became an associate professor at the institute in 1992. That same year he and his team published results confirming the deficit of atmospheric muon neutrinos. They suggested that neutrino oscillations in which the "missing" muon neutrinos changed into the third neutrino flavour, tau (which could not be observed by Kamiokande-II), could be the culprit. Neutrinos were thought to be massless, but, in order to oscillate flavours, they must have a very small mass. In 1994 Kajita and his team found a slight dependence of the number of detected muon neutrinos on direction, with more neutrinos coming down than coming up.

In 1996 Kamiokande-II was replaced by Super-Kamiokande, which contained 50,000 tons of water, and Kajita led the studies of the atmospheric neutrinos. After two years of observations, his team definitively confirmed that the number of muon neutrinos coming down from the atmosphere is greater than the number of muon neutrinos coming up from Earth. Since neutrinos rarely interact with matter, the number of neutrinos observed should not depend on the arrival angle. However, that angle effect proved the existence of neutrino flavour oscillations and thus neutrino mass. The neutrinos coming up through Earth travel a longer distance, thousands of kilometres, than the neutrinos coming down, which only travel a few dozen kilometres. Therefore, the up-going neutrinos have more time to undergo an oscillation into tau neutrinos than those coming down.

Kajita became a professor at the Institute for Cosmic Ray Research and director of the Research Center for Cosmic Neutrinos there in 1999. He became director of the institute in 2008.

Guest of Honour of IEMTRONICS 2022

Nobel Laureate



Prof. Konstantin Novoselov

Honourable Nobel Laureate, Professor at the Centre for Advanced 2D Materials, National University of Singapore, Langworthy Professor in the School of Physics and Astronomy, University of Manchester, Manchester, United Kingdom

Bio: Kostya Novoselov made it into a shortlist of scientists with multiple hot papers for the years 2007–2008 (shared second place with 13 hot papers) and 2009 (5th place with 12 hot papers).

In 2014 Kostya Novoselov was included in the list of the most highly cited researchers. He was also named among the 17 hottest researchers worldwide—“individuals who have published the greatest number of hot papers during 2012–2013”.

Novoselov joined the National University of Singapore’s Centre for Advanced 2D Materials in 2019, making him the first Nobel laureate to join a Singaporean university.

Awards and honours

- 2007 Nicholas Kurti European Science Prize “to promote and recognise the novel work of young scientists working in the fields of Low Temperatures and/or High Magnetic Fields.”
- 2008 Technology Review-35 Young Innovator
- 2008 University of Manchester Researcher of the Year.
- 2008 Europhysics Prize, jointly with Geim, “for discovering and isolating a single free-standing atomic layer of carbon (graphene) and elucidating its remarkable electronic properties.”
- 2008 International Union of Pure and Applied Physics Young Scientist Prize, “for his contribution in the discovery of graphene and for pioneering studies of its extraordinary properties.”
- 2010 Nobel Prize in Physics, jointly with Andre Geim, “for groundbreaking experiments regarding the material graphene.” Novoselov was the youngest Nobel laureate in physics since Brian Josephson in 1973, and in any field since Rigoberta Menchú (Peace) in 1992.
- 2010 Knight Commander of the Order of the Netherlands Lion
- 2010 Honorary Fellow of the Royal Society of Chemistry (HonFRSC)
- 2010 Honorary Professor of Moscow Institute of Physics and Technology
- 2011 Honorary Doctorate from the University of Manchester
- 2011 Honorary Fellow of the Institute of Physics (HonFInstP)
- 2011 Elected Fellow of the Royal Society (FRS)
- 2011 W. L. Bragg Lecture Prize from the International Union of Crystallography “... for his work on two-dimensional atomic crystals”
- 2012 Knight Bachelor in the 2012 New Year Honours for services to science.
- 2012 Chosen among “Britain’s 50 New Radicals” by NESTA and The Observer

- 2012 The Kohn Prize Lecture “...for development of a new class of materials: two-dimensional atomic crystals”
- 2013 Appointed Langworthy Professor of Physics, University of Manchester
- 2013 Leverhulme Medal (Royal Society) “...for revolutionary work on graphene, other two-dimensional crystals and their heterostructures that has great potential for a number of applications, from electronics to energy”
- 2013 Awarded Honorary Freedom of the City of Manchester “for his groundbreaking work on graphene”, see List of Freedom of the City recipients
- 2013 Elected a foreign member of the Bulgarian Academy of Sciences
- 2014 2nd place in the Discovery Section of the National Science Photography Competition.
- 2014 included in a list of the most highly cited researchers. He was also named among the 17 hottest researchers worldwide – “individuals who have published the greatest number of hot papers during 2012–2013”.
- 2014 awarded the Onsager Medal.
- 2015 elected to be a member of the Academia Europaea.
- 2016 awarded the Carbon Medal.
- 2016 awarded the Dalton Medal.
- 2019 elected a foreign associate of the US National Academy of Sciences
- 2019 elected to be a member of the Asia Pacific Academy of Materials
- 2019 Otto Warburg Prize and Lecture by The Otto Warburg Chemistry Foundation “for the discovery of the unusual quantum properties of one atom thick two-dimensional materials”

His certificate of election to the Royal Society in 2011 reads

Kostya Novoselov’s research interests cover a wide range of topics from mesoscopic superconductivity and ferromagnetism to materials science and biophysics. He studied vortex structures in mesoscopic superconductors, observed atomic-scale movements of ferromagnetic walls, monitored heartbeats of individual bacteria and mimicked gecko’s adhesion mechanism. His breakthrough moment was the discovery of graphene. Novoselov is now widely recognised to be one of the pioneers in this field (as a number of international awards prove) and, together with Prof Geim FRS, leads research on various applications of this new material ranging from electronics, photonics, composite materials, chemistry, etc. Prof. Novoselov is strongly committed to disseminating science through public lectures and media interviews.

Keynote speakers



Dr. Xiaodong Wang

Professor, Columbia University, New York

Bio: Dr. Xiaodong Wang is a professor of electrical engineering in Columbia University in the city of New York. His research interest includes statistical signal processing, genomic signal processing, machine learning, wireless communications, and information theory.

Among his publications is a book entitled “Wireless Communication Systems: Advanced Techniques for Signal Reception”, published by Prentice Hall in 2003. He has served as an associate editor for the IEEE Transactions on Communications, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Signal Processing, and the IEEE Transactions on Information Theory. He is a Fellow of the IEEE and listed as an ISI Highly-cited Author.

Wang received the 1999 NSF CAREER Award, the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award, and the 2011 IEEE Communication Society Award for Outstanding Paper on New Communication Topics. Xiaodong Wang received the Ph.D degree in Electrical Engineering from Princeton University. He is a Professor of Electrical Engineering at Columbia University in New York.



Dr. Kenneth Paterson

Professor, ETH Zürich, Switzerland

Bio: Dr. Kenneth Paterson obtained a B.Sc. in 1990 from the University of Glasgow and a Ph.D. from the University of London in 1993, both in Mathematics. He was then a Royal Society Fellow at the Institute for Signal and Information Processing at the Swiss Federal Institute of Technology, Zurich, from 1993 to 1994. After that, he was a Lloyd's of London Tercentenary Foundation Research Fellow at Royal Holloway, University of London from 1994 to 1996.

In 1996, he joined Hewlett-Packard Laboratories Bristol, becoming a project manager in 1999.

He then joined the Information Security Group at Royal Holloway in 2001, becoming a Reader in 2002 and Professor in 2004. From March 2010 to May 2015, he was an EPSRC Leadership Fellow working on a project entitled Cryptography: Bridging Theory and Practice. In May 2015, he reverted to being a Professor of Information Security.

In April 2019, he joined the Department of Computer Science at ETH Zurich. Since 1 January 2021, he has held the role of Associate Department Head. In addition, he is director of the CAS/DAS in Cyber Security and the Masters programme in Cyber Security. His research over the last two decades has mostly been in the area of Cryptography, with a strong emphasis being on the analysis of deployed cryptographic systems and the development of provably secure solutions to real-world cryptographic problems. He co-founded the Real World Cryptography series of workshops to support the development of this broad area and to strengthen the links between academia and industry. From 2014 to 2019, he was co-chair of the IRTF's research group on Cryptography, CFRG. This group is working to provide expert advice to the IETF in an effort to strengthen the Internet's core security protocols.

His research on the security of TLS (the Lucky 13 attack on CBC-mode encryption in TLS and attacks on RC4) received significant media attention, helped to drive the widespread adoption of TLS 1.2 with its support for modern encryption schemes, and was an important factor in the TLS Working Group's decision to abandon legacy encryption mechanisms in TLS 1.3.

He is lucky to have been the recipient of several prizes and awards for my research. These include a Google Distinguished Paper Award for his joint work with Nadhem AlFardan presenting plaintext recovery attacks against DTLS published at NDSS 2012; an Applied Networking Research Prize from the IRTF for his work with Nadhem AlFardan on the Lucky 13 attack; and an Award for Outstanding Research in Privacy Enhancing Technologies for his work with Mihir Bellare and Phil Rogaway on the Security of symmetric encryption against mass surveillance published at CRYPTO 2014. My work with Martin Albrecht, Jean Paul Degabriele and Torben Hansen on symmetric encryption in SSH won a best paper award at ACM CCS 2016. In 2018, his work won best paper awards at CHES and IMC.

Other career highlights include being selected as Programme Chair for EUROCRYPT 2011, being an invited speaker at ASIACRYPT 2014, and being editor-in-chief of the Journal of Cryptology from 2017 to 2020. He was made a fellow of the IACR in 2017.



Dr. Gil Zussman

Professor, Columbia University, New York

Bio: **Gil Zussman** received the B.Sc. degree in Industrial Engineering and Management and the B.A. degree in Economics (both summa cum laude) from the Technion – Israel Institute of Technology in 1995. He received the M.Sc. degree (summa cum laude) in Operations Research from Tel-Aviv University in 1999 and the Ph.D. degree in Electrical Engineering from the Technion – Israel Institute of Technology in 2004. Between 1995 and 1998, he served as an engineer in the Israel Defense Forces. Between 2004 and 2007 he was a Postdoctoral Associate in LIDS and CNRG at MIT.

In 2008 he joined the faculty of the Department of Electrical Engineering at Columbia University where he is now a Professor. His research interests are in the area of networking, and in particular in the areas of wireless, mobile, and resilient networks. He has been an associate editor of IEEE Transactions on Control of Network Systems, IEEE Transactions on Wireless Communications and Ad Hoc Networks, the Technical Program Committee (TPC) co-chair of ACM MobiHoc'15, IFIP Performance 2011, and a member of a number of TPCs (including the INFOCOM, MobiCom, SIGMETRICS, and MobiHoc committees).

Gil received the Knesset (Israeli Parliament) award for distinguished students, the Marie Curie Outgoing International Fellowship, the Fulbright Fellowship, the DTRA Young Investigator Award, and the NSF CAREER Award. He was the PI of a team that won the 1st place in the 2009 Vodafone Foundation Wireless Innovation Project competition. He is a co-recipient of seven best paper awards, including the ACM SIGMETRICS / IFIP Performance'06 Best Paper Award, the 2011 IEEE Communications Society Award for Advances in Communication, and the ACM CoNext'16 Best Paper Award.



Dr. Kin K. Leung

Professor, Imperial College, London, United Kingdom

Bio: **Dr. Leung** had completed his Ph.D. in computer science, Univ. of California, Los Angeles in 1985; M.S. in computer science, Univ. of California, Los Angeles in 1982 and B.S. in electronics, The Chinese Univ. of Hong Kong, Hong Kong in 1980. His major

Honors and Awards are:

- IEEE Communications Society Leonard G. Abraham Prize, 2021
- S.-UK Science and Technology Stocktake Award for the DAIS ITA Team, 2021
- IET Fellow, 2021
- Member of Academia Europaea, 2012
- IEEE Fellow Evaluation Committee for Communications Society: Member 2009-2011, Chairman 2012-2015
- Royal Society Wolfson Research Merit Award, 2004-2009
- IEEE Fellow for contributions to “Performance analysis, protocol design and control algorithms for communications networks,” 2001
- Lanchester Prize Honorable Mention Award, 1997
- Bell Labs Distinguished Member of Technical Staff Award, 1994

He is Journal Editor of the following:

- ACM Computing Survey (2009-now)
- Journal of Sensor Networks (2005-now)
- IEEE Trans. on Mobile Computing, Steering Committee Chairman (2020-2022) and Member (2014-2016)
- IEEE Trans. on Communications (1997-2011)
- IEEE Trans. on Wireless Communications (2001-2009)
- IEEE Journal on Selected Areas in Communications: Wireless Series (1999-2001)
- Guest editor: IEEE Wireless Communications, 2007
- Guest editor: Journal of Wireless Communications and Mobile Computing, 2005
- Guest editor: Journal on Special Topics in Mobile Networking and Application (MONET), 2003
- Guest editor: IEEE Journal on Selected Areas in Communications, 1997

His current Research Interests includes:

- Machine learning, distributed optimization, stochastic modeling and queueing theory.
- Wireless communications: resource allocation, power control, spread spectrum, MIMO/beamforming antennas, cross-layer designs, link adaptation, MAC, wireless TCP/IP, QoS, network protocols, and sensor, vehicular, ad-hoc and mesh networks.
- Wireless technologies: GSM, EDGE, 3G, 4G and 5G cellular networks, and IEEE 802.11, 802.16 and 802.15 networks.
- Communication networks: TCP/IP, mobility management, real-time applications, network control protocols, traffic modeling.



Dr. Dayan Ban

Professor, University of Waterloo, Canada

Bio: Dayan Ban is a full Professor in Electrical and Computer Engineering and is a researcher at the Waterloo Institute for Nanotechnology.

His expertise lies in the conversion of near infra-red light directly to visible light, design and fabrication of high-performance quantum devices and the development of ultra-sensitive surface plasmon sensors.

Professor Ban successfully improved the efficiency of hybrid organic/inorganic devices by more than one order of magnitude and applied time-domain terahertz spectroscopy to study the device physics of terahertz quantum cascade lasers. Professor Ban's research has also accomplished the fabrication of prototype hybrid organic/inorganic devices by direct tandem integration and the study of the effects of interfacial states on device performance. These devices are responsible for the conversion of near-infrared light directly to visible light (green) at room temperature.

Professor Ban pioneered the development of new methods in scanning probe microscopy to observe, with nanometric spatial resolution, two-dimensional profiles of conductivity and potential inside actively-driven lasers. He also resolved the nanoscopic reason for anomalously high series resistance encountered in ridge waveguide lasers. In addition, Professor Ban reported the first direct experimental observation of electron overbarrier leakage in operating buried heterostructure multi-quantum-well –lasers. His work has provided the first experimental visualization of the inner workings of operating semiconductor lasers, and has also provided a platform for enabling tools for quantum semiconductor device and nanotechnology research.

Research Interests

- Semiconductor quantum devices
- Photonics
- THz technology
- Nanotechnology
- Atomic force microscope
- Fiber-optical communication system
- Silicon Devices
- Terahertz Quantum Cascade Lasers
- Biophotonics
- Scanning Probe Microscopy
- Connectivity and Internet of Things
- Nanofabrication
- IoT Devices
- Application domains

Education

- 2003, Doctorate, Ph.D., University of Toronto
- 1995, Master's, MS, University of Science and Technology of China
- 1993, Bachelor's, BA, University of Science and Technology of China



Dr. Torsten Hoefler

Professor, ETH Zürich, Switzerland

Bio: **Torsten Hoefler** directs the Scalable Parallel Computing Laboratory (SPCL) at D-INFK ETH Zurich. He received his PhD degree in 2007 at Indiana University and started his first professor appointment in 2011 at the University of Illinois at Urbana-Champaign.

Torsten has served as the lead for performance modeling and analysis in the US NSF Blue Waters project at NCSA/UIUC. Since 2013, he is professor of computer science at ETH Zurich and has held visiting positions at Argonne National Laboratories, Sandia National Laboratories, and Microsoft Research Redmond (Station Q).

Dr. Hoefler's research aims at understanding the performance of parallel computing systems ranging from parallel computer architecture through parallel programming to parallel algorithms. He is also active in the application areas of Weather and Climate simulations as well as Machine Learning with a focus on Distributed Deep Learning. In those areas, he has coordinated tens of funded projects and an ERC Starting Grant on Data-Centric Parallel Programming.

He has been chair of the Hot Interconnects conference and technical program chair of the Supercomputing and ACM PASC conferences. He is associate editor of the IEEE Transactions of Parallel and Distributed Computing (TPDS) and the Parallel Computing Journal (PARCO) and a key member of the Message Passing Interface (MPI) Forum.

He has published more than 200 papers in peer-reviewed international conferences and journals and co-authored the latest versions of the MPI specification. He has received best paper awards at the ACM/IEEE Supercomputing Conference in 2010, 2013, and 2014 (SC10, SC13, SC14), EuroMPI 2013, IPDPS'15, ACM HPDC'15 and HPDC'16, ACM OOPSLA'16, and other conferences. Torsten received ETH Zurich's Latsis Prize in 2015, the SIAM SIAG/Supercomputing Junior Scientist Prize in 2012, the IEEE TCSC Young Achievers in Scalable Computing Award in 2013, the Young Alumni Award 2014 from Indiana University, and the best student award 2005 of the Chemnitz University of Technology. Torsten was elected into the first steering committee of ACM's SIGHPC in 2013 and he was re-elected in 2016. His Erdős number is two (via Amnon Barak) and he is an academic descendant of Hermann von Helmholtz.



Dr. M. Jamal Deen

Distinguished University Professor , McMaster University, Canada

Bio: **M. Jamal** is a Distinguished University Professor and Senior Canada Research Chair in Information Technology at McMaster University in Hamilton, Ontario, Canada. He is also the Director of the Micro- and Nano-Systems Laboratory. His research specialty are in the broad areas of electrical engineering and applied physics. He has done his Ph.D. (Electrical Engineering and Applied Physics) from Case Western Reserve University, Cleveland, OH, U.S.A in 1985; M.S. (Electrical Engineering and Applied Physics), from Case Western Reserve University, Cleveland, OH, U.S.A in 1982 and B.Sc. (Physics/Mathematics), from University of Guyana, Turkeyen, Guyana in 1978. His expertise includes Micro-/Nano-/Opto-Electronics, Nanotechnology and Data Analytics for Health and Environmental Applications, Bioimagers, Biosensor; his areas of specializations are Imaging, Sensing and Detection, Integrated Systems, Biomedical, Microelectronics, Communications, Biomedical. His achievements includes:

- F.R.S.C., F.C.A.E., M.E.A.S.A., F.N.A.S.I., F.I.N.A.E., F.I.E.E.E., F.A.P.S., F.E.I.C., F.E.C.S., F.A.A.A.S.;
- Distinguished University Professor; Professor and Senior Canada Research Chair in Information Technology ;
- 2017 Distinguished Visiting Fellowship Award from Royal Academy of Engineering, UK;
- 2017 PIFI Distinguished Scientist Award from Chinses Academy of Sciences;
- 2017 Overseas Academic Masters Scholar Award;
- 2014 IEEE Canada Ham Outstanding Engineering Educator Award;
- 2013 IEEE Canada AGL McNaughton Gold Medal;
- 2013 UWI Vice-Chancellor's Award;
- 2013 Faculty of Engineering Research Achievement Award from McMaster University;
- 2011 IEEE Canada R.A. Fessenden Silver Medal Award;
- 2011 Electronics and Photonics Divison (EPD) Award from the Electrochemical Society ;
- 2009 Technology Achievement Award from the Indo-Canada Chamber of Commerce;
- 2008 Eadie Medal from The Royal Society of Canada;
- 2008 Guyana Award from the Academic Excellence Guyana Awards Council – Canada;
- 2006 Humboldt Research Award from the Alexander von Humboldt Foundation;
- 2006 IBM Faculty Award from IBM Corporation, USA;
- 2002 Distinguished Lecturer from IEEE Electron Device Society;
- 2002 Thomas D. Callinan Award from the Electrochemical Society.
- Doctor – Honoris Causa El Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico, 11 Nov 16.
- Doctor – Honoris Causa Universitat Rovira I Virgili, Tarragona, Spain, 7 March 2014.
- Doctor – Honoris Causa Universidad de Granada, Granada, Spain, 25 May 2012.
- Doctor of Engineering – Honoris Causa University of Waterloo, Waterloo, Ontario, Canada, 18 June 2011.

Content:

SL NO.	PAPER NAME	AUTHORS WITH AFFILIATION	PAGE NO.
1	Integrating Mechanistic Information to Predict Drug-Drug Interactions and Associated Relevance for Decision Support	Adeeb Noor (King Abdulaziz University, Saudi Arabia)	1
2	Design and Implementation of a Very-Low-Power Wireless Network of Sensors in an Underground Utility Tunnel for Medium and High Voltage Transmission Lines	Adil Rachid (Polytechnic University of Catalonia, Spain); Antonio Miguel Lopez Martinez (Polytechnic University of Catalonia (UPC), Spain); Sebastian Moreno Garcia (& Infisat, Spain)	5
3	Design of 24 GHz ISM Band Microstrip Patch Antenna for 5G Communication	Debalina Mollik, Rima Islam, Afrin Binte Anwar and Prodig Kumar Saha Purnendu (American International University Bangladesh Address Including Country Name, Bangladesh)	12
4	Detection of Corona Virus Infection Using Convolutional Neural Network	Al Sameera (Birla Institute of Technology and Science Pilani, Dubai Campus, UAE, United Arab Emirates); Vilas H Gaidhane (Birla Institute of Technology and Science Pilani, Dubai Campus, UAE, India)	18
5	A Review of Cognitive Dynamic Systems and Cognitive IoT	Alessandro Giuliano, Waleed Hilal, Naseem Alsadi and Stephen Andrew Gadsden (McMaster University, Canada); John Yawney (Adastra Corporation, Canada)	24

6	IoT Devices Proximity Authentication in Ad Hoc Network Environment	Ali Abdullah S. AlQahtani (North Carolina A&T State University, USA); Hosam Alamleh (University of North Carolina Wilmington & Louisiana Tech University, USA); Baker Al Smadi (Grambling State University, USA)	31
7	Dynamic Modeling of a Micro Solar Electric Vehicle for Pakistan Using Simulink	Ali Husnain and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	36
8	An Optimum Sizing for a Hybrid Storage System in Solar Water Pumping Using ICA	Amirhossein Jahanfar and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	45
9	A Real-Time Parking Space Occupancy Detection Using Deep Learning Model	Raktim Raihan Prova, Title Shinha, Anamika Basak Pew and Rashedur M Rahman (North South University, Bangladesh)	51
10	Solar PV System for Self-Consumption	Ananna Khan, Abdul Kahar Siddiki and Rashedur M Rahman (North South University, Bangladesh)	58
11	Provision of Information and Detection Systems on Two-Wheeled Motorcycle Accidents	Andi Nur Faisal and Amil Ahmad Ilham (Hasanuddin University, Indonesia); Syafaruddin Syafaruddin (Universitas Hasanuddin Makassar, Indonesia)	66
12	A Data-Centric Machine Learning Approach for Controlling Exploration in	Antonio Jose Bolufe-Rohler and Jordan Luke (University of Prince Edward Island, Canada)	72

	Estimation of Distribution Algorithms		
13	Aggregated Modeling of Synchronous Generators Using Transfer Matrices	Arash Safavizadeh and Erfan Mostajeran (The University of British Columbia, Canada); Seyyedmilad Ebrahimi (University of British Columbia, Canada); Taleb Vahabzadeh (The University of British Columbia, Canada); Juri Jatskevich (University of British Columbia, Canada)	81
14	Systematic Analysis and Proposed AI-Based Technique for Attenuating Inductive and Capacitive Parasitics in Low and Very Low Frequency Antennas	Kate G Francisco, R-jay Relano, Mike Louie Enriquez, Ronnie Concepcion II, Jonah Jahara Baun, Adrian Genevie G. Janairo, Ryan Rhay P. Vicerra and Argel Bandala (De La Salle University, Philippines); Elmer P. Dadios (Philippines, Philippines); Jonathan Dungca (De La Salle University, Philippines)	88
15	IoT Enabled Smart Solar Panel Monitoring System Based on Boltuino Platform	Pallav Dutta, Ashim Mondal and Md Jishan Ali (Aliah University, India)	95
16	A Practical Approach to the Development of a Decision-Supporting System Based on Fuzzy Neural Network in Information and Telecommunication Systems	Avaz Ergashevich Kuvnakov (TUIT, Uzbekistan)	102
17	A Universal Method for Solving the Problem of Bending of Plates of Any	Azamatjon Yusupov (Andijan Machine Building Institute, Uzbekistan)	106

	Shape		
18	Development of Stochastic Distribution Model of Contaminated Water Treatment Complex	Bakhadir Begilov (Nukus Branch of Tashkent University of Information Technologies Named After Muhammad-Al Khorezmi, Uzbekistan)	111
19	Predictive Maintenance and Condition Monitoring in Machine Tools: An IoT Approach	Brett Sicard and Naseem Alsadi (McMaster University, Canada); Petros Spachos (University of Guelph, Canada); Youssef Ziada (Ford Motor Company, Canada); Stephen Andrew Gadsden (McMaster University, Canada)	117
20	Research on Ambient Backscatter Communication Signal Detection Algorithm Based on Digital Terrestrial Multimedia Broadcast	Chen Ruan and Hongyi Wang (National University of Defense Technology, China); Zheng Liming (National University of Defence Technology, China); Jianfei Wu (National University of Defense Technology, China); Fengxian Ma (Tianjin Institute of Advanced Technology, China)	126
21	Dynamic Analysis of Demographic Sentiment	Joshua Weston, Brenden Bickert and Caleb Stasiuk (Thompson Rivers University, Canada); Fadi Alzhouri (Trent University, Canada); Dariush Ebrahimi (Thompson Rivers University, Canada)	131
22	Use of Drones (UAVs) for Pollutant Identification in	Deyby Huamanchahua (Universidad de Ingeniería y	139

	the Industrial Sector: A Technological Review	Tecnología - UTEC, Peru); Julio Huamanchahua and Fabiola Flores (Universidad San Ignacio de Loyola, Peru)	
23	Nursery With Automation and Control Systems for the Production of White Chuño (Tunta)	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Alem Huayta Uribe, Jalber Macuri Vasquez and Hitan Cordova Sanchez (Universidad Continental, Peru)	145
24	Biological Signals for the Control of Robotic Devices in Rehabilitation: An Innovative Review	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Luis Alberto Huamán Lévano (Universidad Continental, Peru); Jose Asencios Chávez (Universidad Tecnológica del Perú, Peru); Nicole Caballero Canchanya (Universidad Nacional Mayor de San Marcos, Peru)	150
25	Human Cinematic Capture and Movement System Through Kinect: A Detailed and Innovative Review	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Jhon Rodrigo Ortiz Zacarias, Yerson Taza Aquino and Jhon Quispe Quispe (Universidad Continental, Peru)	157
26	Transtibial Electromechanical Prosthesis Based on a Parallel Robot: A Innovate Review	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Diego Osores Aguilar and Victor André León Sales (Universidad Nacional Mayor de San Marcos, Peru); Yadhira	164

		Samhira Valenzuela Lino (Universidad Continental, Peru); Harold Huallanca Escalera (Universidad Peruana de Ciencias Aplicadas, Peru)	
27	Knee and Ankle Exoskeletons for Motor Rehabilitation: A Technology Review	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Elvis J. de la Torre Velarde (Universidad Peruana de Ciencias Aplicadas, Peru); Ana Quispe Piña (Universidad Tecnológica del Perú, Peru); Cesar Luciano Otarola Ruiz (Universidad de Ingeniería y Tecnología UTEC, Peru)	171
28	Hand Exoskeletons for Rehabilitation: A Systematic Review	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Sayda Huacre (Universidad Nacional Mayor de San Marcos, Peru); Pedro Toledo Garcia (Universidad de Ingeniería y Tecnología - UTEC, Peru); Jack Aguirre (Universidad Nacional de Trujillo - UNT, Peru)	178
29	Artificial Intelligence Applied in Human Medicine With the Implementation of Prostheses	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Ismael Alvarado Landeo, Erick Surichaqui Montalvo and Kener Velasquez Colorado (Universidad Continental, Peru)	184
30	Land-Mobile Robots for	Deyby Huamanchahua	191

	Rescue and Search: A Technological and Systematic Review	(Universidad de Ingeniería y Tecnología - UTEC, Peru); Kevin Aubert, Mirella Rivas, Eduardo Guerreo, Laura Kodaka and Diego Guevara (Universidad de Ingeniería y Tecnología UTEC, Peru)	
31	Efficient Simulation of Variable-Speed Diesel-Engine Generators Using Constant-Parameter Voltage-Behind-Reactance Formulation	Erfan Mostajeran and Arash Safavizadeh (The University of British Columbia, Canada); Seyyedmilad Ebrahimi and Juri Jatskevich (University of British Columbia, Canada)	197
32	Detection and Analysis Types of DDoS Attack	Erkin Navruzov (National University of Uzbekistan, Uzbekistan)	203
33	Suicide Rate and Factors Analysis: Pre and Post COVID Pandemic	Maryam Heidari and EswaraChandraSai Pamidimukkala (George Mason University, USA)	210
34	Employee Turnover Prediction Model for Garments Organizations of Bangladesh Using Machine Learning Technique	Lutfun Nahar, Farzana Tasnim, Zinnia Sultana and Farjana Akter Tuli (International Islamic University Chittagong, Bangladesh)	218
35	Machine Learning Models to Predict COVID-19 Cases in the DC Metro Area	Maryam Heidari, Michael Thompson, Thao Tran, Ghadah Alshabana and Ashritha Chitimalla (George Mason University, USA)	223
36	Server-Side Distinction of User Mobility Using Machine Learning on Incoming Data Traffic	Hosam Alamleh (University of North Carolina Wilmington & Louisiana Tech University, USA); Ali Abdullah S.	230

		AlQahtani (North Carolina A&T State University, USA); Baker Al Smadi (Grambling State University, USA)	
37	Data Centric DAO: When Blockchain Reigns Over the Cloud	Kamal Azghiou (Mohammed First University & El team, Morocco); Ibrahim Mehdi (University Mohammed 1, Morocco); Moussaab Sbai (Mohammed First University (UMP), Morocco); Mohamed Mazlin (Université Mohammed Premier & ENSAO, Morocco)	234
38	Algorithmic Analysis of the System Based on the Functioning Table and Information Security	Inomjon Yarashov (National University of Uzbekistan, Uzbekistan)	241
39	Using Algorithmic Modeling to Control User Access Based on Functioning Table	Islambek Saymanov and Inomjon Yarashov (National University of Uzbekistan, Uzbekistan)	246
40	Sensory Data Fusion Using Machine Learning Methods for In-Situ Defect Registration in Additive Manufacturing: A Review	J Akhavan (Stevens Institute of Technology, USA); Souran Manoochehri (Chair of Dept ME, Stevens Institute of Technology)	251
41	Unmanned Aerial Vehicle Control Using Hand Gestures and Neural Networks	Rocio Alba-Flores and Jack Nemecek (Georgia Southern University, USA)	261
42	Feature Selection Algorithm Characterization for NIDS Using Machine	Jyoti Verma (Punjabi University Patiala & Punjab Institute of Technology, India);	265

	and Deep Learning	Abhinav Bhandari (Punjabi University, India); Gurpreet Singh (Maharaja Ranjit Singh Punjab Technical University, India)	
43	An Approach to Design Keyboard and Mouse Assisting Device for Handicap Users	Kamran Hameed (Imam AbdulRahman Bin Faisal UNiversity, Saudi Arabia); Syed Mehmood Ali and Uzma Ali (Imam Abdulrahman Bin Faisal University, Saudi Arabia)	272
44	Design of Monitoring System for Respiratory Diagnosis	Kamran Hameed (Imam AbdulRahman Bin Faisal UNiversity, Saudi Arabia); Sana Ijlal Shahrukh and Ijlal Shahrukh Ateeq (Imam Abdulrahman Bin Faisal University, Saudi Arabia)	278
45	Enhanced DV-Hop Node Localization Algorithm Based on Nearest Neighbour Distance and Hop-Count Evaluation in WSNs	Kanika Sood (NITTTR, Chandigarh, India); Kanika Sharma (Punjab University, India); Amod Kumar (CSIR-Central Scientific Instruments Organisation, India)	286
46	An Exploration of Mis/Disinformation in Audio Format Disseminated in Podcasts: A Case Study of Spotify	Kevin Matthe Caramancion (University at Albany, SUNY, USA)	293
47	Same Form, Different Payloads: A Comparative Vector Assessment of DDoS and Disinformation Attacks	Kevin Matthe Caramancion (University at Albany, SUNY, USA)	299
48	Mathematical Modeling of	Kuvonchbek Rakhimberdiev	305

	Credit Scoring System Based on the Monge-Kantorovich Problem	(National University of Uzbekistan Named After Mirzo Ulugbek & Nuuz, Uzbekistan)	
49	Forecasting Model Comparison for Soil Moisture to Obtain Optimal Plant Growth	Sachintha Balasooriya (Kyoto University of Advance Sciences, Japan); Lasith Yasakethu (Sri Lanka Technological Campus, Sri Lanka)	312
50	Detection and Quantitative Prediction of Diplocarpon Earlianum Infection Rate in Strawberry Leaves Using Population-Based Recurrent Neural Network	Oliver John Alajas, Ronnie Concepcion II, Argel Bandala, Edwin Sybingco and Ryan Rhay P. Vicerra (De La Salle University, Philippines); Elmer P. Dadios (Philippines, Philippines); Christan Mendigoria, Heinrick Aquino and Leonard Ambata (De La Salle University, Philippines); Bernardo Duarte (University of Lisbon, Portugal)	319
51	Lower-Limb Exoskeleton Systems for Rehabilitation And/Or Assistance: A Review	Dana Terrazas-Rodas, Lisbeth Rocca-Huaman, César Ramírez-Amaya and Angel E. Alvarez-Rodriguez (Universidad Tecnológica del Perú, Peru)	327
52	Deep Learning: An Empirical Study on Kimia Path24	Shaikh Rahman and Hayden Wimmer (Georgia Southern University, USA); Loreen Powell (Bloomsburg University of Pennsylvania, USA)	334
53	Biomechanical Prosthesis With EMG Signal	Deyby Huamanchahua (Universidad de Ingeniería y	343

	Acquisition for Patients With Transradial Amputation	Tecnología - UTEC, Peru); Luis Alberto Huamán Lévano (Universidad Continental, Peru)	
54	Bangla Handwritten Character Recognition Method	Lutfun Nahar (International Islamic University Chittagong, Bangladesh)	350
55	The Iso-RSA Cryptographic Scheme	Mamadou I Wade (Howard University, USA)	355
56	The Iso-ElGamal Cryptographic Scheme	Mamadou I Wade (Howard University, USA)	365
57	Connected and Autonomous Vehicles Against a Malware Spread: A Stochastic Modeling Approach	Manal El Mouhib (El Research Team, Morocco); Kamal Azghiou (Mohammed First University & El team, Morocco); Abdelhamid Benali (El Research Team, Morocco)	373
58	Cassava Leaf Disease Detection Using Deep Learning	Manick Manick and Jyoti Srivastava (NIT Hamirpur, India)	379
59	IoT-Based DDoS on Cyber Physical Systems: Research Challenges, Datasets and Future Prospects	Manish Snehi and Abhinav Bhandari (Punjabi University, India)	387
60	Resilience Evaluation of Cyber Risks in Industrial Internet of Things	Mayer Fernandes Silva (Brazil); Herman Lepikson (CIMATEC, Brazil)	395
61	Detecting Various Chemical Samples and Cancer Cells With a Bio-Chemical Sensor by Using LNOI Based Optical Micro Ring	Md Ashif Uddin (Khulna University, Bangladesh); Uzzwal Kumar Dey (Khulna University of Engineering & Technology, Bangladesh);	401

	Resonator (OMRR)	Moriom Akter (Khulna University, Bangladesh)	
62	Proposing A Cloud and Edge Computing Based Decision Supportive Consolidated Farming System by Sensing Various Effective Parameters Using IoT	Md Ashif Uddin (Khulna University, Bangladesh); Uzzwal Kumar Dey (Khulna University of Engineering & Technology, Bangladesh); Moriom Akter (Khulna University, Bangladesh)	407
63	Artificial Magnetic Conductor Unit Cell Design Using Machine Learning Algorithms	Tasfia Nuzhat (Chittagong Independent University, Bangladesh); Md Nazmul Hasan (The University of British Columbia, Canada)	413
64	Development of an IoT-Based Low-Cost Multi-Sensor Buoy for Real-Time Monitoring of Dhaka Canal Water Condition	Ikbal Hasan, Malobika Mukherjee, Rumi Halder and Farzana Yeasmin Rubina (Independent University, Bangladesh, Bangladesh); Md. Abdur Razzak (Independent University, Bangladesh)	420
65	A High Gain Cascaded DC-DC Boost Converter for Electric Vehicle Motor Controller and Other Renewable Energy Applications	Md. Rezanul Haque (Independent University, Bangladesh); K. M. A. Salam (North South University, Bangladesh); Md. Abdur Razzak (Independent University, Bangladesh)	426
66	Performance Evaluation of Secured Blockchain-Based Patient Health Records Sharing Framework	Meryem Abouali (City College of New York, USA); Kartikeya Sharma (City University of New York, USA); Oluwaseyi Ajayi (Vaughns College of Aeronautics and Technology, USA & City College of New	431

		York, USA); Tarek Saadawi (The City University of New York/The City College, USA)	
67	Input Fuzzing for Network-Based Attack Vector on Smartphones	Hosam Alamleh (University of North Carolina Wilmington & Louisiana Tech University, USA); Micah Noyes (University of North Carolina Wilmington, USA)	438
68	Developments Pertaining to the Characteristics of the Sites of HIV Integration Highlighting Its Role in Clinical Research and Its Future With AI: A Review	Minakshi Boruah and Ranjita Das (NIT Mizoram, India)	442
69	Design and Development of a Smart Garage Door System	Mohamed Imran Mohamed Ariff and Farah Diyana Mohamad Fadzir (UiTM Cawangan Perak, Kampus Tapah, Malaysia); Noreen Izza Arshad (Universiti Teknologi Petronas, Malaysia)	449
70	IronMan: An Android-Web Based Application for Laundry Services	Mohammad Moshfique Uddin, Rohit Roy, Saima Alam Miduri and Rashedur M Rahman (North South University, Bangladesh)	455
71	A Brief Overview on Security Challenges and Protocols in Internet of Things Application	Gajjala Savithri (YSR Architecture and Fine Arts University, India); Bhabendu kumar Mohanta (GITAM Deemed to Be University, India); Mohan Kumar Dehury (Koneru Lakshmaiah Education Foundation, India)	463

72	Brain Waves Pattern Recognition Using LSTM-RNN for Internet of Brain-Controlled Things (IoBCT) Applications	Mokhles Mawlood Abdulghani (University of North Dakota, Canada); Farah Fargo (Intel, USA); Haider Khaleel Raad (Xavier University, USA); Olivier Franza (Intel, USA)	470
73	Low-Power and High Speed SRAM for Ultra Low Power Applications	Neha Meshram and Govind Prasad (IIIT Naya Raipur, India); Divaker Sharma (Jamia Millia Islamia- A Central University, India); Bipin Chandra Mandi (DSPM IIIT Naya Raipur, India)	475
74	Smart Home Automation IoT System for Disabled and Elderly	Nesreen Alsbou (University of Central Oklahoma, USA); Naveen Thirunilath (UCO, USA); Imad Ali (University of Oklahoma Health Sciences Center, USA)	481
75	IoT-Based Smart Hospital Using Cisco Packet Tracer Analysis	Nesreen Alsbou (University of Central Oklahoma, USA); Dakota Price (UCO, USA); Imad Ali (University of Oklahoma Health Sciences Center, USA)	486
76	Simulators and Testbeds for IIoT Development and Validation	Nicholas J Jeffrey (University of Oviedo, Canada); Qing Tan (Athabasca University, Canada); Jose R. Villar (University of Oviedo, Spain)	491
77	Image Captioning- Bangladesh's Heritage Perspective Using Deep Learning	Sarowar Alam, Khalidul Islam, Nishat Sharmila, Ziaur Rahman Sovon and Rashedur M Rahman (North South	496

		University, Bangladesh)	
78	Audio Band Analog Signal Measurement Instrument for Vocational School Practicum Aids	Nyoman Karna, Ridha Negara, Bagus Aditya and Adinda Fatkhah Gifary (Telkom University, Indonesia); Dewa Rahyuni (Universitas Padjajaran, Indonesia)	504
79	Enhancing SARS-CoV-2 Variants Research With Blockchain Architecture	Oluwaseyi Ajayi (Vaughns College of Aeronautics and Technology); Tarek Saadawi (The City University of New York/The City College, USA)	510
80	Convolutional Neural Network Structure to Detect and Localize CTC Using Image Processing	Shorouq Al-Eidi (Memorial University of Newfoundland, Canada); Omar Darwish (Eastern Michigan University, USA); Ghaith Husari (East Tennessee State University, USA); Yuanzhu Chen (Queen's University, Canada); Mahmoud Elkhodr (Central Queensland University, Australia)	517
81	FPGA Implementation of Phase Recovery Technique for Complex Transforms	Poorvi Bhaskar (SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, INDIA, India); Yuvaraj S (SRMIST, India); Palanisamy P (National Institute of Technology, Tiruchirappalli & NIT, Trichy, India); Thilagavathy R (NIT Trichy, India)	524
82	VLSI Implementation of a Real-Time Modified	Pradyut Kumar Sanki (SRM University-AP, India); Bevara	530

	Decision-Based Algorithm for Impulse Noise Removal	Vasudeva (SRMAP, AMARAVATI, India); Medarametla Depthi Supriya, Devireddy Vignesh, Peram Bhanu Sai Harshath and Sravya Kuchina (SRM University AP, India)	
83	Envisioning A Light-Based Quantum-Computational Nano-Cyborg	Pravir Malik (Deep Order Technologies, USA)	540
84	Smart Irrigation Systems: Soil Monitoring and Disease Detection for Precision Agriculture	Premsai Peddi and Anuragh Dasgupta (Birla Institute of Technology and Science Pilani, Dubai Campus, UAE, United Arab Emirates); Vilas H Gaidhane (Birla Institute of Technology and Science Pilani, Dubai Campus, UAE, India)	548
85	Bimetals (Au-Pd, Au-Pt) Loaded WO ₃ Hybridized Graphene Oxide FET Sensors for Selective Detection of Acetone	Radha Bhardwaj (BITS, Pilani, India)	555
86	Intelligent Reflecting Surfaces in UAV-Assisted 6G Networks: An Approach for Enhanced Propagation and Spectral Characteristics	Mobasshir Mahbub (Ahsanullah University of Science and Technology, Bangladesh); Raed Shubair (New York University (NYU) Abu Dhabi, United Arab Emirates)	560
87	Intelligent Reflecting Surfaces for Multi-Access Edge Computing in 6G Wireless Networks	Mobasshir Mahbub (Ahsanullah University of Science and Technology, Bangladesh); Raed Shubair (New York University (NYU)	566

		Abu Dhabi, United Arab Emirates)	
88	Probing the States Around the Charge Neutrality Point of Reduced Graphene Oxide With Time-Resolved Gated Kelvin Probe Force Microscopy	Ragul S (Sardar Patel Road & IIT Madras, India); Soumya Dutta and Debdutta Ray (Indian Institute of Technology Madras, India)	572
89	Broadband Printed Dipole Antennas	Rajendra Ghosh (Vidyasagar University, India)	579
90	SQL ChatBot - Using Context Free Grammar	Rajvardhan Patil, Sorio Boit and Nathaniel Bowman (Grand Valley State University, USA)	587
91	Multiobjective Optimal Control of Power Electronic Loads in Small Scale Power Systems	Ramitha Kalhara Dissanayake (University of Peradeniya, Sri Lanka); Amal Wimalarathna (RMIT University, Australia); Anushka Dissanayake (Schweitzer Engineering Laboratories, USA)	594
92	Explicite Model of a Wheel-Soil Interaction Over a Rough Terrain Using Terramechanics Low	Rania Majdoubi (Mohammed V University in Rabat & LCS Laboratory, Faculty of Sciences, Mohammed V University in Rabat, Morocco)	600
93	Optimal Inventory Policy in Oil Transportation: A Case Study	Rasha Kashef (Ryerson University, Canada); Shuo Xu (Adastra North America, Canada)	607
94	Simulating Software Support Delays in a 24/7 Environment Using	Rasha Kashef (Ryerson University, Canada); Shabbir Mirza (TD Canada Trust,	613

	Discrete Event Simulation	Canada)	
95	Intelligent Feature Selection on Multivariate Dataset Using Advanced Data Profiling	Ritu Chaturvedi (University of Guelph, Canada); Vandana Patnaik (Mass General Brigham, Boston, MA, Canada)	619
96	Optimization of Subsurface Imaging Antenna Capacitance Through Geometry Modeling Using Archimedes, Lichtenberg and Henry Gas Solubility Metaheuristics	Adrian Genevie G. Janairo, Jonah Jahara Baun, Ronnie Concepcion II, R-jay Relano, Kate G Francisco, Mike Louie Enriquez, Argel Bandala, Ryan Rhay P. Vicerra and Melchizedek Alipio (De La Salle University, Philippines); Elmer P. Dadios (Philippines, Philippines)	625
97	Employing Nonlinear Model Based PI Like Controller for the Time Varying System	Atanu Panda (IEM, India); Samrat Banerjee and Indrajit Pandey (Techno International New Town, India); Parijat Bhowmick (IIT Guwahati, India)	633
98	A Novel Multifaceted Deep Learning-Based Mobile Application for Accurate and Efficient Waste Classification and Increased Composting Engagement in Communities	Samyak Shrimali (Jesuit High School, USA)	637
99	Remote Elemental Analysis System for Liquid Using Sonoluminescence	Sardini Sayidatun Nisa Saillellah, Hideharu Takahashi and Hiroshige Kikura (Tokyo Institute of Technology, Japan)	642
100	Detailed Bond Graph	Sayed Arash Omid, Rideout	648

	Modeling of PV-Battery System	Geoff and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	
101	Design and Simulate a 500 MW Grid-Connected PV Farm for Labrador	Sayed Arfat Alam Quadri, Mohamad Mahdi Baalbaki, Andrew Chacko and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	655
102	Coordinated Motion and Force Control of Multi-Rover Robotics System With Mecanum Wheels	Serdar S Kalaycioglu (Toronto Metropolitan University & Canadian Space Research Inc., Canada)	663
103	Analysis of the Influence of Factors on Flight Delays in the United States Using the Construction of a Mathematical Model and Regression Analysis	Sergei Kurashkin, Timofey Kireev, Vladislav Kukartsev, Alesya Pilipenko, Anastasiya Rukosueva and Viktor Suetin (Reshetnev Siberian State University of Science and Technology, Russia)	672
104	Construction of a Factor Model Focused on the Reduction of Difficulties in Road Traffic	Sergei Kurashkin, Natalia Fedorova, Aleksander Myrugin, Elena Filushina, Yuriy Seregin, Elena Vaitekunene and Yuriy Danilchenko (Reshetnev Siberian State University of Science and Technology, Russia)	677
105	Theoretical Foundations of the Development Strategy of the Organization for the Production of the Refrigeration Equipment	Sergei Kurashkin, Aleksander Myrugin, Elena Filushina, Yuriy Seregin, Natalia Fedorova, Elena Vaitekunene and Dmitriy Eremeev (Reshetnev Siberian State University of Science and Technology, Russia)	683

106	Methods and Tools for Developing an Organization Development Strategy	Sergei Kurashkin, Vladislav Kukartsev, Elizaveta Shutkina, Kristina Moiseeva, Ekaterina Volneikina and Timofey Kireev (Reshetnev Siberian State University of Science and Technology, Russia)	690
107	Practical-Oriented Method of Development of Strategy of Development of a Production Enterprise	Sergei Kurashkin, Vladislav Dmitriev, Kristina Moiseeva, Alexander Korostelev and Alexander Stashkevich (Reshetnev Siberian State University of Science and Technology, Russia)	698
108	Road Map "TechNet" National Technological Platform	Sergei Kurashkin (Reshetnev Siberian State University of Science and Technology, Russia); Alena Stupina (Siberian Federal University, Russia); Natalia Fedorova, Yuriy Danilchenko, Dmitriy Eremeev, Elena Vaitekunene and Yuriy Seregin (Reshetnev Siberian State University of Science and Technology, Russia)	705
109	Analysis of Data in Solving the Problem of Reducing the Accident Rate Through the Use of Special Means on Public Roads	Sergei Kurashkin and Vladislav Kukartsev (Reshetnev Siberian State University of Science and Technology, Russia); Anton Mikhalev (Siberian Federal University, Russia); Alexander Stashkevich, Kristina Moiseeva and Igor Kauts (Reshetnev Siberian State University of Science	712

		and Technology, Russia)	
110	Establishment of a Model for Managing Organizational Attendance Based on Data Analysis	Sergei Kurashkin and Viktor Suetin (Reshetnev Siberian State University of Science and Technology, Russia); Anton Mikhalev (Siberian Federal University, Russia); Alexander Korostelev (Reshetnev Siberian State University of Science and Technology, Russia); Vladimir Grishko (Siberian Federal University, Russia)	716
111	Using UML to Describe the Development of Software Products Using an Object Approach	Sergei Kurashkin and Evgeniya Semenova (Reshetnev Siberian State University of Science and Technology, Russia); Vadim Sergeevich Tynchenko (Reshetnev Siberian State University of Science and Technology & Siberian Federal University, Russia); Sofya Chashchina, Viktor Suetin and Alexander Stashkevich (Reshetnev Siberian State University of Science and Technology, Russia)	722
112	IoT-Based Cyber-Physical Distribution System Planning	Shaben Kayamboo and Biplob Ray (Central Queensland University, Australia); Narottam K. Das (CQUniversity Australia, Australia); Mary Tom (Central Queensland University, Australia)	726
113	The Million Improvised	Shahriar Khan (Independent	732

	Electric Rickshaws in Bangladesh; Preliminary Survey and Analysis	University, Bangladesh)	
114	The USB Powered Miniature Tesla Coil, With the Filament Bulb, Fluorescent Lamp and Discharge to Body	Shahriar Khan (Independent University, Bangladesh); Simoom Rahman (Independent University, Bangladesh, Bangladesh)	739
115	The Improvised Three-Wheeler Electric Vehicle Solutions in Bangladesh: Equipment and Experimental Waveforms	Shahriar Khan (Independent University, Bangladesh)	746
116	The Small 3-Wheeler Electric Vehicle Solutions in Bangladesh; Survey, Progress and Documentation	Shahriar Khan (Independent University, Bangladesh)	753
117	Analog Front-End CMOS Temperature Sensor Interface for Optogenetic Devices	Shahrzad Mohammad Ghasemi and Soliman Mahmoud (University of Sharjah, United Arab Emirates)	761
118	A High-Order Temperature-Compensated Bandgap Voltage Reference With Low Temperature Coefficient	Shalin Huang, Mingdong Li, Peng Yin and Fang Tang (Chongqing University, China)	766
119	FPGA Implementation of Artificial Neural Network(ANN) for ECG Signal Classification	Shatharajupally Vinaykumar (NIT, Trichy, India); Thilagavathy R (NIT Trichy, India)	771
120	Dynamic Modelling and	Sheikh Usman Uddin and	777

	Analysis of Solar Powered Reverse Osmosis Desalination System for Pakistan Using the Bond Graph Model	Rideout Geoff (Memorial University of Newfoundland, Canada)	
121	Design and Analysis of a Solar Powered Water Filtration System for a Community in Black Tickle-Domino	Sheikh Usman Uddin, Onyinyechukwu Chidolue, Abdul Azeez and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	783
122	Reconfigurable Star-Delta VBR Induction Machine Model for Predicting Soft-Starting Transients	Sheraz Baig and Taleb Vahabzadeh (The University of British Columbia, Canada); Seyyedmilad Ebrahimi and Juri Jatskevich (University of British Columbia, Canada)	789
123	Comparative Analysis of Deep Learning and Machine Learning Techniques for Power System Fault Type Classification and Location Prediction	Sivaramarao Bodda and Anjali Thawait (IIT Bhilai, India); Prashant Agnihotri (Indian Institute of Technology Bhilai, India)	796
124	Machine Learning Approach to Predict Road Accidents in the United States	Maryam Heidari, Sri Siddhartha Reddy, Yen Ling Chao and Lakshmi Praneetha Kotikalapudi (George Mason University, USA)	805
125	Dynamic Simulation of a Microgrid System for a University Community in Nigeria	Stephen Ogbikaya and Mohammad Tariq Iqbal (Memorial University of Newfoundland, Canada)	811
126	Design of a Computer	Surya Kant (National Institute	818

	Stereo Vision Based Road Marking System	of Technical Teachers Training and Research, India); Amod Kumar (CSIR-Central Scientific Instruments Organisation, India); Garima Saini (NITTTR, India)	
127	A Machine Learning Approach for the Early Detection of Dementia	Sven Broman, Elizabeth O'Hara and Md L Ali (Rider University, USA)	825
128	Semantic Segmentation Using Modified U-Net for Autonomous Driving	T Sugirtha (National Institute of Technology Tiruchirappalli Tamilnadu India, India); M. Sridevi (National Institute of Technology, India)	831
129	Performance Analysis of a New Non-Contact, Potentiometric Angle Sensor	Utpol Tarafdar (National Institute of Technology, Calicut, Kerala, India); Mithun Sakthivel (National Institute of Technology Calicut, India)	838
130	Identifying Functional and Non-Functional Software Requirements From User App Reviews	Dev Dave and Vaibhav Anu (Montclair State University, USA)	845
131	Studying-Alive: A Holistic Wellness Application for College Students	Natasia Fernandez and Vaibhav Anu (Montclair State University, USA)	851
132	A Review of Cognitive Dynamic Systems and Its Overarching Functions	Waleed Hilal, Alessandro Giuliano and Stephen Andrew Gadsden (McMaster University, Canada); John Yawney (Adastra Corporation, Canada)	858
133	Balancing Sampling	Wenwu Deng (University of	868

	Frequencies for Multi-Modality IoT Systems: Smart Shoe as an Example	Science and Technology of China, China); Kangyu Chen (The First Affiliated Hospital of USTC, China); Qijun Ying and Jingyuan Cheng (University of Science and Technology of China, China)	
134	Enhanced Photon Detection Probability Model for Single-Photon Avalanche Diodes in TCAD With Machine Learning	Xuanyu Qian, Wei Jiang and M. Jamal Deen (McMaster University, Canada)	875
135	Design of an Automated Manure Collection System for the Production of Biogas Through Biodigesters	Jhon Rodrigo Ortiz Zacarias, Iraiz Quintanilla Mosquera, Jesus Eduardo Rosales Fierro, Sliver Del Carpio Ramirez, Carlos Coaquira Rojo, Yadhira Samhira Valenzuela Lino and Nabil Moggiano (Universidad Continental, Peru)	881
136	Design of an Automated Feeding and Drinking System for Turkeys at Different Stages of Their Development	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Yadhira Samhira Valenzuela Lino, Jesus Eduardo Rosales Fierro, Jhon Rodrigo Ortiz Zacarias, Nabil Moggiano and Carlos Coaquira Rojo (Universidad Continental, Peru)	887
137	Automated Design of a Cleaning Machine and an Environmental Temperature Controller for Guinea Pig Houses	Jhon Rodrigo Ortiz Zacarias, Yadhira Samhira Valenzuela Lino, Yossef Rojas Tapara, Sliver Del Carpio Ramirez, Carlos Coaquira Rojo and Frank Zarate Peña	893

		(Universidad Continental, Peru)	
138	Machine Learning Performances for Covid-19 Images Classification Based Histogram of Oriented Gradients Features	Yessi Jusman (Universitas Muhammadiyah Yogyakarta, Indonesia); Wikan Tyassari (University of Muhammadiyah Yogyakarta, Indonesia); Difa Nisrina, Fahrul Galih Santosa and Nugroho Abdi Prayitno (Universitas Muhammadiyah Yogyakarta, Indonesia)	898
139	Multilayer Aperture Coupled Single Band Second Order Bandpass Patch Resonator	Mohan K N (Vellore Institute of Technology, Vellore, India); Yogesh Kumar Choukiker (VIT University, India)	904
140	Analysis and Evaluation of A Eod Robot Prototype	Yuri L. Silva, Elvis Supo, Milton Amadeo Ccallata, Jesús Mamani, Mario Betancur, Brunno Pino, Pablo Pari and Erasmo Sulla (Universidad Nacional de San Agustín de Arequipa, Peru)	908
141	A Model of Classification of Consumers on the Retail Electricity Market	Aleksandr V. Belov and Maria Monina (National Research University Higher School of Economics, Russia); Zhenisgul Rakhmetullina (D. Serikbayev East Kazakhstan State Technical University, Kazakhstan)	914
142	Ontology Construction of City Hotline Service for Urban Grassroots Governance	Zuohai Chen (Qilu University of Technology, China)	920

143	Leg Geometry Optimization of Thermoelectric Cooler to Maximize COP Through Gaussian Process Modelling	Ethan Robyn V. Ebuén, La Verne Certeza, Johannes Kurt Tecson, Jowen Louis Francisco, Carl Vincent Villanueva and Jomar Lord Cauton (University of Santo Tomas, Philippines)	930
144	A Sliding Mode Based Finite Time Consensus Protocol for Heterogeneous Multi Agent UAS	Madhumita Pal (Institute of Engineering & Management, Kolkata, India); Titas Bera (Tata Consultancy Services, India)	939
145	Improving Accessibility of Remote Drone Control With a Streamlined Computer Vision Approach	Evan Lowhorn (Georgia Southern University, USA)	946
146	Terminal Sliding Mode Control(TSMC) Based Cooperative Load Transportation Using Multiple Drones	Subhojit Das (Institute of Engineering & Management, India); Madhumita Pal (Institute of Engineering & Management, Kolkata, India); Sagnik Banerjee, Souhardya Das, Rishav Kumar, Sudhanshu Shekhar and Shreyan Ghosh (Institute of Engineering & Management, India)	951
147	Mechanical and Electronic Design of a Prototype of a Modular Exoskeleton for Lower-Limbs	Deyby Huamanchahua (Universidad de Ingeniería y Tecnología - UTEC, Peru); Yerson Taza Aquino (Universidad Continental, Peru)	960
148	Algorithmic Formalization of Risk Synthesis Based on	Inomjon Yarashov (National University of Uzbekistan,	966

	Functioning Table	Uzbekistan)	
149	Intelligent Monitoring and Control of Wind Turbine Prototype Using Internet of Things (IoT)	Pallav Dutta, Md Jishan Ali and Ashim Mondal (Aliah University, India)	970
150	Sentiment Analysis and NLP Models for Identifying Biases of Online News Stations	Maryam Heidari, Anuska Acharya and Grace Cox (George Mason University, USA)	976
151	Customer Relationship Analysis to Improve Satisfaction Rate in Banking	Maryam Heidari (George Mason University, USA)	985
152	Machine Learning Models for Mental Health Analysis Based on Religious Impact	Maryam Heidari, Waseem Ashraf and Krishnasri Dontha (George Mason University, USA)	995
153	A Survey on the Jamming and Spoofing UAV Network Attacks and How Machine Learning is an Effective Against Them	Faisal Alrefaei (Embry-Riddle Aeronautical University, USA); Abdullah Alzahrani (Oakland University, USA)	1001
154	Highly Sensitive Hydrogen Gas Sensor Based on Fe ₂ O ₃ : ZnO Nanostructured Thin Film	Mikayel Seryozha Aleksanyan (Centre of Semiconductor Devices and Nanotechnologies, Armenia); Artak Sayunts, Gevorg Shahkhatuni, Zarine Simonyan and Gohar Shahnazaryan (Center of Semiconductor Devices and Nanotechnologies, Armenia); Vladimir Aroutiounian (Yerevan State University,	1008

		Armenia)	
155	A Review on Trends in the Northern Virginia (NOVA) Housing Market and Understanding Home Characteristics for ML Models	Maryam Heidari, Bethlehem Belaine, Ryan Thomas, Omar Janjua and Paul Karcic (George Mason University, USA)	1013
156	Preliminary Results on Analyzing Credit Card Fraud Detection	Maryam Heidari and Arthi Reddy Kotha (George Mason University, USA)	1022
157	Airbnb Data Analysis of Florida Real Estate	Maryam Heidari, Geetna Penmatsa and Keerthana Vallamkonda (George Mason University, USA)	1029
158	Use Machine Learning Technologies in E-Learning	Taslina Akter (IUB, Bangladesh)	1035

Integrating Mechanistic Information to Predict Drug-Drug Interactions and Associated Relevance for Decision Support

Adeeb Noor

Department of Information Technology, Faculty of Computing and Information Technology

King Abdulaziz University

Jeddah 80221, Saudi Arabia

arnoor@kau.edu.sa

Abstract—While computational methods offer great potential in predicting drug-drug interactions (DDIs), such predictions as of yet have limited utility in supporting clinical decision-making; in particular, there exists especial difficulty in deriving interaction mechanisms from the vast abundance of available information on potential DDIs. Here, we present a backward-chaining inference algorithm that operates on a knowledge graph integrating multiple types of mechanistic information, from metabolizing enzymes to genetic variants. Given two drugs of interest, this algorithm applies complex rules to identify evidence supporting their potential interaction, which in turn suggests their mechanism of interaction. An evaluation of the ruleset using two widely-used drugs with a suspected interaction, the antibiotic levofloxacin and the chemotherapeutic irinotecan, successfully identified pharmacological and biomedical features that support and may explain their interaction. This algorithm represents a first step toward effectively assessing the clinical relevance of identified DDIs, and of identifying pairs of interacting drugs that may be validated in the experimental setting to support clinical decision-making and ultimately improve medication safety.

Keywords—artificial intelligence, decision support, rule-based inference, knowledge graph, drug-drug interactions

I. INTRODUCTION

Adverse drug events contribute to patient morbidity, mortality, and healthcare costs, and are becoming of greater concern as the role of drug therapy expands and polypharmacy becomes more frequent. These include drug-drug interactions (DDIs), for example, toxicity or reduced efficacy, which may result when a patient is co-administered two or more drugs. In the US alone, DDIs were implicated in 231,000 emergency room visits in a 26-month period [1] and nearly a quarter of hospital admissions [2]. Accordingly, there is great interest in using available computational resources and published material to 1) efficiently identify combinations of drugs that can produce clinically meaningful effects and 2) effectively share that information with clinicians so as to make decisions that minimize patient risk. Realizing both of these objectives at once is challenging for extant methods of identifying DDIs.

Classically, DDI discovery has been carried out by means of single-pathway studies using either *in vivo* or *in vitro* approaches. These have often focused on cytochrome enzymes

(CYPs) and disregarded other possible interaction mechanisms; in addition, most are undertaken during clinical trials, hence typically consider few confounding factors and involve small sample sizes [3]. Thus, while traditional methods provide clinically meaningful information, they are low-throughput and limited in scope, making for a considerable bottleneck in DDI discovery. *In silico* methods of DDI prediction have become of great interest as a solution to this bottleneck. Such methods often begin with knowledge of pharmacologic properties and statistical associations of drugs with health outcomes; extant approaches have considered numerous features [4], [5] and utilized a wide variety of techniques [5]–[10]. However, they also have a common flaw in producing vast lists of prospective DDIs with uncertain clinical relevance. New tools are needed to resolve the critical challenge of accurately identifying clinically meaningful DDIs.

Here, we constructed a rule-based inference algorithm with the goal of simultaneously predicting DDIs and explaining their mechanisms, the better to elucidate the clinical significance of any prospective interaction. Rule-based DDI prediction systems have demonstrated considerable promise in this regard [11]; the present algorithm improves on prior efforts by simultaneously accounting for two different mechanistic layers and sourcing supporting evidence from a wide selection of available resources, thereby drawing upon an expansive knowledge base and distilling it into meaningful explanations of a prospective interaction. Thus, this algorithm represents a key step forward in leveraging the abundance of available information to efficiently identify DDI pairs with prospective clinical relevance and appropriately prioritize them for experimental validation.

II. METHODS

A. Creation of the knowledge graph

The inference algorithm applies rules on a knowledge graph encompassing four core categories of information relating to drug interaction mechanisms: biomolecular, physiological, pharmacological, and genetic. Included biomolecular and physiological information originated with Gene Ontology terms, the National Drug File Reference Terminology, and the National Cancer Institute Thesaurus [12]–[14]. Genetic and pharmacological information were

sourced in November 2021 from the Pharmacogenomics Knowledge Base and DrugBank respectively [15], [16]. All items in the graph were either included in the Unified Medical Language System (UMLS) or were mapped to UMLS concept unique identifiers [17]. All told, this graph represented six kinds of mechanistic information: physiology, biological processes, molecular functions, proteins, genes, and SNPs. To validate the graph, we employed the Protégé tool to construct an ontology and checked its consistency with the Pellet reasoner [18], which determined it to be appropriately consistent.

B. Development of the backward-chaining inference algorithm

We developed a rule-based model of the mechanisms by which drugs interact and used a backward-chaining inference algorithm to identify potential interactions through mechanisms. Emulating human decision-making, the algorithm starts from a hypothesis (that a given drug pair interacts), searches the knowledge graph for supporting evidence defined according to a set of if-then rules, and finally either accepts or rejects the hypothesis [19]. Our algorithm considered two potential levels of interaction: convergent pharmacological effects (inhibition or induction of a necessary transporter or enzyme) and sharing of some features in the four biomedical categories.

The ruleset employed by the algorithm was thus designed to elucidate potential DDIs and the mechanism by which they occur, illustrated in Fig. 1. These six rules were validated by a clinician and specifically identified the following situations:

Layer 1: Pharmacological

- Inhibition of one drug by the other at metabolism or transporter levels
- Induction of one drug by the other at metabolism or transporter levels

Layer 2: Biomedical features

- The drug pair has a pharmacological target in common
- The drug pair have common biomolecular features
- The drug pair have common physiological pathways
- The drug pair have common influencing genetic variations

A given pair of drugs were considered prospective interactors if found to have any positive hit in the pharmacological layer AND if they also fully shared features within at least two of the four categories of the biomedical layer.

C. Validation of predictions

To evaluate and validate the performance of the inference algorithm in predicting drug interactions, we tested it on a pair of drugs that are administered to cancer patients and that evidence suggests have potential to interact. The first, irinotecan, is an antineoplastic chemotherapy agent. The second, levofloxacin, is a broad-spectrum quinolone antibiotic frequently used to treat infections in cancer patients and, in those at high risk of febrile neutropenia, as a post-chemotherapy prophylactic [20]. The respective indications of these drugs mean they are likely to be concurrently administered. In addition, both drugs have well-documented pharmacokinetics profiles, which provide a wealth of data for the algorithm to consider. While no study has yet reported a conflicting interaction of irinotecan and levofloxacin in practice, there is some indication that concurrent administration of antineoplastic agents and oral quinolone antibiotics may lead to reduced plasma concentrations of the latter, and hence reduced therapeutic effect [21], [22]. Thus, this drug pair constitutes a likely candidate for interaction.

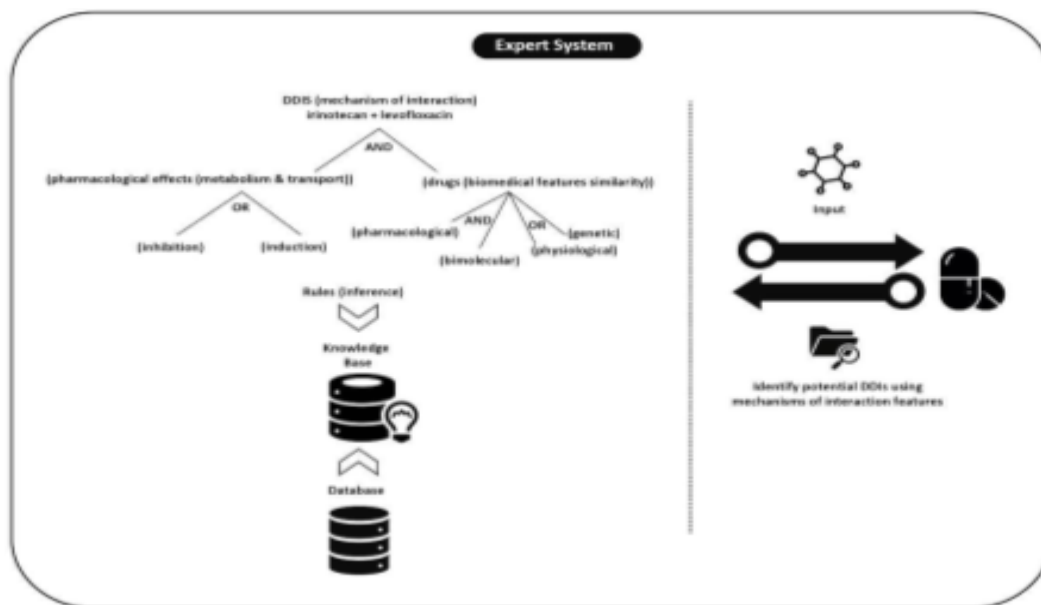


Figure 1: Summary diagram of the rule-based framework for predicting potential drug-drug interactions and their mechanisms.

III. RESULTS

Running the inference algorithm on the candidate interactors levofloxacin and irinotecan successfully identified potential mechanisms of interaction for these drugs on both pharmacological and biomedical feature levels. On the pharmacological level, the algorithm identified two proteins that mutually interact with the two drugs. First, the monooxygenase Cytochrome P450 Family 3 Subfamily A Member 4 (CYP3A4) utilizes irinotecan as a substrate and is inhibited by levofloxacin; and second, the transporter ATP binding cassette subfamily B member 1 (ABCB1, also known as P-glycoprotein) transports irinotecan and is inhibited by levofloxacin. Thus, these proteins represent two points of potential pharmacological conflict for this drug pair when concurrently administered to a patient.

On the biomolecular feature level, our algorithm also identified irinotecan and levofloxacin as sharing three features. First, if a patient is allergic to either of the drugs, administration of the other is also contraindicated; this represents a shared reaction of the body to the two drugs. Second, both drugs adversely affect DNA integrity, which constitutes a shared biomedical impact. Third, the two drugs have in common several classes of molecular groups and structures such as benzene rings, hydroxyl compounds, carboxylic acids and derivatives, hydroxyquinolines, and others; hence, they share chemical characteristics, which may relate to common interactions (as with CYP3A4) or effects of the two drugs.

IV. DISCUSSION

Current state-of-the-art methods for DDI prediction typically yield associations on either pharmacodynamic or pharmacokinetic bases alone; in addition, they consider few features and reference only a fraction of the biomedical resources available [23]–[25]. However, it is also true that existing DDI resources frequently have little overlap in their reported DDIs and may specialize in particular interaction types or mechanisms [26]. Our algorithm thus stands apart from extant methods in that it focuses on the process of evaluation and leverages four different categories of potential interactions to explore and explain DDI mechanisms on multiple levels simultaneously. Applied to irinotecan and levofloxacin as a case study, our algorithm identified two axes through which those drugs may interact: on the pharmacological level, mutual interactions with endogenous proteins, namely the enzyme CYP3A4 and transporter ABCB1, and on the biomedical feature level, commonalities in three feature patterns.

As the incidence of polypharmacy increases, there is potential for exponential growth in associated health risks [27]. It is straightforward to contraindicate concurrent use for drug pairs that have an overt adverse interaction documented in a clinical setting; however, not all DDIs are necessarily overt in their effects, and with the ever-increasing array of drugs on the market and in development, it is not practical to empirically test even a fraction of possible combinations. In addition, as DDIs are an important consideration during drug development and approval, it is imperative to be able to predict interactions before a drug sees wide clinical use. Thus,

it is urgently necessary to understand not only reported but also unobserved and potential DDIs and the mechanisms by which they occur, which in turn necessitates the development of new methods [6]. On top of this, it is essential to identify the clinical relevance of a hypothesized interaction so as to effectively inform regulatory and clinical decision-making. The algorithm presented here represents a first step towards development of a prediction tool that highlights prospective explanatory mechanisms and can be used by clinicians as well as researchers, and so may potentially bring substantial decision support value to present practice.

While this algorithm utilizes a relatively simplistic combined similarity determination in assessing sharing of biomedical feature patterns, that very simplicity represents a springboard on which future work can elaborate and expand. For example, each biomedical feature is currently given the same importance, with the implication that if drugs share a greater number of shared features, there is greater concern of their interaction [28]. As a basic premise this appears sound, and the false alarms reasonably low despite only relatively low similarity being required; however, there may be benefit in employing more comprehensive *in silico* methods to enhance or even replace the current similarity determination. Thus, our work not only constitutes a solid proof of concept but offers potential for further development that can help realize an effective DDI prediction system with utility in accelerating new drug approvals and facilitating clinical decision-making.

REFERENCES

- [1] D. C. Malone *et al.*, “Assessment of potential drug–drug interactions with a prescription claims database,” *Am. J. Health Syst. Pharm.*, vol. 62, no. 19, pp. 1983–1991, Oct. 2005, doi: 10.2146/ajhp040567.
- [2] S. Dechanont, S. Maphanta, B. Butthum, and C. Kongkaew, “Hospital admissions/visits associated with drug–drug interactions: a systematic review and meta-analysis,” *Pharmacoepidemiol. Drug Saf.*, vol. 23, no. 5, pp. 489–497, 2014, doi: 10.1002/pds.3592.
- [3] T. Roblek, T. Vaupotic, A. Mrhar, and M. Lainscak, “Drug–drug interaction software in clinical practice: a systematic review,” *Eur. J. Clin. Pharmacol.*, vol. 71, no. 2, pp. 131–142, Feb. 2015, doi: 10.1007/s00228-014-1786-7.
- [4] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, and C. Friedman, “Drug–drug interaction through molecular structure similarity analysis,” *J. Am. Med. Inform. Assoc.*, vol. 19, no. 6, pp. 1066–1074, Nov. 2012, doi: 10.1136/amiajnl-2012-000935.
- [5] S. Dere and S. Ayvaz, “Prediction of drug–drug interactions by using profile fingerprint vectors and protein similarities,” *Healthc. Inform. Res.*, vol. 26, no. 1, pp. 42–49, Jan. 2020, doi: 10.4258/hir.2020.26.1.42.
- [6] H. Hochheiser *et al.*, “A minimal information model for potential drug–drug interactions,” *Front. Pharmacol.*, vol. 11, p. 608068, 2021, doi: <https://doi.org/10.3389/fphar.2020.608068>.
- [7] S. Liu, B. Tang, Q. Chen, and X. Wang, “Drug–drug interaction extraction via convolutional neural networks,” *Comput. Math. Methods Med.*, vol. 2016, p. e6918381, Jan. 2016, doi: 10.1155/2016/6918381.
- [8] M. Tiftikci, A. Özgür, Y. He, and J. Hur, “Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels,” *BMC Bioinformatics*, vol. 20, no. 21, p. 707, Dec. 2019, doi: 10.1186/s12859-019-3195-5.
- [9] X. Sun, L. Ma, X. Du, J. Feng, and K. Dong, “Deep convolution neural networks for drug–drug interaction extraction,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2018, pp. 1662–1668. doi: 10.1109/BIBM.2018.8621405.
- [10] A. Noor, “A data-driven medical decision framework for associating adverse drug events with drug–drug interaction mechanisms,” *J. Healthc. Eng.*, vol. 2022, p. e9132477, Mar. 2022, doi: 10.1155/2022/9132477.
- [11] B. Abu-Nasser, “Medical expert systems survey,” *Int. J. Eng. Inf. Syst.*, vol. 1, no. 7, pp. 218–224, Sep. 2017.
- [12] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, Art. no. 1, May 2000, doi: 10.1038/75556.
- [13] S. H. Brown *et al.*, “VA National Drug File Reference Terminology: a cross-institutional content coverage study,” *Stud. Health Technol. Inform.*, vol. 107, no. Pt 1, pp. 477–481, 2004.

- [14] S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright, "NCI Thesaurus: using science-based terminology to integrate cancer research results," *Stud. Health Technol. Inform.*, vol. 107, no. Pt 1, pp. 33–37, 2004.
- [15] T. E. Klein *et al.*, "Integrating genotype and phenotype information: an overview of the PharmGKB project," *Pharmacogenomics J.*, vol. 1, no. 3, pp. 167–170, 2001, doi: 10.1038/sj.tpj.6500035.
- [16] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.
- [17] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D267–D270, Jan. 2004, doi: 10.1093/nar/gkh061.
- [18] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: a practical OWL-DL reasoner," *J. Web Semant.*, vol. 5, no. 2, pp. 51–53, Jun. 2007, doi: 10.1016/j.websem.2007.03.004.
- [19] A. Al-Ajlan, "The comparison between forward and backward chaining," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 2, pp. 106–113, Apr. 2015, doi: 10.7763/IJMLC.2015.V5.492.
- [20] A. G. Freifeld *et al.*, "Clinical practice guideline for the use of antimicrobial agents in neutropenic patients with cancer: 2010 update by the Infectious Diseases Society of America," *Clin. Infect. Dis.*, vol. 52, no. 4, pp. e56–e93, Feb. 2011, doi: 10.1093/cid/cir073.
- [21] D. N. Fish and A. T. Chow, "The clinical pharmacokinetics of levofloxacin," *Clin. Pharmacokinet.*, vol. 32, no. 2, pp. 101–119, Feb. 1997, doi: 10.2165/00003088-199732020-00002.
- [22] T. Ito, I. Yano, K. Tanaka, and K.-I. Inui, "Transport of quinolone antibacterial drugs by human p-glycoprotein expressed in a kidney epithelial cell line, LLC-PK1," *J. Pharmacol. Exp. Ther.*, vol. 282, no. 2, pp. 955–960, Aug. 1997.
- [23] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Brief. Bioinform.*, vol. 16, no. 6, pp. 1069–1080, Nov. 2015, doi: 10.1093/bib/bbv011.
- [24] A. Noor, A. Assiri, S. Ayvaz, C. Clark, and M. Dumontier, "Drug-drug interaction discovery and demystification using Semantic Web technologies," *J. Am. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 556–564, May 2017, doi: 10.1093/jamia/ocw128.
- [25] A. Lavertu, B. Vora, K. M. Giacomini, R. Altman, and S. Rensi, "A new era in pharmacovigilance: toward real-world data and digital monitoring," *Clin. Pharmacol. Ther.*, vol. 109, no. 5, pp. 1197–1202, 2021, doi: 10.1002/cpt.2172.
- [26] A. Assiri and A. Noor, "A computational approach to predict multi-pathway drug-drug interactions: A case study of irinotecan, a colon cancer medication," *Saudi Pharm. J.*, vol. 28, no. 12, pp. 1507–1513, Dec. 2020, doi: 10.1016/j.jsps.2020.09.017.
- [27] D. M. Qato, G. C. Alexander, R. M. Conti, M. Johnson, P. Schumm, and S. T. Lindau, "Use of prescription and over-the-counter medications and dietary supplements among older adults in the united states," *JAMA J. Am. Med. Assoc.*, vol. 300, no. 24, pp. 2867–2878, Dec. 2008, doi: 10.1001/jama.2008.892.
- [28] A. Assiri and A. Noor, "Anti-DDI resource: a dataset for potential negative reported interaction combinations to improve medical research and decision-making," *J. Healthc. Eng.*, vol. 2022, p. e8904342, Apr. 2022, doi: 10.1155/2022/8904342.

Design and implementation of a very-low-power wireless network of sensors in an underground utility tunnel for medium and high voltage transmission lines

Adil Rachid

*Department of Electronic Engineering
DEE*

*Polytechnic University of Catalonia
UPC*

*Barcelona, Spain
adil.rachid@upc.edu*

Antonio Miguel Lopez Martinez

*Department of Electronic Engineering
DEE*

*Polytechnic University of Catalonia
UPC*

*Barcelona, Spain
antonio.miguel.lopez@upc.edu*

Sebastián Moreno García

*Department of Electronic Engineering
High Technology Physical Engineering,
S.L. (INFISAT)*

*Barcelona, Spain
sebastian.moreno@infisat.com*

Abstract— The reliability of the electrical network and the need to minimize economic losses due to unexpected power outages have led electricity distribution companies to introduce diagnostic and preventive maintenance programs to assess the condition of facilities under normal working and power conditions in order to be able to react quickly in unexpected conditions (fire and floods). The developed system is composed of different types of sensors characterized by their very low power consumption, which detect anomalies in the operation of medium voltage (30Kv) and high voltage (220Kv) line installations located in underground utility tunnels for electricity distribution.

This article describes and develops a very low power communication system in the IoT field which improves the energy efficiency of radio communications between sensor nodes (WSN) by integrating systems that facilitate the operation of multiple hops of the wake-up signal. This provides a longer overall lifespan in comparison to other monitoring systems.

The developed WSN sensor network is installed and tested in an underground service utility tunnel including medium and high voltage transmission lines that belongs to the Endesa group (ENEL) and is located in the city of Barcelona. A web-type user environment has been designed to view the data sent by the sensor network.

Keywords: *Wake-up, Mesh network, IoT, Low-voltage charge pump, WuRx and WuTx radio.*

I. INTRODUCTION

Energy efficiency is one of the most important criteria in the design of a wireless sensor network (WSN). Nowadays,

WSNs are applied in a wide range of fields, such as environmental monitoring, machine surveillance, health monitoring, and traffic control. However, sensor nodes are generally battery-powered and therefore have a very limited lifespan if power management is not performed. Some problems found in most of the currently available applications are the limited time of operation and the need for short-term maintenance (change of batteries). Radio transmission and reception are the two main sources of power consumption. So when a node is up and waiting to receive data, it wastes energy due to listening idle. Therefore, it is essential to investigate new methodologies to avoid wasting the energy of the entire node.

WSNs are typically operated across multiple hops to provide an extended range of coverage. Multi-hop operation must coordinate multiple nodes to allow data to pass through each hop to the destination node, which requires sensor nodes to switch from receive to transmit modes based on the received trigger signal.

Through the efficient communication mechanisms in the sensor nodes, you can turn off your radio and set your microcontroller (MCU) to sleep power mode when it is idle and wake it up when there are possible transmissions. To wake up a sensor node, two methods can be used: a periodic sleep method programmed by the microcontroller with the help of a timer or the integration of an ultra-low power or passive auxiliary radio called wake-up radio (WuRx) to the sensor node waiting for the trigger signal sent by the trigger radio transmitter (WuTx).

Basically, our aim is to design a multi-hop ultra-low power sensor network in line topology, based on starting from the WuRx passive auxiliary radio integration method. In order to

carry out this method, the characteristics of currently existing WuRx will be improved, so they can be integrated in entation with a coverage range similar to the main radio.

The fact that these underground utility tunnels are not straight along their entire route, having curves to the left and to the right and unevenness, as well as the fact that they are underground difficult the communication between nodes and the output of information to the outside. In addition, inside the underground tunnel, a local network of sensors that communicates with an end point (gateway with internet connection) should be created to process the data flow and send it to the web server.

II. LOCATION OF WSN NODES

A mesh network, capable of routing data based on DigiMesh [1], has been used as topology. Digimesh is an alternative proprietary protocol to Zigbee [2], which uses 2.4GHz XBee S1 pro modules [3] that allow that Packet routers can also sleep and can be synchronized with the rest times of the Network. This technique consists in synchronizing time periodically to save energy in the nodes of the WSN network.

The underground utility tunnel has a route of 2.4 km. The profile is shown in the following figure.

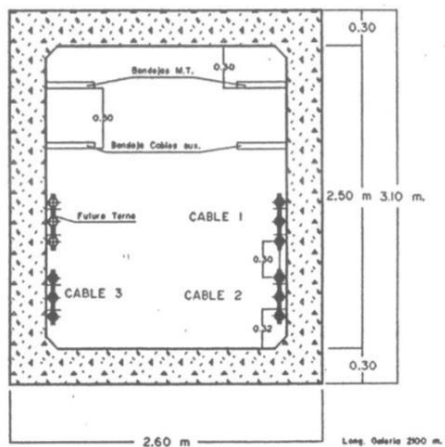


Fig. 1. Undergroun utility tunnel profile.

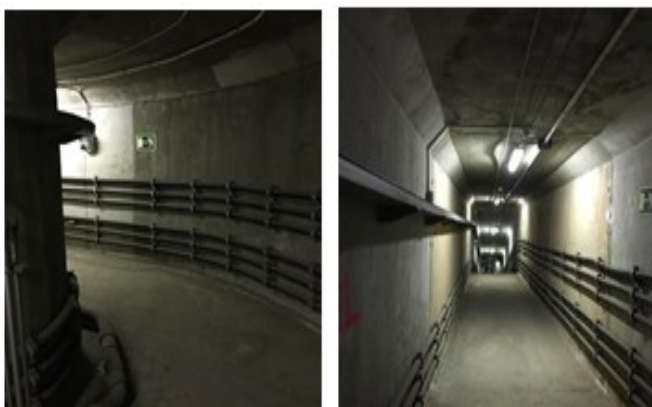


Fig. 2. Some photos of the underground utility tunnel route.

Currently, the network is made up of 27 nodes distributed along 2.4 km, which is the total length of the tunnel. Each node

commercial IoT radios. In addition, this process is required to achieve a multi-hop implem

works in the following way: firstly, the node takes measurements and send them, then enroutes the packets sent by other nodes to the network hub, and finally synchronizes with the sleep times established by the network. The system is divided into three sections, each section is made up of 9 DigiMesh sensor nodes and a gateway. Figure 3 illustrates the block diagram of a section of the implemented sensor network.

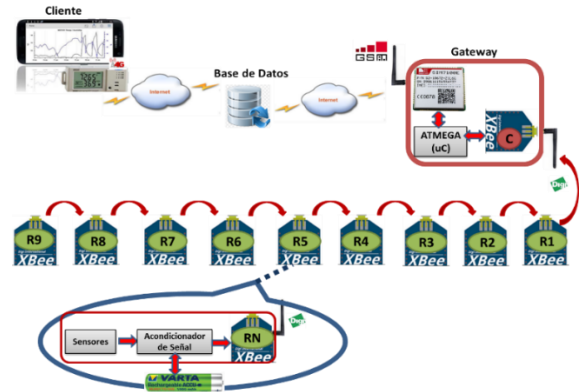


Fig. 3. General diagram of installed WSN

Most of the DigiMesh sensor nodes have been located taking into account the need of measuring environmental conditions in conflicting areas such as junction chambers or water pumps; the rest are located in less conflictive areas and all of them form the DigiMesh network. The gateways have been located at the tunnel entrances to take advantage of the coverage of the mobile network operator. Figure 4 shows the location in meters of each node following the route from the beginning of the tunnel.

SECTION 1	
GSM 1	2150 m
Node 1	2142 m
Node 2	2040 m
Node 3	1950 m
Node 4	1840 m
Node 5	1720 m
Node 6	1670 m
Node 7	1640 m
Node 8	1580 m
Node 9	1490 m
SECTION 2	
Node 10	1420 m
Node 11	1330 m
Node 12	1230 m
Node 13	1110 m
Node 14	1060 m
GSM 2	1050 m
Node 15	990 m
Node 16	910 m
Node 17	830 m
Node 18	780 m
SECTION 3	
Node 1	700 m
Node 2	570 m
Node 3	550 m
Node 4	480 m
Node 5	390 m
Node 6	280 m
Node 7	150 m
Node 8	90 m
Node 9	20 m
GSM 3	0 m

Fig. 4. Location of the sensor nodes installed in the underground utility tunnel

The distribution of the sensor nodes has been based on the simulation of the Wireless InSite® program [4], where it is possible to configure the environment taking into consideration relevant parameters of mobile communications such as reflection and propagation coefficients of the materials, antennas, losses, transmitter and receiver, etc. We have applied the parameters so the transmission conditions that may be found in an underground utility tunnel are reproduced, meaning this that the parameters have been

configured in order to reflect the worst possible transmission conditions. chosen the worst transmission case that can be found inside a tunnel as a 90 degree curve. (figure 5)

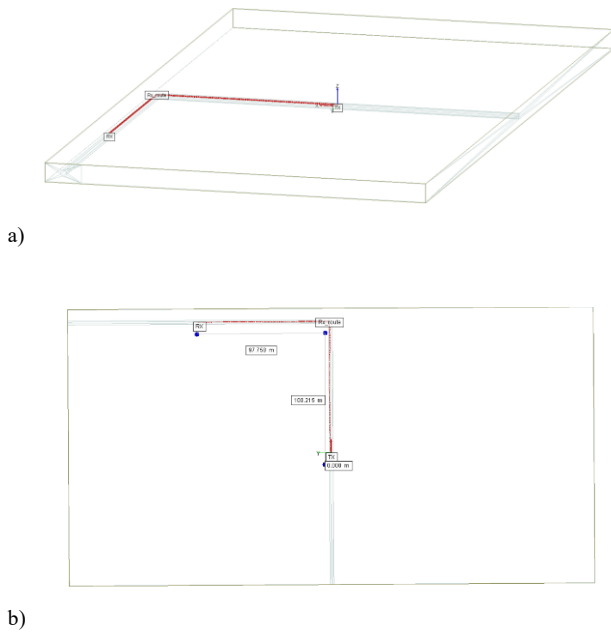


Fig. 5. 90° curve Simulation Scenario. a) 3D design. b) 2D design.

Some simulation results, such as received power and gain versus distance, are illustrated in figure 6 below.

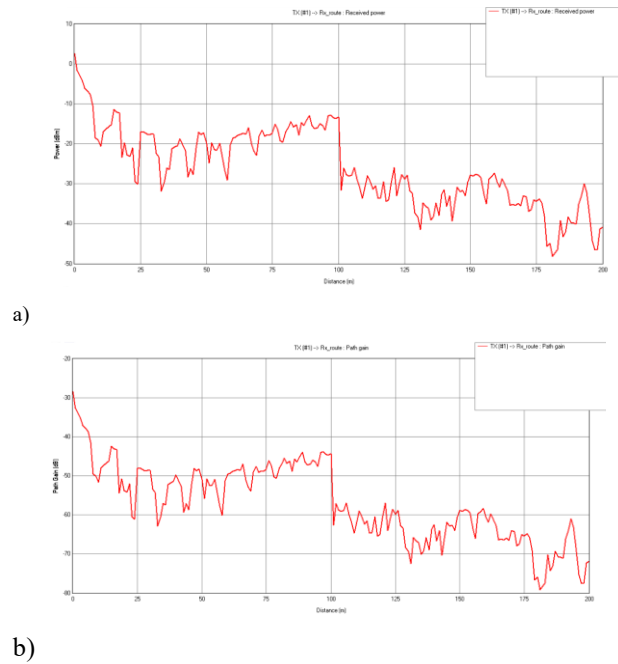


Fig. 6. 90° curve Simulation Scenario. a) Received signal power. b) gain

In the graph of Fig. 6, an attenuation of 19 dBm is observed in the tunnel curve.

III. SENSOR NODE OPTIMIZATION USING ULTRA LOW POWER

The proposal described below aims to reduce energy consumption in the sensor nodes of the system developed in the underground tunnel. The network configuration designed and installed in the tunnel has a line topology, where the nodes have serial communication. Our proposal, unlike other solutions, provides an improvement in range, sensitivity and energy performance, as well as hardware reduction. This can be achieved by eliminating WuTx in WSN sensor nodes in a difficult RF propagation environment, such as an underground electrical utility tun

nel. The designed system could be adaptable to any low consumption WSN network.

In figure 7, the essential block diagram of the proposed sensor node is shown.

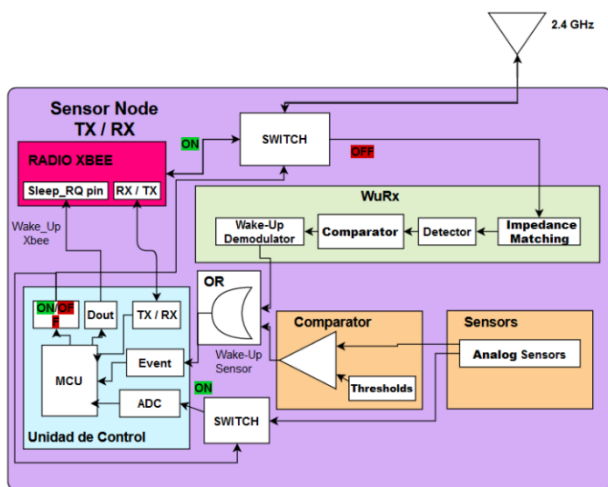


Fig. 7. General block diagram of the proposed sensor node.

Each sensor node proposed in the WSN works as a multi-hop node, both for wake-up signals and data signals. However, first all sensor nodes send a wake-up signal to wake up other nodes in case of detecting anomalies, comparing the sensor measurements of the node with the thresholds set by the client or the reception of a transmitted wake-up signal by another sensor node. In other words, the sensor node wakes up from sleep mode through a wake-up signal received in its WuRx or an anomaly detected by the average measurements of the sensor node. Therefore, two scenarios are contemplated in this design due to the need to incorporate the detection of anomalies for the activation of a sensor node. These are the following:

Activation due to an anomaly: It is a signal to activate the MCU, generated from a continuous comparison of the analog measurements of the sensors integrated in the proposed sensor node with a fixed threshold. Once the MCU is activated, the analog measurements are processed passing through the ADC converter and the antenna is selected to work with the main radio through the switches, then the MCU activates the main radio, Xbee, to send a wake-up signal to activate the next one within range. Then, the Xbee radio waits for an ACK signal and once the signal is received, it sends the data. Finally, it returns to the initial state until listening to a new wake-up signal by WuRx, so that energy is saved.

Activation with a wake up signal: Wake-up signal activates the MCU generated by the WuRx radio alarm clock. When activating the MCU, the sensor and comparator blocks are deactivated and the main Xbee radio is activated and sends the ACK signal, waiting at the same time to receive the measurement data. Then a wake-up signal is sent to the next node to wake it up, and so on until the hub is reached. It should be noted that in this case the sensor node behaves like a data router equipped with the WuRx radio alarm clock.

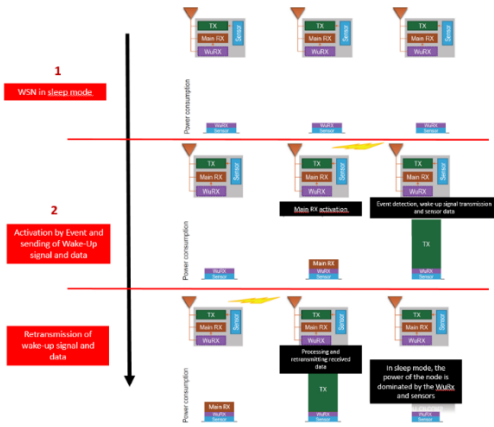


Fig. 8. Different way to activate the transmission node.

The main difficulty of the multi-hop enablement techniques is the complexity of the WuRx and WuTx hardware previously described. However, this level of complexity is necessary in order to guarantee a proper consumption [5]. In this system, the modulation technique used in the transmission of the wake-up signal is the same as the one used in the WuTx and WuRx modes. In this way, the main radio of an Xbee module can be used.

IV. WAKE-UP SIGNAL MODULATION

The proposed radio frequency module, Xbee, is a Digi implementation based on the Zigbee protocol under the IEEE 802.15.4 standard. The IEEE 802.15.4 standard specifies the type of modulation in its physical layer (PHY). In the case of the proposed Xbee of the 2.4Ghz band, the modulation used is the quadrature phase shift O-QPSK (Offset Quadrature Phase Shift Keying) with a bit rate of 250 kb/s, a symbol rate of 62.5 Ksymbol/s and the total number of symbols, 16-array orthogonal.

The wake-up signal transmitted by Xbee is modulated in OQPSK; in theory it would have to be demodulated at the receiver in OQPSK. In this development, OOK (On-Off Keying) digital amplitude demodulation is used for the correct detection of wake-up signals. So, you have to look for an OQPSK modulated signal capable of demodulating it with OOK technology.

OOK modulation is the simplest and most common form of ASK modulation. Its operation can be understood as that of a switch that turns the carrier signal on or off, in such a way that the presence of a carrier indicates a binary 1 and its absence a binary 0. The modulated signal follows the following equation:

$$S(t) \begin{cases} A \sin(2\pi ft) & \text{"1" binary} \\ 0 & \text{"0" binary} \end{cases} \quad (1)$$

An example of OOK modulation is illustrated in Figure 9.

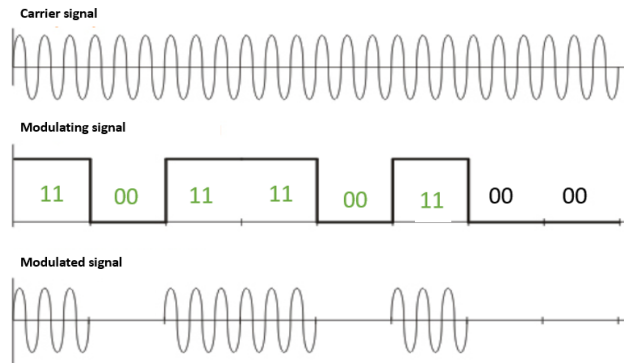


Fig. 9. Amplitude Shift Modulation.

The OQPSK modulated wake-up signal is demodulated with an OOK receiver. The idea is based on taking advantage of the operating characteristics of OOK and OQPSK modulations to build a wake-up signal with the collaboration of a micro and hardware, capable of modulating it again using OOK technology. With OQPSK modulation, a sinusoidal signal can be sent in phase with a fixed amplitude over time, simulating a sequence of high bits "1" following equation 2.

$$S(t) = A \sin(2\pi ft + \frac{\theta}{4}) \quad (2)$$

In this case, the phase shift ($\theta/4=0$) can be neglected because it is not significant when constructing the signal that could be demodulated with the OOK technology. With which, it has been possible to build a carrier signal (equation 3), sending a sequence of binary "1" during a known time (tx).

$$S(t) = \int_0^{tx} A \sin(2\pi ft) dt \quad (3)$$

The next task is based on the OOK modulation of a bit frame to get the wake-up signal using a microprocessor and a time-controlled switch. The construction of a modulated signal in OOK of a frame of bits, is carried out by cutting the transmission of the carrier signal $S(t)$ during the time it takes to send the binary "0"; for this, a switch controlled by the microprocessor is used. An example of a binary frame is shown in Figure 10.

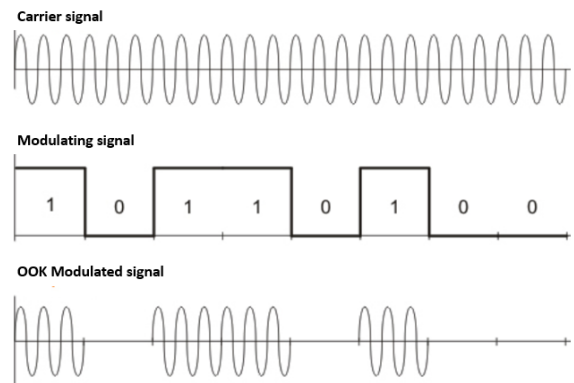


Fig. 10. OOK modulation example with two bits

In figure 11 you can see the block diagram of the WuTx including the controlled switch.

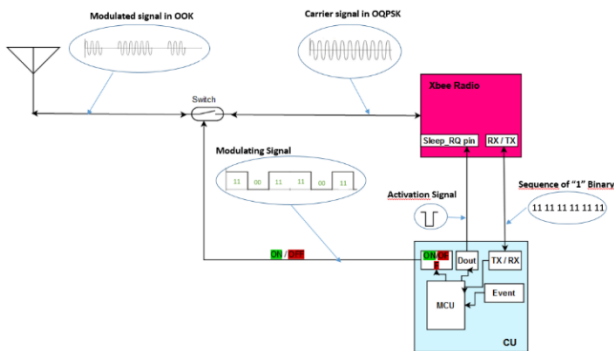


Fig. 11. WuTx block diagram

V. WAKE-UP SIGNAL DEMODULATION

The WuRx receiver design is based on the decoding of the wake-up signal. A wake-up signal is generated by detecting a data signal (wake-up) on a carrier frequency. The communication node has two receive paths, one to process a wake-up signal and one to communicate with the main radio of the node. The receive path is activated depending on the antenna switch. This switch is controlled through an output port of the microcontroller used in the design. Before entering low power mode, the controller configures the switch so that all incoming signals are routed to the WuRx circuit. After impedance matching, rectification, and low-pass filtering, only the envelope signal remains, which is connected to the input of the demodulation block. In the event of a valid wakeup signal and a positive correlation of the sent address with an internally saved bit stream, the WuRx receiver interrupts the microcontroller from its sleep mode. The Figure 12 shows the WuRx receiver block diagram.

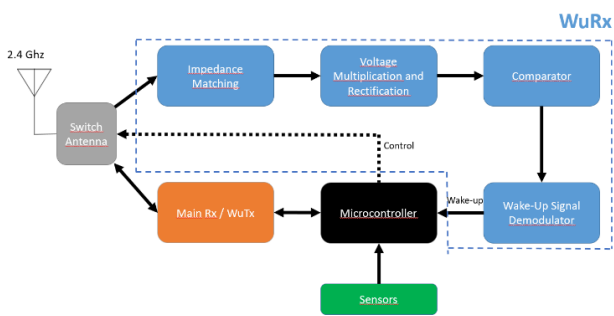


Fig. 12. WuRx receiver block diagram.

The objective of the receiver design is to minimize the energy consumed that is related to the frequency spectrum to be used in the transmission. If low frequencies are used, lower consumption is achieved in the receiver circuit, but larger antennas are required at low frequencies. Therefore, to compensate for this practical situation, a carrier signal of 2.4 GHz with an OOK modulation of 125 KHz is chosen. The 2.4 GHz carrier is then turned on and off so that the resulting envelope represents a 125 kHz square signal. In addition, said signal contains additional modulated address information. Figure 13 shows OOK carrier modulation as described.

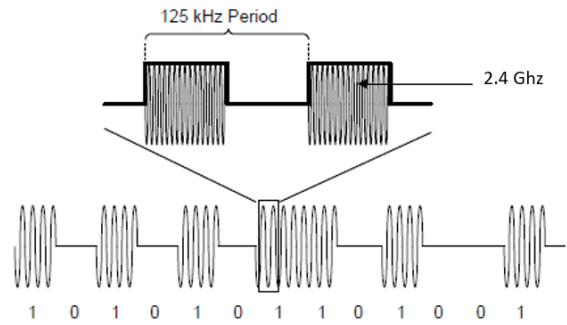


Fig. 13. 2.4GHz carrier and 125 KHz OOK modulation

The WuRx circuit includes the blocks shown in Figure 14.

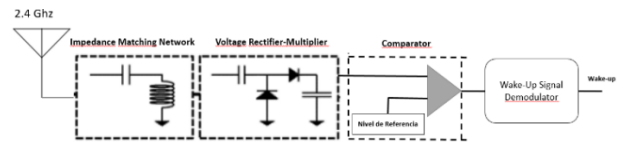


Fig. 14. WuRx circuit blocks. Details.

The Voltage Multiplication and Rectification Block [6] is simply an envelope detector that demodulates the 2.4 GHz wake-up signal sent by WuTx to extract the 125 KHz signal. Since the RF energy at the rectifier input is very weak, a multi-stage rectifier is designed as shown in Figure 15.

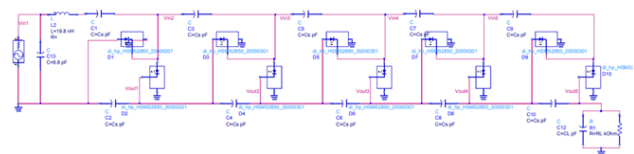


Fig. 15. RF signal collection circuit, 5 stages

If an ideal collection system is considered and it work in open circuit, that is, with an infinite load, the output voltage V_{out} can be expressed as equation 3 where V_p is the peak voltage at the input. With this equation, for a 5-stage multiplier, V_{out} is ideally 10 times the peak voltage.

$$V_{out} = V_0 \times n = 2V_p \times n \quad (4)$$

The Impedance Matching Network Block is formed by an inductive-capacitive filter whose objective is to transfer the maximum power between the antenna and the rest of the circuit.

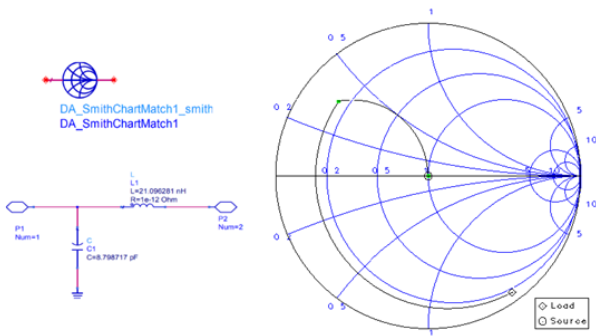


Fig. 16. 5-stage multiplier adaptation simulation.

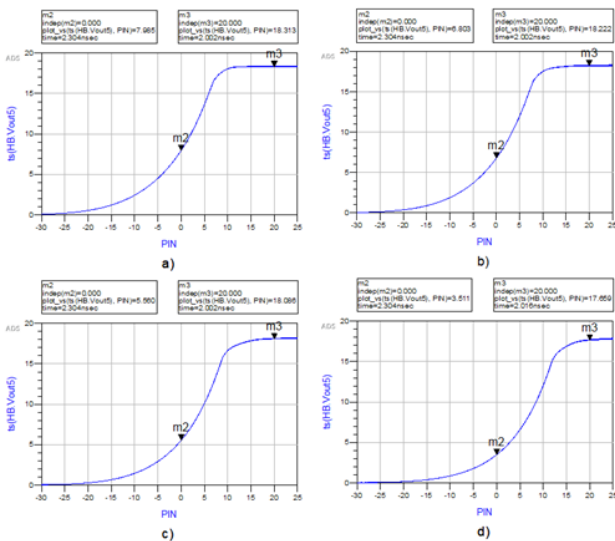


Fig. 17. Plot $V_{out}(V)$ vs $P_{in}(dbm)$, 5 stages with coupling circuit. a) $R_L=300K$, b) $R_L=100K$, c) $R_L=50K$, d) $R_L=20K$

The Voltage Compare Block which uses a comparator device to compare the 2.4 Ghz envelope output signal to a configurable reference level. In addition, it serves to protect the next circuit block against a voltage surge generated by the voltage multiplication and rectification block.

V. MONITORING SOFTWARE

For the treatment of the information, a web application has been designed and developed to monitor the measurements in real time, which is accessible through a browser in the form of a histogram and a table. Figure 18 illustrates the measurements of all the points of the underground utility tunnel in the form of a histogram.

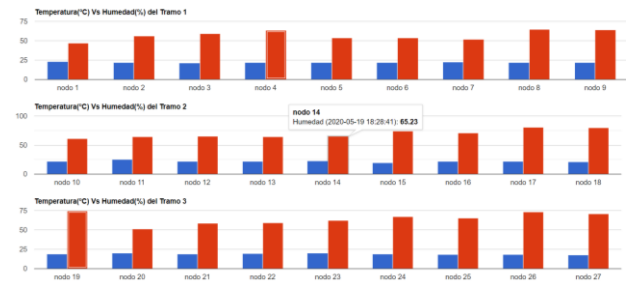


Fig. 18. Histogram of the temperature and humidity measurements of the service tunnel

In addition, scripts have been developed within the web page for the generation of alarms (email and SMS) in case of exceeding the thresholds and the configuration of sensor thresholds (temperature, humidity and temperature increase) and registration of new accounts. of user or administrator type through authentication for, for example, the reception of alarm notifications.

VI. FEATURED RESULTS

The maximum energy consumption in the transmitting node is generated in the RF emission period. This consumption is much higher than that produced in the sensor block and the microcontroller. The estimation of the autonomy of the device in different scenarios and a comparison with other devices is illustrated below (Table 1). The best performance is due to the use of the developed Wake-up techniques.

Table 1. ESTIMATION OF THE AUTONOMY OF THE DEVICE IN DIFFERENT SCENARIOS

	Battery (mAh)	Scope (m)	Sleep time (h)	Enable time (s)	Autonomy(dies)
MH-REACH-Mote [7]	3600	9,1	4	3	413
prototype (a)	3600	100	4	3	631,34 (+53%)
prototype (b)	19000	100	24	2	3492 (+750%)
Other Prototypes (State of art)	3600	100	4	3	265,07 (-36%)

VII. CONCLUSIONS

A monitoring system has been developed using very low consumption communication nodes. Its operation has been installed and verified in an underground utility tunnel including medium and high voltage lines. This scenario requires a line network topology that makes the design difficult. Performance has been increased by lowering the power consumption required by each transmission node; Wake-up techniques are used to achieve this. There is a passive auxiliary radio in all nodes, WuRx, capable of receiving the wake-up signals, and also another radio, WuTx, to transmit another wake-up signal in order to wake up the other nodes. To take advantage of a previous system, the main radio of the sensor node has been redesigned so that it transmits the Wake-up signals. A passive energy capture system has been incorporated, taking advantage of radio frequency to activate the sensor node. Although the transmission range between nodes can be improved, the tests are positive. As a line of future research, it is desired to deepen the study of the RF Wake-up system, which is essential to reduce energy consumption. It will also be studied how to improve the renewable energy system constituted by the capture of tuned radiofrequency to enable the transmission node that is part of the IoT structure.

REFERENCES

- [1] <https://www.digi.com/blog/post/what-are-the-differences-between-digimesh-and-zigb>
- [2] <https://users.eecs.northwestern.edu/~peters/references/ZigbtbeeIEEE802.pdf>
- [3] https://www.digi.com/resources/documentation/digi_docs/pdfs/90000982.pdf
- [4] <https://www.remcom.com/wireless-insite-em-propagation-software>
- [5] E. Lopez-Aguilera, M. Hussein, M. Cervia, J. Paradells, and A. Calveras, "Design and Implementation of a Wake-Up Radio Receiver for Fast 250 kb/s Bit Rate," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 6, pp. 1537–1540, 2019.
- [6] Yi Li, Sheng ming Huang and Quanzhen Ducean, "A low input voltage charge pump for energy harvesting", *Journal of Physics: Conference Series*, vol. 1550, 2020.
- [7] L. Chen, J. Warner, W. Heinzelman, and I. Demirkol, "MH-REACH-Mote: Supporting multi-hop passive radio wake-up for wireless sensor networks," *IEEE Int. Conf. Commun.*, vol. 2015-Septe, pp. 6512–6518, 2015.

Design of 24 GHz ISM Band Microstrip Patch Antenna for 5G Communication

Debalina Mollik
Department of CoE
American International
University-Bangladesh (AIUB)
Dhaka, Bangladesh
promidebolina@gmail.com

Rima Islam
Department of EEE
American International
University-Bangladesh (AIUB)
Dhaka, Bangladesh
rima.islam012018@gmail.com

Afrin Binte Anwar
Department of EEE
American International
University-Bangladesh (AIUB)
Dhaka, Bangladesh
afrinanwar144@gmail.com

Prodip Kumar Saha Purnendu
Department of EEE
American International
University-Bangladesh (AIUB)
Dhaka, Bangladesh
purnendu7927@gmail.com

Md. Azad Hossen Shanto
Department of EEE
American International
University-Bangladesh (AIUB)
Dhaka, Bangladesh
shanto01749@gmail.com

Mohiuddin Ahmad
Dept. of Electrical and Electronic
Engineering
Khulna University of Engineering
& Technology
Khulna-9203, Bangladesh.
mohiuddin.ahmad@gmail.com

Abstract— With the growth of world communication, wireless communication has become one of the most prominent sectors. The design and simulation result of the 24 GHz (ISM band) antenna has been developed in this article. Many publications have addressed this research and developed antennas utilizing arrays. However, the proposed design provides a novel model of a small size patch antenna that is easier to build and has improved return loss, efficiency, and gain than previous research studies. The recommended patch antenna design is $18\text{mm} \times 2.451\text{mm} \times 0.257\text{mm}$. The substrate material for this antenna is Roger RT5880 (lossy), which has a permittivity of 2.2. According to the simulation results, this design has a return loss of -28.96 dB , a VSWR (Voltage Standing Wave Ratio) of 1.08497, a gain of 5.152 dB, and an antenna efficiency of 67.026%. The patch antenna is expected to function effectively and may be used for wireless communication based on simulation results.

Keywords—CST, 5G, Gain, ISM band, Radiation efficiency, Surface current, VSWR.

I. INTRODUCTION

Wireless communication technology is getting developed with the enhancement of 5th generation communication as it has the benefit of having large bandwidth, and a high-speed data rate. To support this, it is required to develop and design more improved characteristics of antenna such as high gain, beam width, VSWR, radiation efficiency, and return loss and needed to balance them with antenna size, cost, and design [1-3]. In addition, the Microstrip antenna contains low characteristics, light weight, moderate cost, light volume, low fabrication expense, smaller size, and flexible design. [4]. The proposed antenna is designed at 24 GHz which is within the range of industrial as well as scientific (ISM) bands. The ISM band 24-24.25 GHz is available for worldwide application, and licensed users of this frequency include earth exploration satellite services and armature satellite radiolocation. This frequency band is quite useful in high-performance and lower-cost health care sensors, radar applications, and communications [5-7]. The

recommended Microstrip patch antenna in this article has a high gain of 5.152 dB and a significant return loss of -28.91 dB , which are required for future 5G Wireless Communication Systems. The recommended model has an efficiency of 67.026% along with a bandwidth of 0.769 GHz. Among several previously completed works, the antenna model has a strong gain and a larger return loss, implying that the suggested design work is more efficient.

This paper is organized as follows: Section II describes the theory and methodology, section III illustrates the proposed antenna design, Section IV describes the simulation and analysis, and a comparative study is given. Finally, section V concludes the paper.

II. THEORY AND METHODOLOGY

The frequency range taken for the proposed Microstrip patch antenna is from 23 GHz to 25 GHz where the operating frequency is 24.068 GHz to have better performance and this is under the ISM band. Here, Fig. 1 delineates the geometric configuration of the designed antenna in free space.

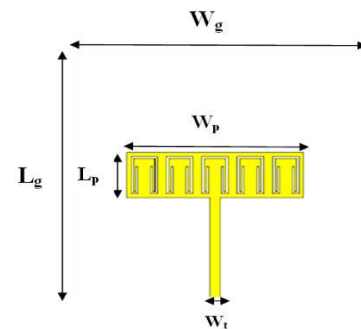


Fig. 1. Structure of the recommended microstrip patch model for free space.

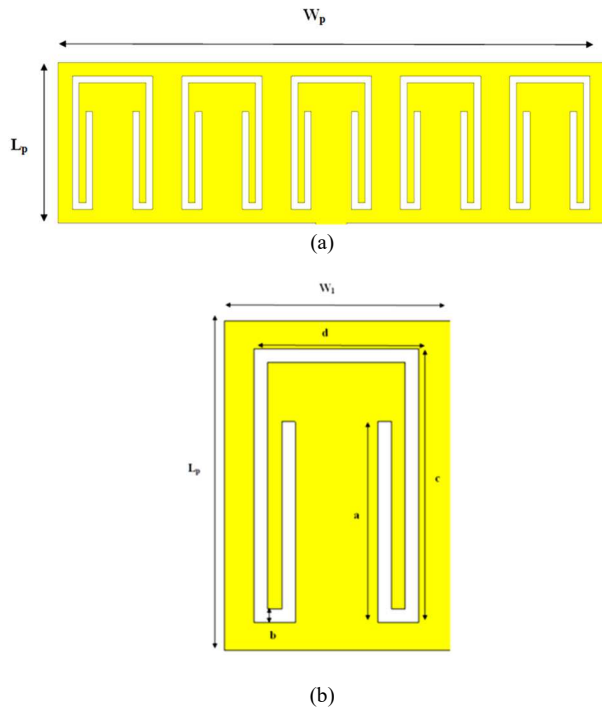


Fig. 2. (a) and (b) are the configurations of the designed antenna including slot size, depth, and feedline width in free space.

Roger RT5880 (lossy) is chosen as substrate with permittivity 2.2. Necessary formulas for the calculation of parameters are given below [8]. Patch antenna's width and length are being measured by using Eq. (1).

$$W = \frac{c_0}{2f_r \sqrt{\frac{\epsilon_r + 1}{2}}} \quad (1)$$

Here,

W= Patch's width

C = Light Velocity = 3×10^8 m/s

f = Resonant frequency

ϵ_r = Substrate's dielectric constant

The effect permittivity ϵ_{reff} and the effective length, L_{eff} are calculated by equations (2) and (3) respectively.

$$\epsilon_{reff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + 12 \frac{h}{W}\right)^{-0.5} \quad (2)$$

$$L_{eff} = \frac{c_0}{2f_r \sqrt{\epsilon_{reff}}} \quad (3)$$

Where L_{eff} = Effective length. The extension of length is given by Eq. (4).

$$\Delta L = 0.412 \frac{\left(\frac{W}{h} + 0.264\right) (\epsilon_{reff} + 0.3)}{(\epsilon_{reff} - 0.258) \left(\frac{W}{h} + 0.813\right)} \quad (4)$$

ΔL = extension length.

The length of the patch is calculated by Eq. (5) below.

$$L = L_{eff} - 2\Delta L \quad (5)$$

III. ANTENNA DESIGN

CST software is used to design the antenna. Designed antenna views from different angles such as perspective, front, and rear, left and right, and top and bottom views are shown in Figures 3 to Figure 8.

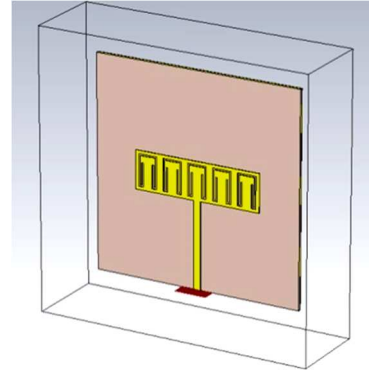


Fig. 3. Standpoint vision of the antenna.

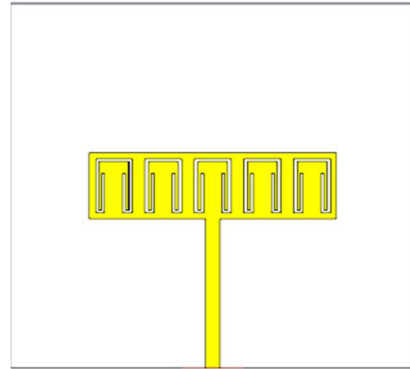


Fig. 4. Antenna's front view.

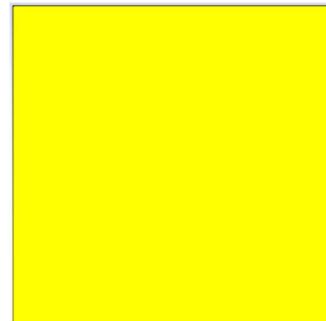


Fig. 5. Rear view of the antenna.

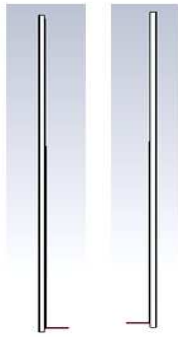


Fig. 6. Left & Right vision of designed antenna sequentially.



Fig. 7. Antenna's top view design.



Fig. 8. Recommended antenna's bottom view design.

Related specifications of the proposed antenna are shown in the following Table I.

TABLE I. THE PROPOSED MODEL'S PARAMETERS

Parameter descriptions and symbols	Measurements (mm)
Antenna Patch's width, W_p	18
Width of Single Patch, W_1	3.6
Length of Patch, L_p	2.451
Ground plane's width, W_g	20
Ground plane's length, L_g	20
Slot size, a	2.2
Slot size, b	0.15
Slot size, c	3
Slot size, d	1.8
Height of Substrate, h_s	0.257
Feedline width	0.695
Ground Thickness, h_t	0.045

IV. SIMULATION AND ANALYSIS

After designing the antenna, to run the simulation, two options can be selected. From the home section, an author can select setup solver or start simulation. For setup solver, the accuracy level has been kept at -40 dB. After that start option has been pressed. Besides, the start simulation option also can be pressed which will lead directly to the simulation process.

A. S-Parameter

S-parameter or return loss gives information about the amount of reflected power caused by impedance discontinuity and the acceptance range for return loss of an antenna is greater than -10 dB [9]. The value of return loss of this antenna is -28.96 dB which is quite good for a Microstrip antenna at the operating frequency of 24.07 GHz shown in Fig. 9. The bandwidth is frequency ranges within it, and the antenna radiates. Return loss is -10 dB at 23.4 GHz and -10dB at 24.169 GHz, as indicated

in Fig. 10. To determine the antenna's bandwidth, an axis marker is placed on both frequencies. The maximum frequency is 24.169 GHz, while the lowest frequency is 23.4 GHz. The bandwidth has been determined as the difference between the highest and lowest frequencies, which is 0.769 GHz.

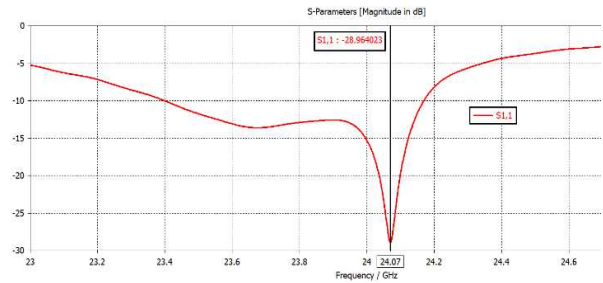


Fig. 9. $S_{1,1}$ parameter in free space.

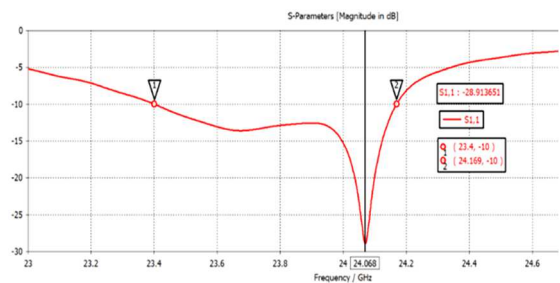


Fig. 10. Bandwidth in free space.

B. VSWR

The definition of voltage standing wave ratio-VSWR is the determination of the amount of matched impedance between antenna and transmission line. The least acceptable value is 1 for the VSWR [10]. For the proposed antenna the VSWR value is 1.08497, supporting the value shown in Fig. 11.

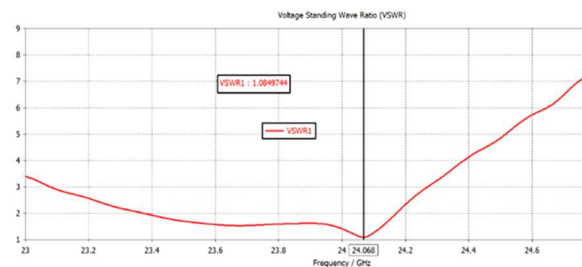


Fig. 11. VSWR vs. the Frequency in port 1.

C. Gain & Directivity

Gain and directivity are one of the most important parameters for antenna performance. The antenna's gain determines the quantity of power transmitted at the peak radiation direction [11]. Directivity is the term that explains about directivity of antenna radiation patterns [12]. Gain for the designed antenna is 5.152 dB and directivity 7.699 dBi shown in Figures 12-16.

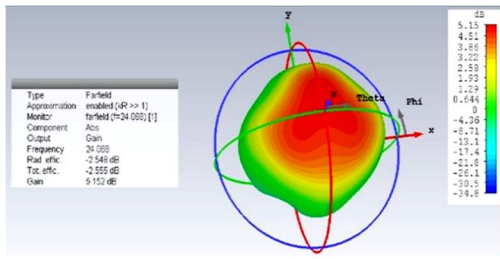


Fig. 12. Farfield gain plot of the designed antenna at 24.068 GHz (3D).

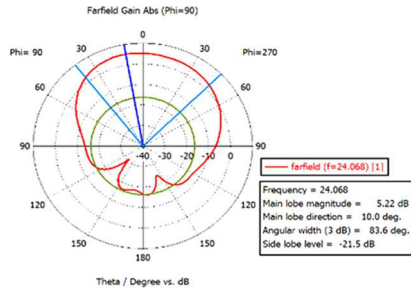


Fig. 13. Farfield gain plot of the antenna at 24.068 GHz (2D).

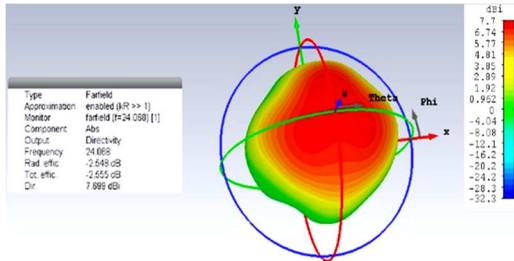


Fig. 14. Radiation-Pattern in three dimensions at 24.068 GHz (Directivity).

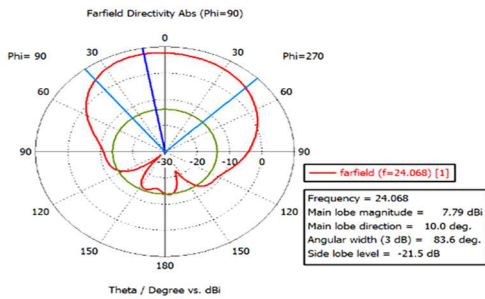


Fig. 15. Radiation pattern in 2D at 24.068 GHz (Directivity).

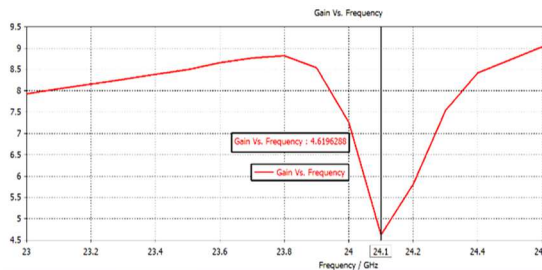


Fig. 16. Gain vs. Frequency Curve of the antenna in free space.

D. Overall and Radiation Efficiency

From Fig.17, the overall efficiency is -2.57 dB, while the radiation efficiency is -2.568 dB. The ratio of gain to directivity can be used to calculate antenna efficiency.

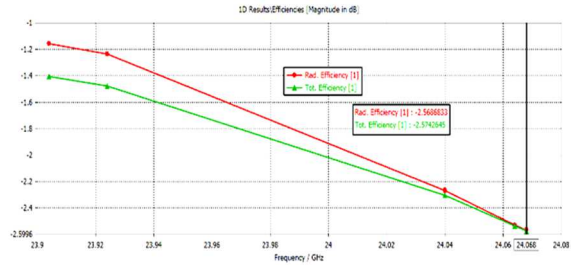


Fig. 17. Radiation and Total Efficiency Curve of the designed antenna in free space.

E. Surface Current

The definition of surface current is the electric current induced in a metallic antenna due to an enforced electromagnetic field that drives charges at its surroundings [13]. Here, the surface current has shown in fig.18 which is 1360 A/m.

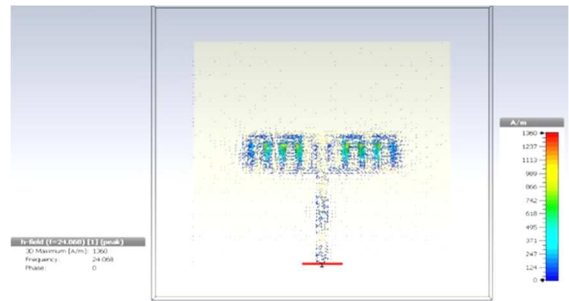


Fig. 18. Surface current.

F. Smith Chart

Smith's chart illustrates the relationship between transmission line impedance and antenna via frequency. This helps to understand more about changes in transmission line impedances [14]. Smith chart for this antenna is 49.8 ohm shown in Fig. 19. Impedance mismatches are noticed as a result of shape complexity. Finally, Table II gives the parameters summary of the proposed designed antenna.

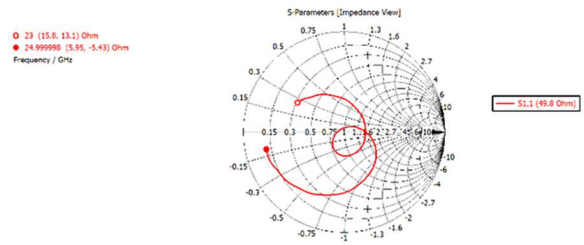


Fig. 19. Smith chart of the antenna (Impedance View).

TABLE II. OBTAINED PARAMETER'S SUMMARY OF THE DESIGNED ANTENNA.

Specifications of Antenna	Values
$S_{1,1}$	-28.96 dB
Bandwidth	0.769 GHz
VSWR	1.08497
Gain	5.152 dBi
Directivity	7.699 dBi
Efficiency	67.026%
Surface Current	1360 A/m
Radiation Efficiency	-2.569 dB
Total Efficiency	-2.574 dB

V. COMPARATIVE STUDY

Table III summarizes and compares different parameters of the antenna. Different antenna characteristics are compared and examined with other research articles, including $S_{1,1}$, operating frequency, bandwidth, band, gain, and efficiency. Circular Polarization Antenna Array is designed for 24 GHz wireless communication and has 42% efficiency and a high gain of 20dBi, according to ref. [15]. The array antenna for 24 GHz has been presented in the majority of research articles. In ref. [6], a Microstrip patch antenna was built for 24GHz, although it has a return loss of -22dB.

TABLE III. COMPARATIVE ANALYSIS WITH OTHER RESEARCH PAPERS

Ref.	$S_{1,1}$	Resonant Frequency	Bandwidth	Band	Gain	Efficiency
[15]	-19 dB	24.5 GHz	2 GHz	-	20 dBi	42%
[6]	-22.1 dB	24.2 GHz	1.2 GHz	ISM	5.95 dBi	-
[7]	-25dB	2.4Ghz	160 MHz	ISM	-	-
[16]	-24 dB	24 GHz	-	-	13.5 dBi	-
[17]	-23dB	24.15 GHz	(24.07-24.27) GHz	ISM	18 dBi	-
This work	-28.96 dB	24.068 GHz	0.769 GHz	ISM	5.220 dB	67.02 6%

In contrast to other research papers, this article has a better return loss, a better gain, and a higher efficiency. An array antenna's design is complicated, but a Microstrip patch antenna's design is simple and easy to fabricate. This research has resulted in improved simulated findings for the 24GHz (ISM band), which will be advantageous for wireless communication and medical applications.

TABLE IV. COMPARISON OF OBTAINED RESULTS WITH RECENT RESEARCH PAPERS.

Ref.	Architecture of ground (length × width)	Substrate height (mm)	Substrate materials	Dielectric permittivity, ϵ_r	Software
[15]	30mm×30mm	-	Roggers/duroid 5880	2.2	HFSS
[6]	-	0.787	Roggers/duroid 5880	2.2	HFSS
[7]	30mm×40mm	1.63	FR04	4.4	HFSS
[16]	65mm×23mm	0.813	Rogers RO4003C	3.38	HFSS
[17]	-	0.508	Teflon - fiberglass	2.2	3D electromagnetic simulator
This work	20mm×20mm	0.257	Roggers 5880 (lossy)	2.2	CST

Table IV evaluates the ground architectures, substrate heights, materials, dielectric permittivity, and software applied in antenna design. The majority of research articles used HFSS software to build antennas, while this work has used CST software to simulate and design [15, 6]. Roggers/duroid 5880 was chosen as a substrate material in the majority of the research publications. Different substrate materials, such as Rogers RO4003C [16] and Teflon –fiberglass [17] have been implemented in numerous research publications. A very thin substrate is recommended for a low-profile construction and lightweight antenna; this paper has also proposed an antenna with thin substrate materials and a lower substrate height. [15] However, recent investigations have found that different substrate materials provide different results. As a consequence of the aforementioned comparisons, it can be determined that the projected antenna would be a preferable alternative for the wireless network to earlier versions.

VI. CONCLUSION

The proposed microstrip patch antenna can be considered to be compatible with forthcoming 5G wireless communication and medical applications. The recommended antenna's working frequency is 24.07GHz, with a return loss of -28.91 dB, a bandwidth of 0.769 GHz, a gain of 5.152 dB, directivity of 7.699 dBi, surface current of 1360 A/m, and 67.026% efficiency. The operational frequency of 24.07GHz is part of the ISM band, which spans 24 GHz to 24.25 GHz. Most of the researchers have focused on array antennas, but minorities have focused on Microstrip patch antennas, which is a novel idea in the 5G communication system. However, the proposed antenna in this study has a basic design that is easy to fabricate and improves overall simulation results. After analyzing all of the findings and simulations, it can be concluded that this study can

be used as a model for future communication and other research works.

REFERENCES

- [1] Q. Wang, N. Mu, L. L. Wang, S. Safavi-Naeini, and J. P. Liu, "5G MIMO Conformal Microstrip Antenna Design," *Wireless Communications and Mobile Computing*, 17-Dec-2017. [Online]. Available: <https://www.hindawi.com/journals/wcmc/2017/7616825/>. [Accessed: 14-Jul-2021].
- [2] R. Przesmycki, M. Bugaj, and L. Nowosielski, "Broadband Microstrip Antenna for 5G Wireless Systems Operating at 28 GHz," *Electronics*, vol. 10, no. 1, p. 1, 2020.
- [3] D. Imran, M. M. Farooqi, M. I. Khattak, Z. Ullah, M. I. Khan, M. A. Khattak, and H. Dar, "Millimeter wave microstrip patch antenna for 5G mobile communication," 2018 International Conference on Engineering and Emerging Technologies (ICEET), 2018.
- [4] Y. Jia, Y. Liu, and Y. Zhang, "A 24 GHz microstrip antenna array with large space and narrow beamwidth," *Microwave and Optical Technology Letters*, vol. 62, no. 4, pp. 1615–1620, 2019.
- [5] Y. Yan, Y. B. Karandikar, S. E. Gunnarsson and H. Zirath, "24 GHz balanced self-oscillating mixer with integrated patch antenna array," 2011 41st European Microwave Conference, 2011, pp. 404-407, doi: 10.23919/EuMC.2011.6101836.
- [6] G. Christina, A. Rajeswari, and S. Mathivanan, "Real Time Analysis of a 24 GHz Planar Microstrip Antenna for Vehicular Communications," *Wireless Personal Communications*, vol. 97, no. 1, pp. 1129–1139, 2017.
- [7] "Design & Analysis of Microstrip Antenna in Spiral Structure for Medical Application", *International Research Journal of Engineering and Technology (IRJET)*, vol. 06, no. 052019, pp. 1663-1666, 2019. Available: <https://www.irjet.net/archives/V6/i5/IRJET-V6I5332.pdf>. [Accessed 14 July 2021].
- [8] A. Balanis, "Antenna theory: a review," *Proceedings of the IEEE*, vol. 80, no. 1, pp. 7–23, 1992.
- [9] P. Bevelacqua, "Welcome to Antenna-Theory.com!", *The Antenna Theory Website*. [Online]. Available: <https://www.antenna-theory.com/>. [Accessed: 14-Jul-2021].
- [10] P. Bevelacqua, "VSWR (Voltage Standing Wave Ratio)," *VSWR*. [Online]. Available: <https://www.antenna-theory.com/m/definitions/vswr.php>. [Accessed: 14-Jul-2021].
- [11] P. Bevelacqua, *Antenna Gain*. [Online]. Available: <https://www.antenna-theory.com/basics/gain.php>. [Accessed: 14-Jul-2021].
- [12] P. Bevelacqua, "Directivity," *Antenna*. [Online]. Available: <https://www.antenna-theory.com/basics/directivity.php>. [Accessed: 14-Jul-2021].
- [13] P. Patel, "What is surface current?," *ResearchGate*, 23-Jun-2014. [Online]. Available: https://www.researchgate.net/post/What_is_surface_current. [Accessed: 14-Jul-2021].
- [14] "The Smith Chart," *Smith Charts*. [Online]. Available: <https://www.antenna-theory.com/tutorial/smith/chart.php>. [Accessed: 14-Jul-2021].
- [15] S. Ladan, A. B. Guntupalli and K. Wu, "A High-Efficiency 24 GHz Rectenna Development Towards Millimeter-Wave Energy Harvesting and Wireless Power Transmission," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 12, pp. 3358-3366, Dec. 2014, doi: 10.1109/TCSI.2014.2338616.
- [16] C. -A. Yu, K. -S. Chin and R. Lu, "24-GHz Wide-Beam Patch Antenna Array Laterally Loaded With Parasitic Strips," 2019 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), 2019, pp. 1-3, doi: 10.1109/CSQRWC.2019.8799203.
- [17] I. Radnovic, B. Jekanovic and A. Boryszenko, "Circularly Polarized Patch Antenna Array at 24 GHz for Radar Applications," 2018 26th Telecommunications Forum (TELFOR), 2018, pp. 1-4, doi: 10.1109/TELFOR.2018.8611820.

Detection of Corona Virus Infection using Convolutional Neural Network

Al Sameera

Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science Pilani, Dubai
Campus, UAE
E-mail:nabash.sameera@gmail.com

Vilas H Gaidhane

Department of Electrical and Electronics Engineering and APPCAIR
Birla Institute of Technology and Science Pilani,
Dubai Campus, UAE
E-mail:vilasgd612@gmail.com

Abstract—Coronavirus 2019, also known as COVID-19, has recently had a negative influence on public health and human lives. Since the second world war, this disastrous consequence has changed human experience by initiating an increasingly more devastating and unexpected health calamity. The world's condition has become a catastrophic epidemic due to uncontrolled infectious properties within society. Therefore, the initial stage and accurate identification of the virus may be a good strategy for tracking and suppressing the illness from spreading due to the lack of medication. During the pandemic, computed tomography (CT) imaging has been extensively used to detect the percentage of infection. Artificial intelligence-assisted CT- image analysis could be a better option which can be achieved using a convolutional neural network (CNN). This is one of the prominent modes that can be effectively used in such applications. In this paper, an artificial-intelligence-based approach has been presented to investigate the coronavirus infection in the human body. The various experiments are carried out on the freely available dataset. It has been observed that the results are better indicating adequate performance for prediction.

Keywords— *CT scan images; Convolutional neural network; Coronavirus; Dense-Net; Image Processing.*

I. INTRODUCTION

Recently, the coronavirus disease outbreak has already been identified as a significant worrying patient safety diversion. This virus is first discovered in late 2019 in Wuhan, China, and quickly spread to over 200 nations, prompting the Department of Health (WHO) to declare a pandemic emergency. Medical professionals, as well as regulations, were unable to control the spreading of the virus in the society which results in the quick death of individuals, due to significant infectious qualities among people in intimate contact. Since subsequent illnesses in communities are caused by close personal dealings, it is critical to quickly isolate and characterize the infectious disease and institute a social lockdown. As a result, early diagnosis of virus occurrences is critical for help in managing efforts to reduce infectious hazards and spreading, organize clinical treatment, and coordinate prompt care assistance. This could be crucial in improving national healthcare. This will have a direct impact on the virus's removal from the planet.

During the pandemic time, due to the lack of particular treatments and vaccinations, it was critical to identify the sick person so that prompt isolation can be taken. The real-time polymerase chain reaction analysis is widely adopted for the

identification of the virus in healthcare. Nevertheless, a low incidence of positive RT-PCR results and findings could be attributed to any evidence collecting and transportation bottlenecks, which are both time-consuming procedures. The patients with severe situations (those in the emergency ward (ICU)) may be missed by this method. An additional disadvantage of the RT-PCR evaluation methods is that traditional PCR tests take more time to detect disease. As this COVID-19 virus has a powerful transmitted character, an infected individual could become a carrier and pass the infection to other healthy, common citizens or medical practitioners. On the other hand, emerging economies have inadequate technical capabilities and a lack of healthcare technicians and specialists. Although there are few resources available to combat a pandemic, the COVID-19 performance indicators are a potential solution that reflects real requirements. Imaging techniques are a conventional technique for finding pneumonia that gives a more reliable judgement than RT-PCR. In this regard, CT image classification become a potential alternative approach for detecting viruses, particularly in patients with acute illnesses. According to the research, the accuracy of RT-PCR is lower than the accuracy of CT [1].

The major goal of this study is to develop an accurate method for identifying the coronavirus infection in the patients using CT images and a convolutional neural network (CNN). In recent days, several machine vision models using CNN presented good results [2-3]. Because of its opportunity for self-training, CNN has become more popular in medical image processing. Based on its superiority, several convolutional neural techniques for COVID- 19 detection have been presented in the literature [4-6]. Unlike all the other image datasets, the imaging techniques database uses a small number of training examples. CNNs have now been boosted with different techniques such as data augmentation to reach ground-breaking performance upon those datasets [7].

In this work, to investigate the patients with COVID-19, a relatively new strategy called CNN is used, wherein the process of learning is constructed using the interconnection of densely linked CNN architectures commonly called as Dense-Net. The fundamental justification for using DenseNet-121 is that it solves the degradation problem, increases feature reuse, and reduces variable usage, all of which are beneficial for creating deep learning algorithms.

DenseNet-121 has also been shown to be useful in identifying disorders using imaging technology [8]. To optimize the design flow, Dense-Net would've been driven by linking all components to just about every other layer below it. This method allows CNN to make judgements based across all levels rather than just one previous layer. In comparison to typical image recognition technologies, Dense-Net is much more advanced and can gather sensory information in a bigger sense.

II. CNN MODEL

A convolutional neural network (CNN) is a multilayer perceptron artificial neural network and it is superior in processing large dataset. It has structure consisting of different structures such as convolution, pooling, flattening, and connected layers. A generalized architecture of CNN is demonstrated in Fig.1.

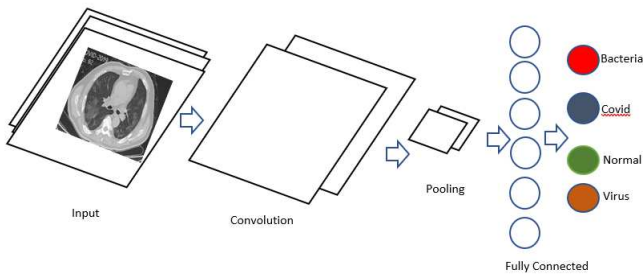


Fig. 1. Convolution neural network architecture

In the CNN architecture, the learnable filters in the convolution layer plays an important role and demonstrate the learning skill of network. Fig. 2 explain the operation of convolution operation.

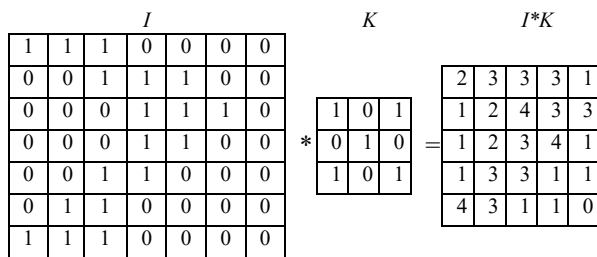


Fig. 2. An example of a convolution operation

In Fig. 2, the notations I , K and $I * K$ represent the pixel values of an image, filter and feature map, respectively. Another crucial component in a network architecture is pooling layer that lowers the computational issues. Here, average and minimum pooling are generally preferred in many applications. Fig. 3 depicts the operation of max-pooling.

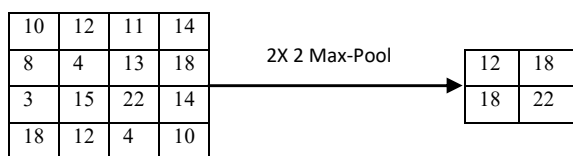


Fig. 3. Max pooling operation

The next block after max-pooling in CNN architecture is fully connected layer. The previous layer neurons are fully interconnected to the next layer. In this, the input data need to be flattened before entering to this layer. Fig. 4 shows such data flattening operation.

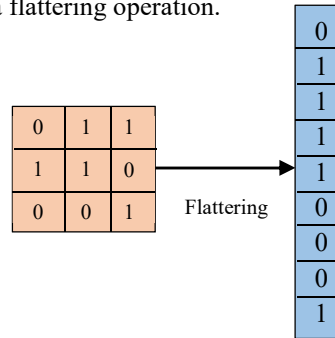


Fig. 4. Flattening operation

From literature, it has been observed that the development of CNN model is big challenge for researches and it is a time-consuming task. However, use of pre-trained network architecture may be a reasonable solution to these issues. The pretrained model can be effectively used for different applications particularly for feature extraction and classification applications. The pre-trained networks such as ResNet 18, GoogleNet and ResNet 50 are most comely used in the literature. The details pre-trained models are summarised in Table I.

TABLE 1: FEW FEATURES OF THE PRE-TRAINED MODELS FOR THE INPUT IMAGE SIZE 224x224 PIXELS

Model	Depth	Size (MB)	Parameters (Millions)
ResNet 18	18	44	11.7
GoogleNet	22	27	7
ResNet 50	50	96	25.6

III. PARAMETERS OPTIMIZATION ALGORITHM

In CNN, the various parameter needs to be optimized to the make the model more efficient for the particular application. The grid search (GS) approach is one of the effective techniques which can be used for the optimization of the parameters. In this approach, the different possible values of the parameters are calculated in the required or predefined range and feed to the defined model to start the optimization process [9].

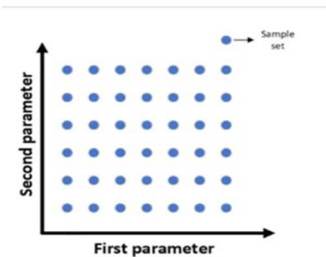


Fig. 5. Parameter sets in the search space for GS

IV. GRID SEARCH BASED CNN ALGORITHM

An effective model based on the GS and CNN approach is shown in Fig. 6. It consists of image dataset, parameters set, pre-trained supported model, and evaluation phase.

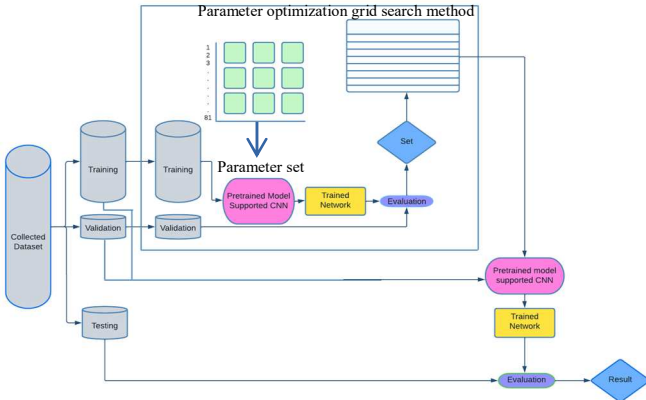


Fig. 6. GS and transfer learning supported CNN model

The first step in the presented model is the image data set collection. Here, freely data set available online websites like Kaggle or GitHub is used for the experimentations. The dataset needs to split into three subsets: training, validation, and testing and randomly percentage of each subset can be selected. The next task is the parameter sets optimization. It can be achieved using grid search algorithm. This step, help to optimized the parameters and can be used as an input to the CNN model. In the next phase, the CNN model is trained using the training data and optimized parameters. In the last phase, the trained CNN is evaluated on testing dataset and conclude the final results in the form of classification and confusion matrix.

Moreover, extensive work in literature depicts that most of the approaches to holding down a process before using the pre-trained models. As data augmentation is one of the important steps of the training procedure for profound learning models, it is used for our benefit and employed as an extra measure. In addition, the Dense-Net study obtained quality precision for CT images. Compared to other learning models, Dense-Net enhances a smaller set of parameters.

For greater accuracy, the Dense Net model with 121 perceptions can be used in the application [8]. It has been observed that the Dense Net-based CNN approach might be a useful detection approach for identifying infected patients using CT images. The CNN model could attain favorable outcomes. However, a CNN model-based application with sizeable training data for classification remains a challenge and hence, the motivation for the researchers. Therefore, Dense-net can be used as an alternative to achieve good results.

A. Dense-Net - A Pre-Trained Network

A Dense-Net is a modified CNN generally exploited in object identification [10]. It is somewhat analogous to Res-Net. Dense-Net combines the future layer and previous layer

output with its concatenated (\cdot) attributes. The conventional CNN aims to obtain the i^{th} output layers using a non-linear transformation $Q_i(\cdot)$ with respect to the previous layer $(P_i - 1)$ output [2].

$$P_i = Q_i(P_i - 1) \tag{1}$$

Dense-Net achieve an efficient information flow between the different layers. Eq. (1) can be modified [11]

$$P_i = Q_i[P_0, P_1, P_2, \dots, P_{i-1}] \tag{2}$$

where $[P_0, P_1, P_2, \dots, P_{i-1}]$ represents a concatenation of previous layer output maps [14].

Data augmentation is one of the approaches that facilitate the user to expand the diversity of data available for training sets, eliminating the need to acquire more data. Data augmentation is a strategy for increasing the number of training specimens by modifying photos whilst preserving semantic features, and it allows for bias-free image data. Basic changes such as horizontal flipping, random cropping, and color augmentations are generally used for model training. Further, data augmentation allows the designer to understand a more diversified set of characteristics, which expands the dataset and helps to keep the model from getting overfitted. The data augmentation to increase image categorization accuracy has also been made.

In this paper, image augmentation is employed to boost training benefits while reducing network regularization. A commonly known data augmentation method is used. Images were arbitrarily trimmed to the 64×64 original image size, with a random rotational range of 3600 pixels on each side. Horizontally and vertically duplicated images are mixed throughout.

B. Image Dataset

Large image datasets are available on websites like Kaggle or GitHub. This is an open source dataset that contain 349 Covid-19 CT images 463 non-Covid CT images from 216 individuals. These images are different in size and shapes and categorized into two types: effected and non-affected. Fig. 7 shows a few cases of CT images for patients that are COVID-19 positive/negative that compose the dataset. The patient with positive results has age, male and female information [16].

C. Model and Applied mechanism

In this work, to categorize corona using DenseNet-121 architecture-based CNN, a dataset comprising images of real patients is used in this study.

The pre-processing can be carried out by normalizing and changing the size of the image for further processing. There are many kinds of pre-processing techniques that can be used for developing our model. In the presented model, the process of image resizes, and normalization is being done on MATLAB through image processing techniques.

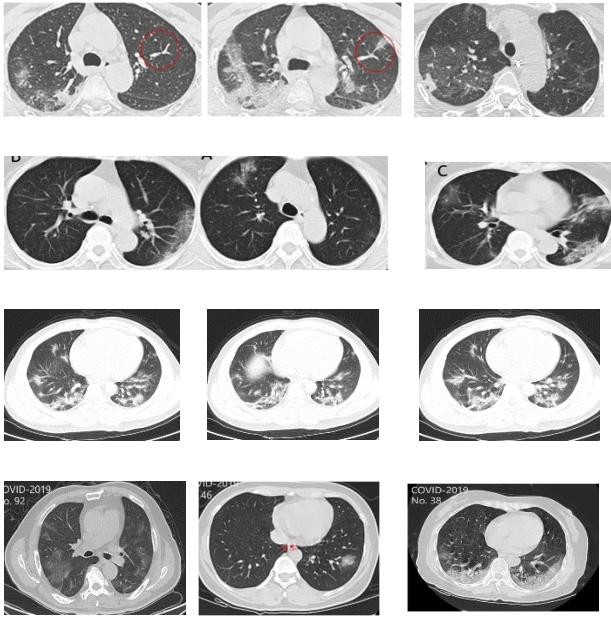


Fig. 7. An example of Covid 19 CT images

In the presented model, 36×36 pixel values are being used in the dataset. Typically, most of the dataset values lie between 0 and 255 but due to the network model, it is preferred to perform in the range 0 and 1 which will be the best fit for the model [2]. The technique helps in the reduction of the complexity of the model. Eq. (3) can be used to normalize the images [12-15].

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (3)$$

The presented COVID-19 classification model is designed using MATLAB software with required packages such as neural networks and image processing. The aim of COVID-19 patients' identification is to assure the positive/negative test. For experimentations, dataset is subdivided into validation (15%), training (70%), and testing (15%) [2]. The systematic procedure for COVID-19 prediction is presented in Fig. 8.

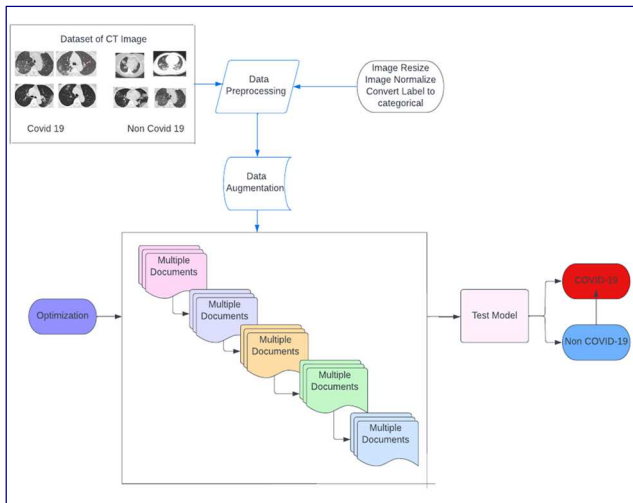


Fig. 8 procedure for predicting COVID-19 is presented

The results of specimens of each category correctly and incorrectly classified can be compiled as a confusion matrix. Accuracy can be determined based on the confusion matrix. The performance of the model is calculated using four performance parameters: recall, precision, G-Mean and F_1 -Measure. All these performance parameters are defined using the following equations.

$$Accuracy = \frac{T_p + T_N}{T_p + T_N + F_p + F_N} \quad (4)$$

where T_p , T_N , F_p and F_N represents true positive, true negative, false positive and false negative, respectively.

$$F_measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

$$G_mean = \sqrt{\frac{T_p + T_N}{T_p + T_N + F_p + F_N}} = \sqrt{Precision \times Recall} \quad (6)$$

V. RESULTS AND ANALYSIS

The various experiments are carried out on the publicly available dataset. Each DenseNet-121-based CNN architecture is trained for various epochs on grayscale image data sets. Grayscale test images were used to evaluate DenseNet-121's performance. Table 2 shows the performance of the presented model.

TABLE 2: PERFORMANCE PARAMETERS FOR DENCENET

Data set	Accuracy	Recall	Gmean	F_Score
Non-covid	0.963	0.852	0.92	0.9
covid	0.857	0.953	0.9	0.89

Fig. 9 (a) and Fig. 9 (b) show the comparison of image prediction for COVID-19 along with non-COVID-19.

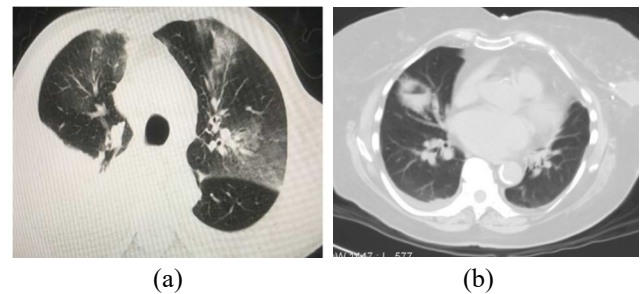


Fig. 9. (a) affected by Covid-19, (b) Non-affected by COVID-19

Generally, the performance calculation and the verification of the presented approach can be demonstrated using the concept of confusion matrix [17]. Thus, the confusion matrix can be used to show the accuracy of the presented approach. The confusion matrix is particular square table which demonstrates and visualized the classification accuracy of multiple classes. In this paper, two classes are classified: Covid positive and Covid negative, respectively. Fig. 10 shows classification accuracy based on the confusion matrix for 25 sample images. It is observed from Fig. 10 that

the DenseNet-121 performs better as classifier and achieved 96% to 100% accuracy.

Confusion Matrix

Output Class	Negative	23 46.0%	0 0.0%	100% 0.0%
	Positive	2 4.0%	25 50.0%	92.6% 7.4%
		92.0% 8.0%	100% 0.0%	96.0% 4.0%
		Negative	Positive	
		Target Class		

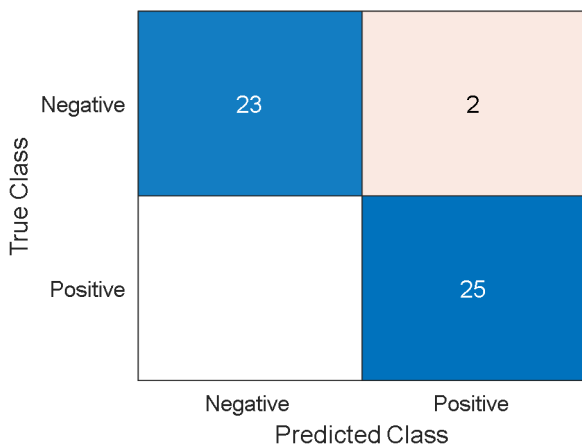


Fig. 10 confusion matrix.

Further, the evaluation of the proposed approach can be carried out using the Receiver Operating Characteristics (ROC) curve [12]. A binary classification is the basis for the ROC which results into four possibilities: true positive, true negative, false positive and false negative, respectively. Fig. 11 shows the results of four classes in the form of ROC curve.

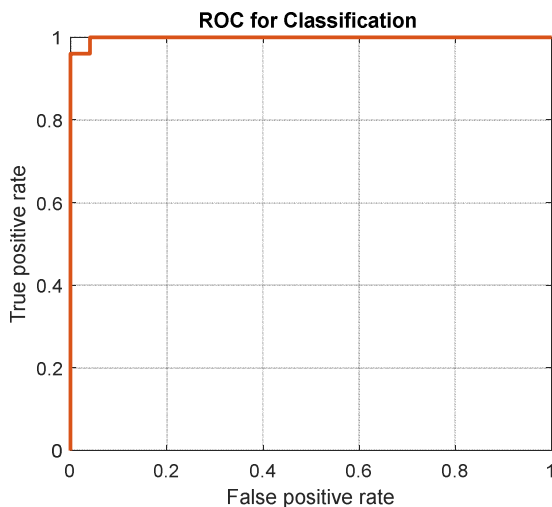


Fig. 11. ROC curve.

The current techniques generally used a smaller dataset, whereas the applied DenseNet-121 model used a relatively

large dataset. Various photos and tactics have been validated using various state-of-the-art approaches. Some multi-class classification models achieved an accuracy of up to 95% or higher while being more difficult and computationally expensive. However, a fair comparison of performance evaluation results and validation is not possible due to the differences in data sets.

VI. CONCLUSION

In this paper, an efficient and systematic COVID-19 infection identification is presented. In this, the infected region can be identified from two different modalities of medical images. COVID-19 patients are identified based on the available CT image dataset. In addition, the qualitative results in the form of four performance parameters: recall, precision, G-Mean and F1-Measure as well as a confusion matrix. The results demonstrate higher accuracy in the classification and detection of infected COVID-19 in CT image datasets.

REFERENCES

- [1] M. D. Mair, M. Hussain, S. Siddiqui, S. Das, A. Baker, P. Conboy, T. Valsamakis, J. Uddin, and P. Rea, "A systematic review and meta-analysis comparing the diagnostic accuracy of initial RT-PCR and CT scan in suspected COVID-19 patients," *The British J. Radiol.*, vol. 94, no. 1119, p.20201039, 2021.
- [2] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi and H. Ghayvat, "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope," *Electron.*, vol. 10, no. 20, p.2470, 2021.
- [3] A. Bhargava and A. Bansal, "Novel coronavirus (COVID-19) diagnosis using computer vision and artificial intelligence techniques: a review," *Multimedia Tools Appl.*, vol. 80, no. 13, pp.19931-19946, 2021.
- [4] M. M. A. Monshi, J. Poon, "Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, p.104375, 2021.
- [5] M. Polsinelli, L. Cinque and G. Placidi, "A light CNN for detecting COVID-19 from CT scans of the chest," *Pattern Recognit. Lett.*, vol. 140, pp.95-100, 2020.
- [6] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu and B. Zheng, "Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms," *Int. J. Med. Inf.*, vol. 144, p.104284, 2020.
- [7] D. X. Xue, R. Zhang, H. Feng and Y. L. Wang, "CNN-SVM for microvascular morphological type recognition with data augmentation," *J. Med. Biol. Eng.*, vol. 36, no. 6, pp.755-764, 2016.
- [8] Q. Zhou, W. Zhu, F. Li, M. Yuan, L. Zheng and X. Liu, "Transfer Learning of the ResNet-18 and DenseNet-121 Model Used to Diagnose Intracranial Hemorrhage in CT Scanning," *Curr. Pharm. Des.*, vol. 28, no. 4, pp.287-295, 2022.
- [9] Y. Sun, S. Ding, Z. Zhang and W. Jia, "An improved grid search algorithm to optimize SVR for prediction," *Soft Comput.*, vol. 25, no. 7, pp. 5633-5644, 2021.
- [10] M. Haris and A. Glowacz, "Road object detection: a comparative study of deep learning-based algorithms," *Electron.*, vol. 10, no. 16, p.1932, 2021.
- [11] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp.1285-1298, 2016.
- [12] N. Kumar, V. H. Gaidhane, and R. K. Mittal, "Cloud-based electricity consumption analysis using neural network," *Int. J. Comput. Appl. Technol.*, vol. 62, no. 1, pp. 45-56, 2020.
- [13] V. H. Gaidhane, N. Kumar, R. K. Mittal and J. Rajevenceltha, "An efficient approach for cement strength prediction," *Int. Journal of Comput. Appl.*, pp. 1-11, 2019.

- [14] V. H. Gaidhane, Y. V. Hote, and V. Singh, "Emotion recognition using eigenvalues and Levenberg–Marquardt algorithm-based classifier," *Sādhanā*, vol 41, no. 4, pp. 415-423, 2016.
- [15] V. H. Gaidhane and Y. V. Hote, "An efficient edge extraction approach for flame image analysis" *Pattern Anal. Appl.*, vol. 21, no. 4 pp. 1139-1150, 2018.
- [16] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang and P. Xie, "COVID-CT-dataset: a CT scan dataset about COVID-19,". *arXiv preprint arXiv:2003.13865*, pp. 1-14, 2020.
- [17] V. H. Gaidhane, Y. V. Hote and V. Singh, V. (2018) 'An efficient similarity measure approach for PCB surface defect detection', *Pattern Anal. Appl.*, vol. 21, pp. 277–289, 2018.

A Review of Cognitive Dynamic Systems and Cognitive IoT

Alessandro Giuliano
McMaster University
Hamilton, ON, Canada
giuliana@mcmaster.ca

Waleed Hilal
McMaster University
Hamilton, ON, Canada
hilalw@mcmaster.ca

Naseem Alsadi
McMaster University
Hamilton, ON, Canada
alsadin@mcmaster.ca

S. Andrew Gadsden
McMaster University
Hamilton, ON, Canada
gadsdesa@mcmaster.ca

John Yawney
Adastra Corporation
Toronto, ON, Canada
john.yawney@adastragr.com

Abstract— This review covers the topics of cognitive dynamic systems and their definition following Simon Haykin's work in the field as well as their application through cognitive radar, cognitive radio and cognitive control. Furthermore, the article presents the topic of cognitive IoT and discusses it under the lens of cognitive dynamic systems referencing research in the field. It also discusses the needs for interoperability between IoT architectures and the need to integrate cognitive radio with future IoT frameworks developments.

Keywords—Cognitive Dynamic Systems, Cognitive IoT, Cognitive Radio

I. INTRODUCTION

The augmentation of physical objects with the power of the internet has become commonplace in the midst of the fourth industrial revolution. Many elements are becoming increasingly intertwined, including wearable technology, healthcare, home appliances, and transportation. The Internet of Things (IoT) has been characterised as a deeper integration of all physical objects with the digital realm, including the advancement from simple control systems sensing devices and effectors to more complex systems capable of exchanging data between devices connected to the internet for more timely and productive decision-making. The magnitude of today's IoT applications has posed several issues in terms of building a system-to-system interoperability framework. Developing IoT services that are intended to adapt to the circumstances and self-adjust in response to unforeseen situations through cooperation. These should also use the acquired data to extract semantic notations and optimize the system's effectiveness. In a general context, the IoT model has been defined as a globally connected network of uniquely addressable devices following established communication protocols. This term refers to a theoretical model of complex multidimensional systems made up of interconnected and interdependent items [6]. Smaller subsystems collaborate to achieve results in the most efficient way possible. Social network analysis is a study tool that is used to explain the network of relationships that exists among the numerous objects that make up the larger Internet of Things, as well as to investigate the effects on data processing, context extrapolation, and semantic

derivation. This is especially beneficial in time varying IoT systems like smart cities.

In a point-of-view essay published in the Proceedings of the IEEE in 2006, Dr. Simon Haykin initially suggested the concept of a dynamic cognitive system (CDS) [3]. He went on to write "Cognitive Radio: Brain Enabled Wireless Communication" [2] and "Cognitive Radar: A Way of the Future" [3], both of which were hugely significant. The author of these defines CDS as systems that learn from repeated enduring interactions with the environment to develop norms of behaviour over time, allowing them to deal with environmental uncertainty. Following Fuster's work on cognition, Haykin refined this concept in [4], outlining the distinction between adaptation and cognition by outlining the norms by which a cognitive dynamic system is defined, the perception action cycle (PAC), memory, attention, language, and intelligence.

Novel cyber-physical systems (CPS) are continually being launched in this new era of greater connectivity, upgrading things with the ability to control the world around them, compute the data acquired, and share it through the internet. This transformation affects a wide range of sectors and services, prompting the development of novel IoT architectures targeted at efficiently capturing and processing data to improve process efficiency. The concept of cognitive IoT (CIoT) has been introduced as a way to enhance current IoT systems with cognitive capabilities in order to better leverage the vast volumes of data being collected and tackle scalability issues. To better examine variables that impact functionality and data gathering, semantic computing, cognitive computing, and perceptual computing can be used [5]. The goal of CIoT is to make IoT systems capable of understanding environmental elements and capable of contextual awareness. This new paradigm aims to apply the principles of human cognition to IoT dynamic systems. The process of learning, reasoning, and understanding the physical and social environments by embedding cognition processes into IoT seeks to build a new class of systems capable of operating with minimal human intervention [7]. Some present obstacles for a scalable and reliable IoT must

also be faced and answered beforehand in order to build such systems. The existing constraints of wireless technology and mobile networks are the first major worry in terms of scalability of such systems. Regarding future large-scale IoT systems, the restricted range, data capacity, and spectrum availability are major concerns [8]. These difficulties will likely intensify in the coming years, given the rapid development of IoT.

Cognitive Radio (CR) and Cognitive Radio Networks (CRN) have sparked the interest of academic and business communities in recent years as a potential solution to many of these issues [8]. Spectrum sensing, spectrum decision, spectrum management, and spectrum mobility are the main processes of cognitive radio, with the goal of taking full advantage of licenced spectrum bands through the effective application of dynamic allocation to fill present spectrum gaps.

A further constraint in vast IoT network is the complexity of aggregating data from multiple sources. Since data collected in a multi-sensor IoT system can be heterogeneous, adaptive analytics approaches must be considered. Collecting such diverse datasets to get a holistic picture of the system can be difficult. The data collected can also be nonlinear, multidimensional, or partial, making its use for intelligent decision-making and services provisioning much more difficult [7][1]. This problem is addressed by Cognitive IoT, which adapts to the data type, situation, and setting, utilising techniques like association analysis, clustering analysis, and regression, for contextual data analysis. Big-data-driven applications, on the other hand, necessitate more intelligent decision-making to enable more efficient and flexible operations via cooperative self-organized and self-optimized behaviours. Moreover, for large-scale IoT systems, centralized data handling is a significant barrier. The challenges associated with central data processing include single-node failure, restricted scalability, and massive trade overhead [7].

II. RELATED WORK AND MOTIVATION

Because of the wide range of disciplines to which IoT may be applied, developments in architectural design for IoT systems have primarily been targeted to individual applications. As a result, cooperation amongst IoT systems is constrained, thereby restricting advancement toward a bridge architecture [9].

Although cognitive IoT is still a new topic, it is growing in prominence because of scientific research in cognitive dynamic system and cognitive control. This new paradigm could be used as a model for developing new IoT designs and as a framework for addressing specific IoT concerns. In principle, CIoT aspires to provide IoT systems with a cognition component that allows them to learn, reason, and comprehend both physical and social realms [7], bringing

together a variety of professions and areas such as computer science, mathematics, cognitive science, neurology, and engineering. CIoT may improve the interconnection of diverse IoT networks and be applied beyond disciplines and sectors, spanning the physical and cyber worlds to improve smart distribution of resources, autonomous process controls, and intelligent service provisioning.

The Internet of Things can be viewed as a macro-level development of ubiquitous computing combined with CPS. At the moment, IoT can only use permitted spectral bands, which are presumably already being used, posing an impediment for large-scale IoT implementation. As trillions of objects become more interconnected in the near term, we can anticipate the issue to intensify [14]. In turn, by significantly increasing IoT data transmission using cognitive radio and incorporating machine learning, signal processing, and other technologies, effective distribution of data transmission within 4G and 5G licenced available spectrum could be an answer [13].

Cognitive Radio is a promising enabling communication technology for IoT, dealing with issues such as wireless access network conflict and severe congestion, as well as automaticity, scaling, dependability, energy consumption, and service quality [14]. The most immediate advantage of CR for IoT is that it allows for more efficient spectrum allocation and administration, which improves accessibility, usability, flexibility, and interconnectivity. Addressing efficient and flexible networks and addressing heterogeneity concerns are the two types of CR techniques, with flexible networking referring to the optimal use of available spectrum via spectrum aware optimization to enhance QoS. While addressing diversity, the aim is to strengthen environment discovery, self-organization, adaptability, and nodal cohabitation.

Devices in both centralised and decentralised networks will require routing in order to convey data to a predetermined destination. However, traditional single-hop and multi-hop routing algorithms are incompatible with cognitive radio systems because they lack additional functions like flexible spectrum allocation. CR mesh networks (semi-static) and CR ad hoc networks (adaptable and self-reconfiguring using P2P interactions) have both been discussed in the literature [14]. By accessing data about interference zones and relaying this to the underlying common infrastructure or cluster leader, spectrum knowledge would be readily available to all nodes in centralised networks. Delay, hop count, energy usage, bandwidth, and route stability will be among the network parameters which will be collaboratively optimised. Simultaneously, single-hop, D2D networking will most likely be used in centralised networks provided that the transmitter's range is not surpassed.

A. Cognitive Dynamic Systems and Cognitive Control

Dr. Simon Haykin was the first to propose the concept of a cognitive dynamic system as an answer to the radio spectrum's utilisation inefficiencies. Prospective bandwidth users' capacity for using non - utilized bandwidths is constrained due to government organisations' control of electromagnetic bands in the nature of licences [2]. Cognitive radio was created to maximise the usage of existing radio frequencies by taking advantage of spectral gaps. The study describes it as a smart wireless communication technology that learns from the surroundings and adjusts its settings in real-time using the approach learning by building.

By assessing the electromagnetic environment, finding channels, and transmitting data through dynamic bandwidth control, cognitive radio seeks to maintain high reliability in connectivity utilising the radio spectrum efficiently. The radio scene analysis is performed by the receiver, which surveys the surroundings to detect spectrum holes and calculates the interference temperature. Dynamic beamforming is used for inference control in this passive activity, which requires interpreting non - stationarity temporal signals to accommodate for the spatial characteristics of radiofrequency inputs. This strategy relies on constant spectrum observation and the computation of alternate paths to recognised spectrum holes since it offers resilience whenever a main user requires the spectrum for its own purpose.

Difficulties in the channel estimation problem are solved by adopting semi-blind receiver training, resulting in a receiver with two modes: supervised learning and tracking. The first option acquires and estimates the channel value using a quick training cycle. The other, on the other hand, is intended to be used in operating condition and repeatedly evaluates the channel state. The computations are performed using a state-space model of the channel parameters, with the premise of linearity, utilising the process equation and measurement equation. Choosing an effective monitoring approach and filter choice addresses AR coefficients, dynamic noise, and measurement noise.

Cognitive radio might have to function in a distributed mode via a cooperative system that accomplishes collaboration among nearest neighbors in constant interaction, widening the reach of its application and making it easier for multiple users to adopt and implement the technology to existing networks. The challenge of transmit-power regulation for different users can indeed be viewed as a game-theoretic problem. The author suggested Nash equilibrium and water-filling procedures as remedies. The transmit-power regulation problem affects the dynamic spectrum management system in a similar fashion. The transmitter handles each of these components of the operation, thus it employs the very same methods to address it.

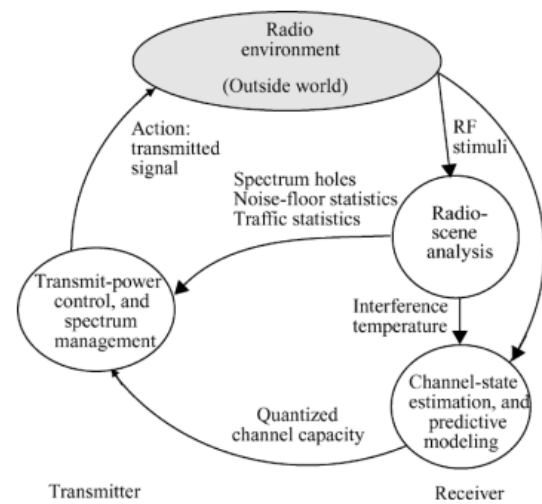


Figure 1: Cognitive Radio operational representation. [2]

In [15] authors considerations are raised regarding modulation techniques and traffic control, with a focus on OFDM. Orthogonal frequency-division multiplexing (OFDM) is a powerful modulation approach for cognitive radio and a cost-effective way to enable dynamic bandwidth allocation. Co-channel interference must be avoided when using an OFDM, which necessitates the inclusion of a traffic control system in the dynamic spectrum management algorithm. This system will be able to anticipate the length of time the spectrum gap would be empty, as well as predicting traffic patterns, based on previous data. More information regarding routing protocol, gateways and more can be found in [16]-[19]

Following the new paradigm outlined by the five principles of cognition, a system is deemed cognitive if it can perform the five basic cognitive processes: perception-action cycle (PAC), memory, attention, language, and intelligence. The perception-action's cycle concept ties to feedback loops, using sensing devices to extract data regarding the system's condition and functioning, which is then used to trigger predetermined events that affect the environment and the system itself within the context of the Internet of Things. To reach intelligent decision-making, this approach employs advanced data analytics and the other components of cognitive dynamic systems. Expanding on the PAC, by using relevant stored information about the surroundings, the system, and past behavior, which are stored in order to improve the system's reaction to hypothetical situations. perceptual, executive, and working memory are the three types of memory.

The PAC and memory elements of CDS are responsible for attention. This refers to the cognitive system's ability to comprehend data and properly optimise all preceding operations. In a cognitive dynamic system, attention is the systematic method for prioritising the distribution of

computational capabilities to alleviate the problem of information overload. As a result, dynamically filtering processed data by significance to aid learning and cognitive controller enhancement are employed. In CDS, attention is not defined by a physical state, but rather by an artificial process that shows itself within system. Network protocols used by devices to interact and send information to other systems represent speech in engineering systems. To share data, cognitive systems should be able to adjust to any communication protocol. Nevertheless, initiatives to standardise such protocols in the IoT are aimed at finding an easier solution. Intelligence is built on the preceding four cognitive operations and incorporates them into an analytical process capable of choosing the best decisions. Intelligence can carry out an assessment and develop appropriate action in the face of unanticipated conditions and uncertainty to then learn from it.

This innovative feature to Haykin's realisations sparked the special necessity to incorporate cognitive control further into conceptual frameworks of cognitive dynamic systems [10]. Fatemi et al. present a fresh perspective on cognitive control in this study, concentrating on two key components: training and planning, both of which are based on two basic ideas. Firstly, the global perception-action cycle, which in this case refers to a cyclic controlled stream of environmental information and is the basic premise of cognition. Secondly the two-state model is composed of the target state, or target of interest, and the current entropic state of the preceptor, which can be viewed as a measure of the lack of sufficient data in the cyclic flow of information from the global PAC. Mathematically it is represented by a state-space model of the environment defined by a process equation and a measurement equation. The target state, or target of concern, is made up of the goal state and the entropic state of the preceptor, which again can be thought of as a gauge of the insufficient data in the cyclical flow of information from the global PAC. A state-space model of the environment characterised by a process equation and a measurement equation is used to describe it analytically.

The purpose of cognitive control is to optimise cognitive strategy, which is defined as the probability density function of cognitive activities, such as the impact of past behaviour on present state, via learning and scheduling. The current state of the preceptor is described using Shannon's entropy notion, which quantifies the disturbance existing as a probability distribution depending on acquired data. The system attempts to estimate future entropic state of the system and apply it in the planning stage of the cycle through gradual variations, formalised by an automatic rewarding mechanism at the end of every iteration [10]. The core of the cognitive controller capabilities is the cognitive control algorithm, which is described in the article as a method to converge towards the optimal policy. This is defined and proved as a special example of dynamic programming that inherits the fundamental traits of convergence and optimality.

The authors explain a combined strategy of pure exploration and pure exploitation dubbed the $\epsilon - greedy$ strategy, which they adapted from Powell et al. [11], to accelerate the learning and convergence of the algorithm to optimized parameters. This balance among exploitation (selecting activities based on the highest value criterion) and exploration (purposeful sampling actions arbitrarily) could be considered as an attentiveness enhancer. Assigning computing resources to sustain developing awareness of the environment while avoiding local sub-optimal solutions [10]. Figure 2 depicts the global feedback loop as well as the interconnections between both the cognitive perceptor and the cognitive controller. The implementation of cognitive control studied in [10] uses this novel cognitive controller idea to solve a radar surveillance challenge. The cognitive controller adjusts the parameters in the system to enhance the prediction of the object's position, velocity, and ballistic coefficients. When contrasted to a static waveform radar, dynamically modifying the waveform led to a four-order-of-magnitude enhancement in performance.

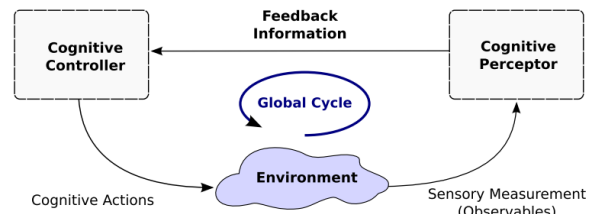


Figure 2: Cognitive Control flow diagram. [10]

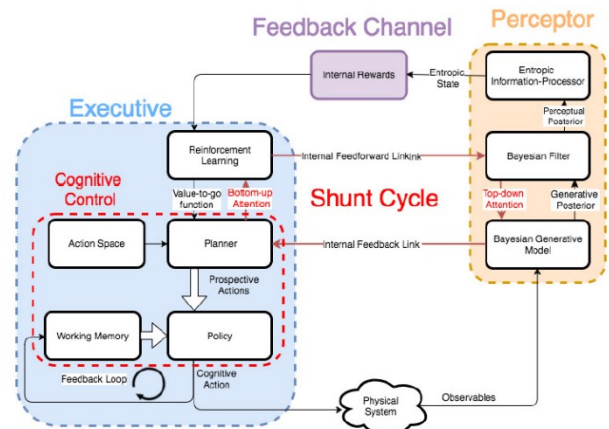


Figure 3: Architectural composition of CDS inclusive of Cognitive Risk Control mechanisms. [12]

In the face of future unforeseen uncertainty, cognitive control, as previously established, misses a strategy for anticipating undesirable outcomes or impediments, often known as risk. The researchers in [12] go even farther, recognising the necessity for a component able of interfacing with multiple parts of CDS, like the preceptor, working memory, and executive memory, in order to predictably adjust the system to the new unpredictable environment. To drive CDS via timely risk-avoidance actions, this redefined

subsystem employs a Bayesian filter algorithm and a Bayes generative model [12]. Figure 3 depicts the perceptor's engagement with the Bayesian-based subsystem. The posterior is calculated on the latest state derived from previous iterations in this generative model, with the caveat that each PAC cycle may have multiple repetitions.

The article's Bayesian filtering is the sub-optimal Kalman filter, which is applied under premise of a linear model. For nonlinear applications, the cubical Kalman filter is recommended. The screening phase' goal is to gather up valuable information from the generative model and discard useless details, all while refining the important information given to the entropic information processor in a top-down attention process. In addition, a shunt cycle is established to transfer bottom-up attention from the scheduler to that same reinforcement learning algorithm, culminating in localized feedback between the systems. The entropic state computation flows through into internal rewarding mechanism, which feeds into the executive for reinforcement learning and task switch control. This has two key qualities that let it distinguish between different scenarios: internal rewards are always positive in the absence of uncertainty and persistently negative in the presence of uncertainties. The reinforcement learning mechanism computes and transforms them into a value-to-go function, which is used as input to the cognitive controller. The action space (which contains all theorised activities), internal incentives, discount factor (a weight provided to progressively discount prior actions), and policy, all impact this.

The cognitive controller is made up of the planner and the policy, as defined earlier in this section (the function that leads to decision making) and of a classifier involved in decision-making, selecting risk-sensitive cognitive actions when there is ambiguity. The classifier assigns a specific posterior to prior disrupted cognitive activities stored in executive memory based on N past events. Moreover, a task switching mechanism is provided to avoid the disrupted cognitive operations from impacting executive memory; this directly connects with the internal reward systems' dual composition. Pre-adaptation is accomplished by correctly identifying events that took place in hazardous uncertain situations versus those that did not. In the CDS framework, a set of gates are employed to divert the flow of data to other regions, necessitating additional investigation if disruptive cognitive actions are required [12]. The implementation of effective policy decision is of major relevance in the administration of these systems, hence cognitive control is highly pertinent in the IoT field.

The following Section will discuss some attempts to apply the CDS framework to the IoT and more general trials to bring cognition into these systems.

B. Cognitive IoT

Incorporating a cognitive element to the Internet of Things advances existing studies in the areas, which is focused on enabling generic objects to detect their surroundings and share their findings with a central administrator. According to Wu et al. [7], simply being connected is insufficient. IoT systems should be able to learn,

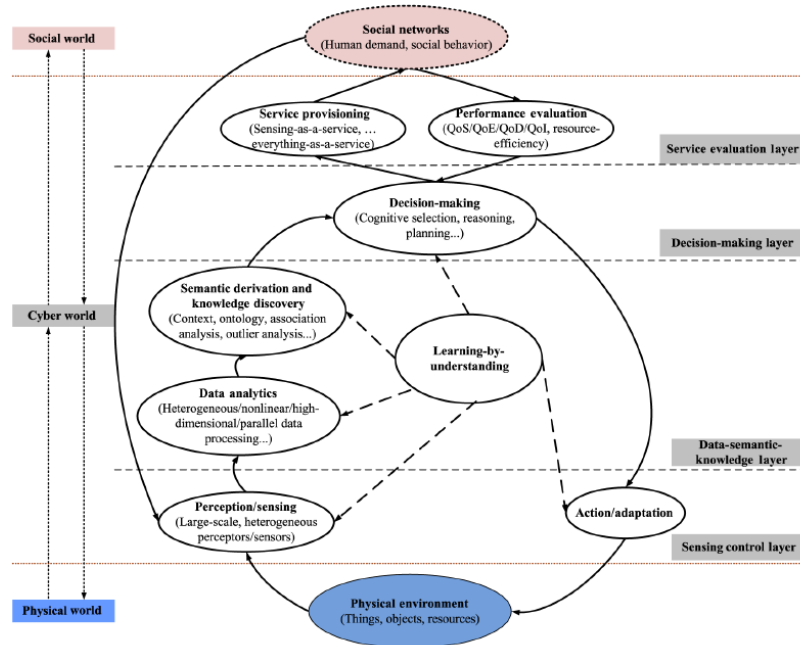


Figure 3: Cognitive Control flow diagram. [7]

think, and comprehend the physical and social world on their own, giving them "high-level intelligence" [7]. Based on past work by Haykin and Fuster, the study proposes a new concept, the Cognitive Internet of Things (CIoT). It presents a new implementation strategy based on interactions between five basic cognitive tasks: the perception-action cycle, large data analysis, semantic extraction, knowledge extraction, intelligent decision-making, as well as on demand resource provisioning [7].

By bridging the tangible, virtual, and social worlds and enabling smart distribution of resources, active network operation, and smart service provisioning, the researchers introduce a new network framework wherein physical/virtual objects are interconnected and act as autonomous agents with minimal intervention. The structure can be separated into layers. Through processing incoming inputs and feedback data, the sensory control layer, which is linked to the global PAC cycle, directly interacts with the surroundings. The semantic knowledge layer, which is concerned with semantic and ontological derivations, further analyses the input in order to provide context awareness. The decision-making layer reasoned, planned, and selected the best appropriate action for the interacting parts to take to use the information extracted from the preceding layer. The service evaluation layer evaluates the services provided and the feedback, using innovative social world-related performance measures.

Planning and choosing are the two aspects of decision-making generally. The article refers to the process of selecting an action from a range of options based on gathered data and deduced knowledge in Cognitive Radio Networks (CRN), which is driven by their learning capabilities. The ability to cognitively modify based on previous and present data is known as cognitive selection. The research highlights three types of cognitive selection: Markovian decision processes, multibandit armed problems, and multiagent learning. Since a distributed IoT architecture is likely to have a high number of decision-makers, the authors focus on the last described method, modelling it using game theory and studying the learning approach with ambiguous, volatile, and incomplete data [7].

Noncooperative game theories, which characterise exchanges between individual decision whereby each player optimises its utility function, are indeed a good fit for the challenge. The development of this systems is primarily concerned with constructing a utility function and achieving acceptable stable solutions. Local relationships amongst actors and spatial game models present additional hurdles in sizable CIoT systems. While global information exchange is impossible in a traditional large-scale IoT system, local interactions between agents can be achieved via regional collaboration, resulting in near-optimal solutions. The paper does not address whether blockchain may be a plausible solution for the worldwide flow of information in an IoT

ecosystem and could be a future topic for research in this field.

In CIoT, evaluating the overall system performance is a difficult operation that is reduced by classifying the data into two dimensions: cost and profit. Three primary measures are presented in the profitability dimension. The quality of data (QoD) is the initial parameter, which assesses the data gathering procedure as well as the reliability of sensed data. Furthermore, the QoD ought to be able to measure data completeness, veracity, and timeliness. The next indicator is quality of information (QoI), which reflects the amount of useful data obtained over a certain task based on precision, accuracy recall, and volume by the decision-maker. These specifics define the quality of the information presented. Finally, the profit dimension's quality of experience (QoE) indicator assesses customer experience relating to access, steady operation, speed, and requirements. Device usage performance, computational complexity, energy efficiency, and storage efficiency, on the other hand, are the cost aspect measures provided in the study.

Home automation provide an ideal platform for analysing the prospects of CIoT as a people centric IoT that enhances the quality of life by dynamically modifying the living spaces in the context of linking the cyber-physical and social worlds, as described in [21]. Additionally, the growing presence of intelligent sensors in households makes it feasible to introduce intelligence to today's smart homes, buildings, automobiles, and, eventually, cities.

In a larger sense, CIoT could be deployed to smart cities in a variety of ways. Feng et al. detail a test case employing the CDS concept towards the Internet of Vehicles (IoV) in smart cities in [22], claiming that modernising the transport network has the potential to decrease traffic, vehicle crashes, and commuting expenses. CAVs (connected autonomous vehicles) are ideal for this since they can change their activities in response to perceived environmental data. To keep up with recent advancements in the use of electric automobiles, the article expands this description to RACE vehicles. The adoption of vast CAV networks could help both commercial and public mobility but would also expose them to cyber-attacks. While delving into the CDS framework for smart vehicles, the authors explain the cyber dangers that such networks may face and recommend countermeasures to maintain the system's resilience, security, and privacy. This study looks at active attacks like jamming, binding, and FDI attacks, as well as passive attacks such eavesdropping and stalking [22]. Due to complex, variable, and hostile environment in which CAV function, adding CDS as a proactive supervisor of all components existing in a car is desired to improve risk management reduction via joint interoperability and adaptability. Relying on the context extrapolation features of CRC, operational sensors like LIDAR, video cameras, radio receivers, and radar receivers might be dynamically modified to the scenario, increasing

their functionality. The authors offer an improved CRC framework that is based on a Bayesian generative model and entropic information processing, which uses a task switch method of control to change mode of operation situationally. The reinforcement learning and scheduler, action library, policy, classifier, working memory, and executive memory all make up the executive element of CDS, that would estimate the optimum cognitive action or policy based on the adaptive and filtered feedback information. Further studies in the application of CIoT to smart cities, smart manufacturing and smart energy grids can be found in [23]-[25].

III. CONCLUSION

This article provides an overview of Cognitive Dynamic Systems and how they can be used in the Internet of Things. The standardisation initiatives aimed at expediting the creation of new IoT designs were also discussed in order to highlight the current issues with interoperability in IoT architectures, which is required to develop larger scale and comprehensive IoT and CIoT architectures. In addition, based on current research, the usage of Cognitive Radio for IoT was examined. Finally, IoT was used to improve cognitive policy selection and how CIoT is approached.

IV. REFERENCES

- [1] S. Haykin, "Cognitive Dynamic Systems," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1910–1911, Sep. 2008, doi: 10.1109/jproc.2006.886014.
- [2] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb. 2005, doi: 10.1109/JSAC.2004.839380.
- [3] S. Haykin, "Cognitive radar," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30–40, 2006, doi: 10.1109/MSP.2006.1593335.
- [4] S. Haykin, "Cognitive Dynamic Systems," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 5, no. 4, pp. 33–43, Oct. 2011, doi: 10.4018/jcini.2011100103.
- [5] A. Sheth, "Internet of Things to Smart IoT Through Semantic, Cognitive, and Perceptual Computing Targeted and Limited Use of IoT Data," *IEEE Computer Society*, vol. 16, pp. 1541–1672, 2016.
- [6] A. Zelenkauskaitė, N. Bessis, S. Sotiriadis, and E. Asimakopoulou, "Interconnectedness of complex systems of internet of things through social network analysis for disaster management," in *Proceedings of the 2012 4th International Conference on Intelligent Networking and Collaborative Systems, INCoS 2012*, 2012, pp. 503–508. doi: 10.1109/iNCoS.2012.25.
- [7] Q. Wu *et al.*, "Cognitive internet of things: A new paradigm beyond connection," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 129–143, Apr. 2014, doi: 10.1109/JIOT.2014.2311513.
- [8] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-Radio-Based Internet of Things: Applications, Architectures, Spectrum Related Functionalities, and Future Research Directions," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 17–25, 2017, doi: 10.1109/MWC.2017.1600404.
- [9] S. Krco, B. Pokric, and F. Carrez, "Designing IoT architecture(s): A European perspective," in *2014 IEEE World Forum on Internet of Things, WF-IoT 2014*, 2014, pp. 79–84. doi: 10.1109/WF-IoT.2014.6803124.
- [10] M. Fatemi and S. Haykin, "Cognitive control: Theory and application," *IEEE Access*, vol. 2, pp. 698–710, 2014, doi: 10.1109/ACCESS.2014.2332333.
- [11] W. B. Powell, *Approximate dynamic programming: solving the curses of dimensionality*. Wiley, 2011.
- [12] S. Haykin, J. M. Fuster, D. Findlay, and S. Feng, "Cognitive Risk Control for Physical Systems," *IEEE Access*, vol. 5, pp. 14664–14679, Jul. 2017, doi: 10.1109/ACCESS.2017.2726439.
- [13] F. Li, K. Y. Lam, X. Li, Z. Sheng, J. Hua, and L. Wang, "Advances and Emerging Challenges in Cognitive Internet-of-Things," *IEEE Transactions on Industrial*
- [14] P. Rawat, K. D. Singh, and J. M. Bonnin, "Cognitive radio for M2M and Internet of Things: A survey," *Computer Communications*, vol. 94, pp. 1–29, 2016, doi: 10.1016/j.comcom.2016.07.012.
- [15] W. Lu, S. Hu, X. Liu, C. He, and Y. Gong, "Incentive Mechanism Based Cooperative Spectrum Sharing for OFDM Cognitive IoT Network," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 662–672, Apr. 2020, doi: 10.1109/TNSE.2019.2917071.
- [16] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-Radio-Based Internet of Things: Applications, Architectures, Spectrum Related Functionalities, and Future Research Directions," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 17–25, 2017, doi: 10.1109/MWC.2017.1600404.
- [17] S. Franklin, T. Madl, S. D’Mello, and J. Snaider, "LIDA: A systems-level architecture for cognition, emotion, and learning," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 1, pp. 19–41, Mar. 2014, doi: 10.1109/TAMD.2013.2277589.
- [18] K. R. Chowdhury and I. F. Akyildiz, "CRP: A routing protocol for cognitive radio ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 794–804, Apr. 2011, doi: 10.1109/JSAC.2011.110411.
- [19] F. Jalali, O. J. Smith, T. Lynar, and F. Suits, "Cognitive IoT gateways: Automatic task sharing and switching between cloud and Edge/Fog computing," in *SIGCOMM Posters and Demos 2017 - Proceedings of the 2017 SIGCOMM Posters and Demos, Part of SIGCOMM 2017*, Aug. 2017, pp. 121–123. doi: 10.1145/3123878.3132008.
- [20] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-Radio-Based Internet of Things: Applications, Architectures, Spectrum Related Functionalities, and Future Research Directions," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 17–25, 2017, doi: 10.1109/MWC.2017.1600404.
- [21] S. Feng, P. Setoodeh, and S. Haykin, "Smart Home: Cognitive Interactive People-Centric Internet of Things," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 34–39, Feb. 2017, doi: 10.1109/MCOM.2017.1600682CM.
- [22] S. Feng and S. Haykin, "Cognitive Dynamic System for Future RACE Vehicles in Smart Cities: A Risk Control Perspective," *IEEE Internet of Things Magazine*, vol. 2, no. 1, pp. 14–20, Sep. 2019, doi: 10.1109/iotm.2019.1900008.
- [23] P. Vlacheas *et al.*, "Enabling smart cities through a cognitive management framework for the internet of things," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 102–111, Jun. 2013, doi: 10.1109/MCOM.2013.6525602.
- [24] H. S. Park and N. H. Tran, "An autonomous manufacturing system based on swarm of cognitive agents," *Journal of Manufacturing Systems*, vol. 31, no. 3, pp. 337–348, Jul. 2012, doi: 10.1016/j.jmsy.2012.05.002.
- [25] D. N. Molokomme, C. S. Chabalala, and P. N. Bokoro, "A review of cognitive radio smart grid communication infrastructure systems," *Energies*, vol. 13, no. 12. MDPI AG, Jun. 01, 2020. doi: 10.3390/en13123245.

IoT Devices Proximity Authentication In Ad Hoc Network Environment

Ali Abdullah S. AlQahtani
 Computer Systems Technology
 North Carolina A&T State University
 Greensboro, North Carolina
 alqahtani.aasa@gmail.com

Hosam Alamleh
 Computer Science
 University of North Carolina Wilmington
 Wilmington, North Carolina
 hosam.amleh@gmail.com

Baker Al Smadi
 Computer Science
 Grambling State Univeresity
 Grambling, Louisiana
 bakir_smadi@hotmail.com

Abstract—Internet of Things (IoT) is a distributed communication technology system that offers the possibility for physical devices (e.g., vehicles, home appliances sensors, actuators, etc.), known as *Things*, to connect and exchange data, more importantly, without human interaction. Since IoT plays a significant role in our daily lives, we must secure the IoT environment to work effectively. Among the various security requirements, authentication to the IoT devices is essential as it is the first step in preventing any negative impact of possible attackers. Using the current IEEE 802.11 infrastructure, this paper implements an IoT devices authentication scheme based on something that is *in the IoT device's environment* (i.e., ambient access points). Data from the broadcast messages (i.e., beacon frame characteristics) are utilized to implement the authentication factor that confirms proximity between two devices in an ad hoc IoT network.

Index Terms—Internet of Things, IoT, ad hoc, proximity, Beacon Frame, IoT Authentication

I. INTRODUCTION

THE Internet over the last four decades has developed from peer-to-peer networking (P2P), world-wide-web (WWW), and mobile-Internet to the IoT. The IoT is a network that might consist of animals, people, objects, physical devices (e.g., home appliance sensors, actuators, vehicles, digital machines, etc.) that can collect and exchange data with each other without human intervention. The way of communication in an IoT network can be between people, between people and devices, and between devices themselves, also called machine-to-machine (M2M).

The IoT environment has three major components; IoT devices, Cloud, and Client applications. IoT devices are responsible for sensing and measuring the world around them, taking local action as necessary (turning off/on or opening/closing an object, sharing data, etc.). The cloud is where the powerful applications reside and can collect data from IoT devices, combine IoT device data with other data sources, perform data analytics to reveal trends, identify problems, and predict the future. The client applications enable users to access and view data processed in the cloud issue commands to remote IoT devices.

In 2015, approximately 15 billion IoT devices worldwide were in use [1], which was doubled in 2020 (approximately around 31 billion) and might be around 60 billion by 2024 [2].

In general, IoT is continuously evolving and has become an actual attractive area of attack for hackers. The number of cyber-attacks on IoT devices is raised by 600% in only one year, from 2016 to 2017, corresponding to 6000 and 50,000 reported attacks, respectively [3]. IoT devices are usually subject to Distributed Denial-of-Service (DDoS) and ransomware attacks due to the fact that these attacks take advantage of storage limitations and their internet-supported connectivity [4] which grants hackers the opportunity to interact with devices remotely.

One way IoT networks can mitigate cyber-attacks is to establish authentication, which is based on trusting the other end before communicating with. As of today, there is a number of ways to establish authentication in IoT networks:

- **On-way authentication:** in the case before two IoT devices start to communicate with each other, only IoT device authenticates itself to the other, while the other IoT device will not be authenticated.
- **Two-way authentication:** both IoT devices must authenticate themselves to each other prior to the communication.
- **Three-way authentication:** a service provider is involved in this type, which authenticates the two IoT devices and assists them to authenticate each other.
- **Distributed:** this method utilizes a distributed authentication technique between the IoT devices prior to the communication.
- **Centralized:** a trusted third end is utilized to distribute and manage the authentication certificates used.

IoT devices can be connected to a centralized network, in which all devices are connected directly the gateway. Alternatively, IoT devices can be connected to each other in ad hoc networks, where communications between IoT devices is used to relay information of other devices to

the gateway.

We propose a technique to authenticate IoT devices in ad hoc networks to verify proximity. This is done in a way that only devices within a certain distance from other authenticated IoT devices will be able to connect to the network. Meanwhile, devices that are far from an authenticated device or not physically in the area will fail in the proximity authentication. The proposed system enforces the security in ad hoc IoT networks.

II. CONTRIBUTION

Using the current IEEE 802.11 infrastructure, this paper implements an IoT devices proximity authentication scheme in IoT ad hoc networks. This based on something that is *in the IoT device's environment* (i.e., ambient access points). In this paper, data (a Wi-Fi *footprint*) from the broadcast messages are utilized to implement the proximity by determining whether two devices are within a certain range for an authenticated IoT device unobtrusively.

III. GUIDE TO PAPER

This paper is organized as follows; *Section IV* reviews the previous research on IoT devices authentication. *Section V* describes the proposed scheme system for beacon frame-based IoT devices authentication. The experiment of the proposed system and results are presented in *Section VI*. The proposed scheme is analysed in *Section VII*. Security analysis is discussed in *Section VIII*. Lastly, *Section IX* illustrates the conclusion for the proposed method.

IV. RELATED WORK

This section provides an analytical overview of the literature proposing proximity-based authentication. Some systems use GPS for proximity verification [5], [6]. This is mainly done by comparing the GPS location calculated at two device. Few researches propose using GPS for location verification. However, using GPS suffers from performance limitations indoors [7]. Other systems use wireless sensor network for localization [8]. Moreover, utilizing wireless sensor network requires additional hardware installed on IoT devices, which can be costly to deploy and maintain. On the other hand, Bluetooth had been a popular choice for proximity-based authentication [9], [10]. In addition, the weakness in these approaches is that Bluetooth typically has a short-range and requires additional hardware that is not always guaranteed to be in the infrastructure or in every user's device [7]. Wi-Fi is a popular solution for proximity-based authentication are ubiquitous and widely used for connectivity in many users [11]–[13] and IoT devices. A few works use channel characteristics to verify proximity using signal information received from access points such as SSID, MAC address, and Received Signal Strength Indicator (RSSI) [14], multipath profiles [15], or mathematical modeling

on channel characteristics [10], [16]. The proposed system uses Wi-Fi measurements to achieve authentication in the context of ad hoc IoT networks. This can be done by utilizing the role of Representational State Transfer (REST) API in the IoT Systems, which is able to record and count everything [17]. The proposed system provides a complete framework for IoT devices proximity verification and management, which is done by utilizing Wi-Fi Beacon frame information to perform proximity-based authentication.

V. THE PROPOSED SCHEME

In this section, we will present the proposed scheme in detail. The proposed scheme aims to authenticate new IoT devices joining and ad hoc network. The type of authentication that is carried out is proximity-based. This aims to prevent adding new IoT nodes that are outside the "allowed area". This would also eliminate adding fake IoT nodes, or adding IoT node outside the allowed physical boundaries. The proposed system utilizes Wi-Fi beacons in the area to achieve this goal. In general, Wi-Fi access points periodically broadcast their own beacon frame, including Service Set Identifier (SSID) and Basic Service Set Identifier (BSSID). Moreover, Using the Wi-Fi footprint, we can measure the RSSI value of every presented access point in the location.

Before a new IoT device joins an IoT environment, an authenticated IoT device must verify whether or not the new device is within the boundary of operations using their site following the steps below:

- 1) An authenticated IoT device in the system scan for the beacon frame of every Wi-Fi access point in the environment then collect it. Also, it measure the RSSI value of every presented access point as follows:

$$B_1 \begin{cases} Tuple_1 = \{SSID_1, BSSID_1, RSSI_1\} \\ Tuple_2 = \{SSID_2, BSSID_2, RSSI_1\} \\ \vdots \\ Tuple_n = \{SSID_n, BSSID_n, RSSI_n\} \end{cases} \quad (1)$$

- 2) When a new device wants to be added to the network the authenticated device verifies whether the new device is within the boundary of operations.
- 3) Similar to step, 1 the new device scan for the beacon frame of every Wi-Fi access point in the environment and then collect it. Also, it measures the RSSI value of every presented access point as

follows:

$$B_2 \begin{cases} Tuple_{e_1} = \{SSID_1, BSSID_1, RSSI_1\} \\ Tuple_{e_2} = \{SSID_2, BSSID_2, RSSI_1\} \\ \vdots \\ Tuple_{e_n} = \{SSID_n, BSSID_n, RSSI_n\} \end{cases} \quad (2)$$

Then it sends it to the authenticated device.

- 4) When the authenticated device receives the data in step 3, it calculates the Euclidean distance between the two data sets, B1 and B2, using equation 3:

$$D = \sqrt{\sum_{i=1}^n [B_1 Tuple_{e_i}(RSSI) - B_2 Tuple_{e_i}(RSSI)]^2} \quad (3)$$

Where D is the Euclidean distance.

If D is below a certain defined threshold, the proximity authentication is successful. If D is above this threshold, proximity authentication fails. The threshold is determined based on a calibration experiment. Such calibration can be done by the vendor for similar devices. Alternatively, it can be calculated at the area of operation for different devices.

The proposed system facilitates proximity authentication for multiple nodes. Authenticated devices are able to verify new device proximity before authenticating them to the network. This is done using the threshold, which is determined by the Euclidean distance in equation 3. When the calculated distance is below the threshold, it means that the accepted proximity reflects the most efficient data transmission in the ad hoc network. Also, it enforces security in the system to make sure far or imposter devices are not able to authenticate. Figure 1 shows how authentication works in the proposed system.

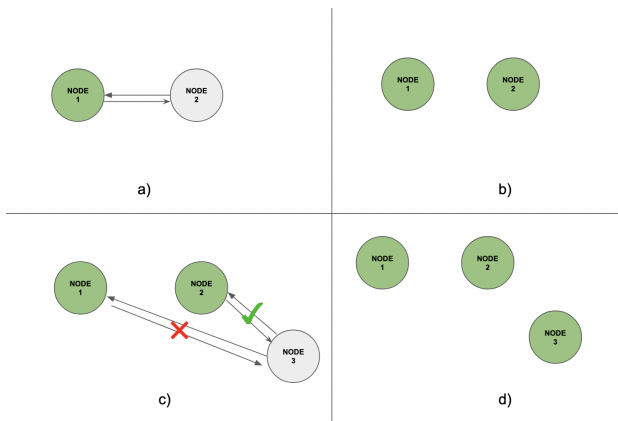


Fig. 1. Authenticating new IoT nodes

As can be seen in Figure1 (a). Node 1 is an authenticated node in the system. The color green denotes an authenticated node. Node2 enters the system and wants to be authenticated. It scans for Wi-Fi Beacon and broadcasts this data in the request to join the network. Node1 received this request along with the scan data. Node1 verifies whether Node2 meets the proximity threshold by calculating the Euclidean distance and comparing the distance with the pre-defined threshold. Figure1(b) shows that the authentication is successful. In Figure1 (c), a Node 3 enters the area. Node3 is more proximate to Node2 than Node1. Node 3 broadcasts the request to join the network along with the Wi-Fi scan data, which is received by Node1 and Node2. Both Node1 and Node 2 verify whether Node3 meets the proximity threshold in reference to these nodes. Figure1(d) shows that it met the threshold in reference to Node2 but not Node1. For this case, the new device meet the threshold with two devices. It connects to the device with the least Euclidean distance.

VI. EXPERIMENTS

To test the proposed system, two experiments were conducted. Experiment 1 calculates the accuracy of the proposed proximity-based authentication. Experiment 2 simulates the system operation with several nodes.

A. Two-device proximity authentication

In this experiment, two Raspberry Pis were placed within two meters of each other. Each Raspberry Pi scanned and collected the Wi-Fi beacon frames and RSSI in the area following equation 1 and equation 2. The collected data was used to calculate the value of the threshold using equation 3. In the experiment, the ten access points with the highest RSSIs were used in the equation. To test the proposed system, the two Raspberry Pis were placed at ten different locations around the building. Ten authentications were attempted at each location. Five of these attempts were conducted when the two devices were less than two meters apart. Then, five more times where the two devices are more than two meters apart. The data was collected for each attempt and the Euclidean distance was calculated using equation3. This distance was compared to the threshold to determine successful and failing authentications. The results of the experiment are shown in Table I.

TABLE I
SUCCESS RATE

N = 100	Actual	
	Success	Failure
True	TS = 45.54%	TF = 42.36%
False	FS = 4.46%	FF = 7.64%

The experiment returned authentication accuracy of 87.9%. The accuracy can be increased if the tolerance of the threshold value is increased. When 20% tolerance is considered, the accuracy jumps to 94.5%.

B. Several nodes simulation

This experiment is to simulate the system’s operations when several nodes are involved. The simulation was conducted utilizing python. The threshold calculated in the experiment above was considered. A function was written to perform the authentication following the steps above. The nodes in the simulation were configured with these Actual RSSI values at ten different locations collected in the experiment above. The simulation had each node to authenticate with the other through an iteration. Each node in the simulation would attempt to connect to the node where there are the nodes that meet the threshold and with the least Euclidean distance. In the simulation, each device is connected to a node with the least distance, as shown in figure 2. The experiment returned authentication accuracy of 90%, see Table II.

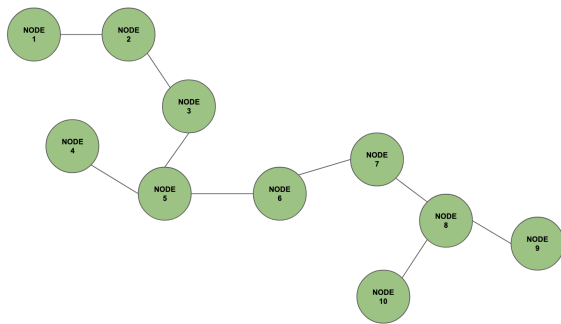


Fig. 2. Experiment 2: Ten nodes simulation

TABLE II
SUCCESS RATE

		Actual	
N = 10		Success	Failure
True		TS = 5	TF = 1
False		FS = 0	FF = 3

VII. SYSTEM ANALYSIS

A. Continuous Authentication

Continuous authentication is a feature where the authentication process every predefined time. The service provider will check continuously if the two communicated IoT devices are in the same IoT network in order to maintain the session. The frequency with which the IoT devices collect and send the new data (SSID, BSSID, RSSI values) can be varied based on requirements and/or an administration’s desire. The continuous authentication will check if both communicated IoT devices are still in the same IoT environment in order to keep the session alive. If not, the session can be automatically terminated. This feature could mitigate and, at some level, prevent

some types of cyber-attacks (more details are provided in the next section).

B. Usability

Due to the proposed scheme using the existing IEEE 802.11 infrastructure (i.e., Wi-Fi access points already in the IoT environment, network interface cards already in IoT devices), no new hardware is required to be installed. Moreover, the proposed scheme is considered to be readily applicable in a scalable manner because it relies on ubiquitous Wi-Fi access points. The internet can be found almost everywhere people live [18]. Due to its ubiquitous nature, the internet is an essential and robust platform for education, business, and entertainment. It has been noted that locations with reliable internet connectivity are also where access points are commonly established [19].

VIII. SECURITY ANALYSIS

A. Environment Simulation Attack

A chance of simulating the IoT Environment is possible. The attacker scans the IoT environment (i.e., beacon frame characteristics and RSSI value) and then replicates it elsewhere, where the attacker has full control of the simulated environment. In the proposed research, the IoT environment can be scanned and replicated; however, the attacker’s IoT device’s unique identifier (e.g., IMEI, UUID, MAC address, etc.) will not match the IoT’s unique identifier in the database. Moreover, the administration will be notified because the authentication entity will reject the request. Also, The continuous authentication feature will mitigate and, at some level, will prevent the attack.

B. Insider attacks

An attacker can be inside the IoT environment, which means he/she will be able to scan the IoT environment (i.e., beacon frame characteristics and RSSI value) and then utilizes it. Beginning inside the IoT environment and scanning it, is something easy to obtain. However, the continuous authentication feature will mitigate and, at some level, will prevent the attack. In addition, the service provider will notice that the malicious IoT device’s unique identifier does not match the one in its database, which results in rejecting the request and notifying the administrator.

IX. CONCLUSION

IoT is one of many buzzwords in Information Technology (IT), which will transform our daily lives into smart systems. To guarantee the security of the wireless communications in an IoT environment, devices must build trust in the identity of each other, *authentication*. This paper proposes a technique to authenticate IoT devices in ad hoc networks to verify proximity. This is done in a way that only devices within a certain distance from other authenticated IoT devices will be able to

connect to the network. Meanwhile, devices that are far from an authenticated device or not physically in the area will fail in the proximity authentication. The proposed system enforces security in ad hoc IoT networks. Also, it figures the more suitable device to connect to in an ad hoc network that would reflect the most suitable Radio frequency conditions to communicate. The experiment showed an adequate accuracy of proximity authentication that can be increased with configuring the tolerance in the threshold.

REFERENCES

[1] C. Hodson, *Cyber Risk Management: Prioritize Threats, Identify Vulnerabilities and Apply Controls*. Kogan Page, 2019. [Online]. Available: <https://books.google.com/books?id=yuWYDwAAQBAJ>

[2] M. Liyanage, A. Braeken, P. Kumar, and M. Ylianttila, *IoT Security: Advances in Authentication*. Wiley, 2020. [Online]. Available: https://books.google.com/books?id=Bk6_DwAAQBAJ

[3] A. Braeken, P. Kumar, M. Ylianttila, and M. Liyanage, *IoT Security: Advances in Authentication*. Wiley, 2019. [Online]. Available: <https://books.google.com/books?id=DdHGDwAAQBAJ>

[4] Y. Al-Hadhrani and F. K. Hussain, "Ddos attacks in iot networks: a comprehensive systematic literature review," *World Wide Web*, vol. 24, no. 3, pp. 971–1001, 2021.

[5] H. Takamizawa and K. Kaijiri, "Reliable authentication method by using cellular phones in wbt," 05 2006, pp. 200 – 200.

[6] H. Alamlah and A. A. S. AlQahtani, "A cheat-proof system to validate gps location data," in *2020 IEEE International Conference on Electro Information Technology (EIT)*, 2020, pp. 190–193.

[7] A. A. S. AlQahtani, Z. El-Awadi, and M. Min, "A survey on user authentication factors," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2021, pp. 0323–0328.

[8] G. Mao, B. Fidan, and B. D. Anderson, "Wireless sensor network localization techniques," *Computer Networks*, vol. 51, no. 10, pp. 2529–2553, 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389128606003227>

[9] W. Jansen and V. Korolev, "A location-based mechanism for mobile device security," in *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 1, 2009, pp. 99–104.

[10] C. Y. Leong, T. Perumal, K. W. Peng, and R. Yaakob, "Enabling indoor localization with internet of things (iot)," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, 2018, pp. 571–573.

[11] A. Honnef, E. Sawall, M. Mohamed, A. A. S. AlQahtani, and T. Alshayeb, "Zero-effort indoor continuous social distancing monitoring system," in *2021 IEEE World AI IoT Congress (AllIoT)*, 2021, pp. 0482–0485.

[12] E. Sawall, A. Honnef, M. Mohamed, A. A. S. AlQahtani, and T. Alshayeb, "Covid-19 zero-interaction school attendance system," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–4.

[13] A. A. S. AlQahtani, "0e2fa: Zero effort two-factor authentication," *Louisiana Tech University*, 2020.

[14] Y. Agata, J. Hong, and T. Ohtsuki, "Room-level proximity detection using beacon frame from multiple access points," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 941–945.

[15] C. Wullems, M. Looi, and A. Clark, "Enhancing the security of internet applications using location: a new model for tamper-resistant gsm location," in *Proceedings of the Eighth IEEE Symposium on Computers and Communications. ISCC 2003*, 2003, pp. 1251–1258 vol.2.

[16] T. Guelzim and M. Obaidat, "Novel neurocomputing-based scheme to authenticate wlan users employing distance proximity threshold." 01 2008, pp. 145–153.

[17] H. Garg and M. Dave, "Securing iot devices and securely connecting the dots using rest api and middleware," in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, 2019, pp. 1–6.

[18] M. M. Singh, K. Adzman, and R. Hassan, "Near field communication (nfc) technology security vulnerabilities and countermeasures," *International Journal of Engineering & Technology*, vol. 7, no. 4.31, pp. 298–305, 2018.

[19] X. Zhu, S. Yu, and Q. Pei, "Quickauth: Two-factor quick authentication based on ambient sound," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

Dynamic Modeling of a Micro Solar Electric Vehicle for Pakistan using Simulink

Ali Husnain

Department of Electrical Engineering
Memorial University of Newfoundland
St. John's, NL, Canada.
ahusnain@mun.ca

M. Tariq Iqbal

Department of Electrical Engineering,
Memorial University of Newfoundland,
St. John's, NL, Canada.
tariq@mun.ca

Abstract— While the global electric vehicle (EV) adoption is on rise, developing countries like Pakistan have been facing many obstacles in the face of EV adoption such as shortfall of electricity, high EV prices, low average income and absence of commercial and residential charging infrastructure. In this paper, we propose a design of a micro solar electric vehicle, which would help overcome all of these problems and provide an economical yet feasible solution. The design and system sizing of the micro solar EV was done in HOMER Pro. A dynamic model of the micro solar EV was created in MATLAB/Simulink, which implemented PV generation, maximum power point tracking, battery charging and discharging, DC-motor operation and speed control of the electric vehicle, while taking into consideration environmental factors like irradiance and temperature.

Keywords— solar EV, solar car, electric vehicle, EV design, dynamic modeling

I. INTRODUCTION

Climate change is one of the most pressing issues of the modern world and steps are being taken around the world to help slow down and mitigate climate change. The use of electric vehicles is one such bold step that helps reduce the GHG emissions of the global transport sector and slow down the impact of climate change. Global EV adoption has been growing at a rapid pace, in developed countries. However, the same cannot be said about developing countries, such as Pakistan.

The EV adoption in Pakistan is facing many obstacles including but not limited to electricity shortfall, high prices of electric vehicles, low average income, absence of commercial charging infrastructure and the inability of residential electrical infrastructure to charge an electric vehicle. At typical electrical service is 5A or 10A at 220V to supply house load, which is not enough to charge an EV at home.

Therefore, in this paper, we propose the design of a micro solar electric vehicle, which holds against all the above-mentioned challenges. This solar EV would have the ability to take power from the sun in the form of solar energy, throughout the daytime. It would also be very economical in price and would easily be charged from the existing residential electrical infrastructure.

II. LITERATURE REVIEW

While it has roughly been less than a decade since EV commercialization has become mainstream and automakers are selling fully functional EVs to consumers. There has been

plenty of research on EVs before that as well. During the literature review part of this research we came across multiple designs and dynamic models of all types of EVs from micro electric vehicles, to passenger sedans, vans, buses and other commercial vehicles.

Qingqing Xie et al. present a MATLAB/Simulink model for advanced vehicle dynamic model for EV emulation, while taking into consideration actual environmental conditions. The model also takes into account the rotational inertia of each component involved in developing the model [1]. Another research published in University of Washington discusses the dynamics of an electric vehicle by preparing a model which uses enhanced techniques such as torque vectoring [2]. Similarly, another research uses modelica methodologies to model EV pickup attributes to study the dynamic response of an electric vehicle [3]. Other such models for battery dynamics and solar hybrid vehicles have been given in [4] and [5] respectively. In conclusion, there is enough research material that can be found for developing a model to study the vehicle dynamics of an electric vehicle.

Talking about solar electric vehicles, there is limited research present on the design of solar electric vehicles, let alone, dynamic models of solar electric vehicles.

As outlined in [6] the research on solar electric vehicles can be divided into three major segments a) commercial solar cars, b) solar electric vehicles for solar races and c) small solar cars. Research groups and universities from all around the world take part in solar car design for solar races such as World Solar Challenge [7] and American Solar Challenge [8].

Talking about the second segment of solar electric vehicles, there are multiple automakers, who have undertaken the task of designing and selling commercial solar electric vehicles [9] [10] [11]. Finally some interesting designs have come forward for the small solar electric vehicle segment as well [12] [13].

However, there exists a huge gap, when it comes to developing and studying a dynamic model of a solar electric vehicle. In our previous research, we proposed a design for a micro-solar electric vehicle for application in Pakistan [6]. In this paper, we are going to develop a dynamic model of our micro-solar electric vehicle using MATLAB/Simulink.

III. DESIGN OF MICRO SOLAR EV

Before we talk about the dynamic model of our micro solar electric vehicle, we will briefly discuss the system design and PV sizing. The specifications of the vehicle used in the design are as following;

Table 1: Vehicle Specifications used for the design [6]

Item	Description
Max. Range	100 km
Dimensions	2450 mm*1350 mm*1750 mm
Wheelbase	1500 mm
No. of seats	3
Motor Size	1200 W
Total Motor Torque	100 Nm
Max Speed	43 km/hr

The design and PV sizing was carried out in HOMER Pro and the system can be seen in the figure below:

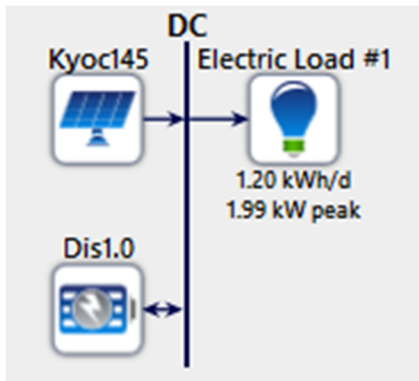


Figure 1: System Design of micro-solar electric vehicle in HOMER Pro [6]

The sizing of the system was as follows:

Table 2: System sizing carried out in HOMER Pro [6]

Item	Quantity
Solar Panels/ Advance Solar Hydro Wind Power API – 150 W	3
24V, 1kWhr Li-Ion battery, 72V bus	3
Cost/NPC (Rs)	177013.7
Cost/COE (Rs)	54.62492
Cost/Operating cost (Rs/yr)	961.2305
Cost/Initial capital (Rs)	184128
System/Ren Frac (%)	100
Kyoc145/Capital Cost (Rs)	21000
Kyoc145/Production (kWh/yr)	766.9749
Dis1.0/Autonomy (hr)	57.6
Dis1.0/Annual Throughput (kWh/yr)	392.1335
Dis1.0/Nominal Capacity (kWh)	2.88

Based on this system design and sizing, we have developed a detailed dynamic model of our solar electric vehicle.

IV. DYNAMIC MODELING OF MICRO SOLAR ELECTRIC VEHICLE IN MATLAB/SIMULINK

Electrical dynamic modeling is an important part of research and helps us understand how a system would be working in real-life. A dynamic model of a PV system would not only help us understand how the system would perform, but also how it would reach to the changes in the environment. This dynamic model would also help us simulate the system in various conditions that an electric vehicle has to go through during its course of operation.

Figure 2 shows a block diagram of the dynamic model of our system. As shown in the figure, our system consists of seven major blocks, which are further made of tens of components, all coming together to form a dynamic model of our micro-solar electric vehicle. The first block i.e. PV arrays generate power by converting solar energy into electrical energy, which is then supplied to the DC-DC converter. This converter is then connected to the battery storage block of our system, which not only stores this power, but also supplies to the DC motor, when in operation. The battery storage block can also be charged by an external AC source i.e. the electrical grid. There are two control blocks used in our system, one is an MPPT controller, which helps the system extract the maximum amount of power from the PV panels, and the second one is a speed controller, which helps the DC machine operate at desired speeds, as per the speed reference provided.

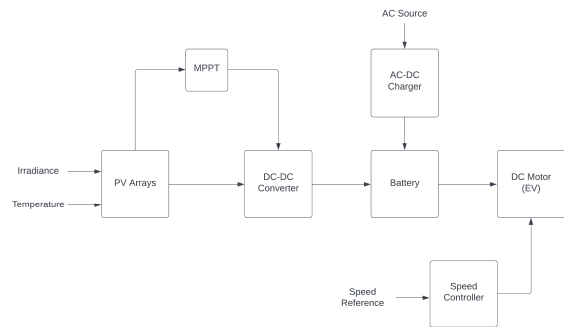


Figure 2: Block Diagram of the dynamic model

A. PV Array

PV arrays are the most fundamental part of our dynamic model, as we are developing a micro solar electric vehicle, meaning our focus is on extracting the most amount of power from the PV array, given the available space to mount PV panels. It is important to note that the non-linear output of PV modules is affected by environmental parameters such as irradiance, temperature, clearness index, dust, cloud and other shading effects etc. However, in this model, we are only taking into account irradiance and temperature.

A PV array can be made up of multiple PV modules, which in turn consist of series of PV cells. To better understand the performance of a PV module, we will take a look at the equivalent circuit of a PV module as show in figure 3.

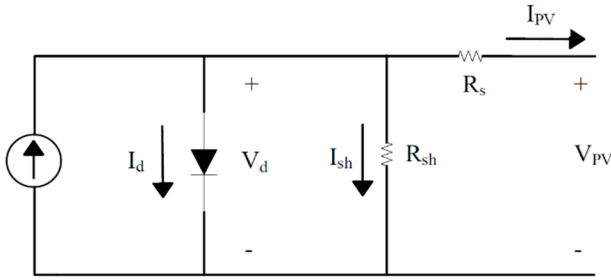


Figure 3: Equivalent circuit of a PV module

This equivalent circuit gives us some basic equations to help calculate the values of current and voltages under consideration.

$$I_d = I_o \left[\exp\left(\frac{V_d}{V_T}\right) - 1 \right] \quad (1)$$

$$V_T = \frac{KT}{q} \times A \times N_{cell} \quad (2)$$

$$V_d = V_{PV} + R_S I_{PV} \quad (3)$$

$$I_{PV} = I_{ph} - I_D - \frac{V_D}{R_{sh}} \quad (4)$$

In the above equations Eq. (1) to Eq. (4) [14], I_d is the diode current, I_{ph} is the light-generated current, I_o is the diode saturation current, V_T is thermal voltage equivalent, V_d is the diode voltage, V_{pv} and I_{pv} are module voltage and current, respectively, K is the Boltzman constant equal to $1.3806 \times 10^{-23} \text{J/K}$, A is the diode ideality factor, T is the cell temperature, q is the electron charge equal to $1.602 \times 10^{-19} \text{C}$, and N_{cell} is the number of cells connected in series in a module. As evident, from equations 1 to 4, almost every parameter of the PV module is subject to change as per the variations in temperature and irradiance.

Based on the results from HOMER Pro [6] the model consisted of three 24V, 150W solar PV (Advance Solar Hydro Wind Power API – 150) modules connected in series, making the bus voltage to be 72V.

B. Maximum Power Point Tracking (MPPT) Controller

As discussed before, the performance of a PV module is very much dependent on environmental factors like irradiance and temperature. For each PV module, there is a point in the P-V curve, where the PV module is giving the maximum amount of power. This point is called Maximum power point, and it can be different for different condition. This can be observed in figure 4, where two different temperatures result in two different MPPs.

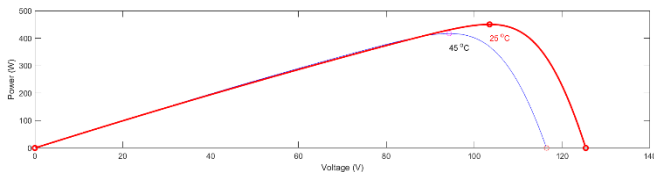


Figure 4: PV Curve of our chosen PV Module

Maximum power point tracking is a process, controls the power generation from a PV panel in a manner that it is always operating at maximum power point. There are numerous techniques to achieve maximum power point tracking such as incremental conductance method, perturb and observe (P&O) method etc. In this model, we are going to use the perturb and observe (P&O) method, in order to implement maximum power point tracking.

The way P&O technique is used is that a minor perturbation is introduced into the system, which causes the power of the PV module to vary. After the perturbation is introduced, the algorithm then observes the output power periodically and compares it with previous power. A similar perturbation is introduced to vary the voltage of the PV array. By doing so, the algorithm identifies, where the system is lying on the PV curve. Depending on that, the duty cycle is either increased or decreased.

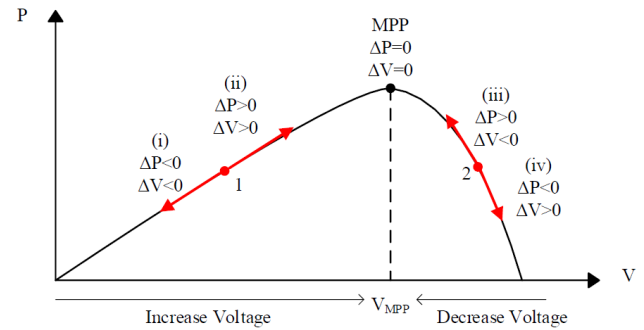


Figure 5: P&O algorithm working principle [15]

As seen in the figure 5, our system can have two operation points. The first one is indicated as point 1 and the second one is indicated as point 2 on the graph. In the case that the system was on point 1, the voltage needs to be increased. Similarly, if it was on point 2, the voltage need to be decreased. In order to implement this operation, we used a code-based function, according to the flowchart shown in figure 6.

The MPPT starts by measuring the current and voltage of the PV modules using sensors, and outputs a duty cycle. This duty cycle is going to be used to generate a PWM signal, which will enable or disable the switch used in our DC-DC converter. A change in the duty cycle generated by the MPPT means that the voltage of the module would change and subsequently, the power of the PV module would also change. The P&O algorithm can be best described by the flowchart given in figure 6.

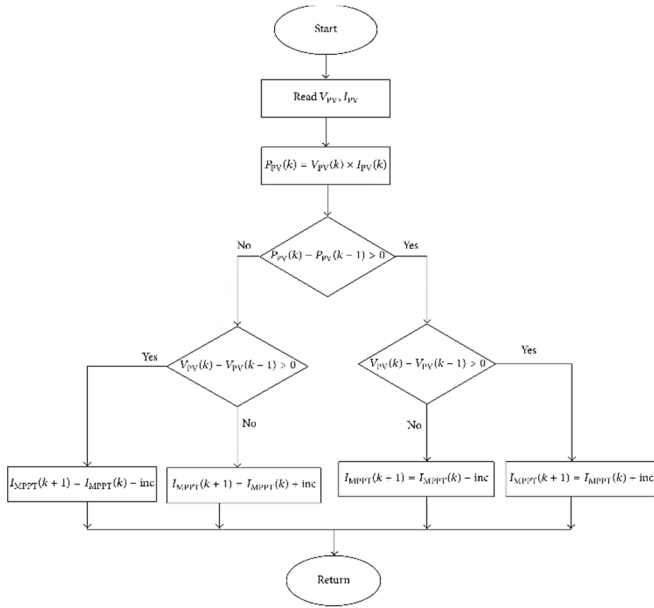


Figure 6: P&O Algorithm flowchart

C. DC-DC Converter

As evident from the name, DC-DC converters are used where both the input voltage and output voltage are in DC. The main purpose of using a DC-DC converter is to produce a regulated voltage from an uncontrolled source to a load that may or may not be constant, in a manner that is efficient. A DC-DC converter is a high-frequency power conversion circuit, which makes use of high-frequency switching, transformers, inductors, and capacitors.

There are many different topologies of DC-DC converter including buck, boost, buck-boost and SEPIC converters. A buck converter is used in applications where the output voltage needs to be lower than the input voltage, boost converter is used when output voltage needs to be higher than input voltage, whereas, buck-boost and SEPIC can be used in both situations. In this model, we are going to use a buck converter since the input voltage or PV module voltage in our case is higher than what is needed at the battery terminal. Figure 7 shows the circuit of a buck converter.

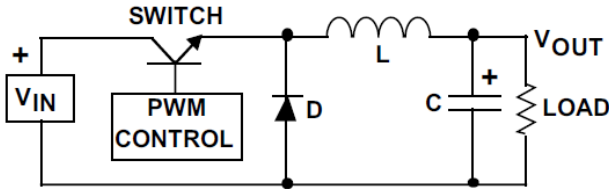


Figure 7: DC-DC Converter (Buck Converter)

The buck converter operates in two steps. The first step of operation is when the switch is turned ON. During this step, the current is flowing to the output capacitor, thus charging it up. Since the voltage of a capacitor does not rise instantly, as well as the inductor limiting the charging current, the capacitor voltage is not the full voltage of the source during the switching

cycle [16]. This is depicted in figure 8, the highlighted part of the circuit shows the flow of current.

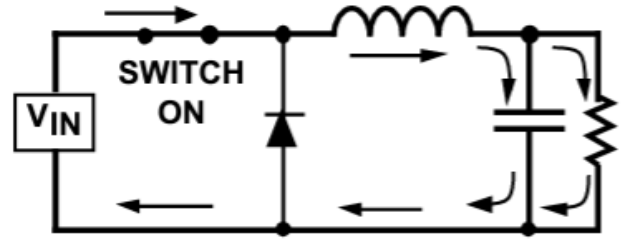


Figure 8: Buck Converter Operation (Switch ON)

During step two, the switch turns off and a voltage is created across the inductor due to the fact that inductor cannot change current suddenly. This voltage allows the capacitor to charge and power the load when switch is off [16]. This can be seen in figure 9.

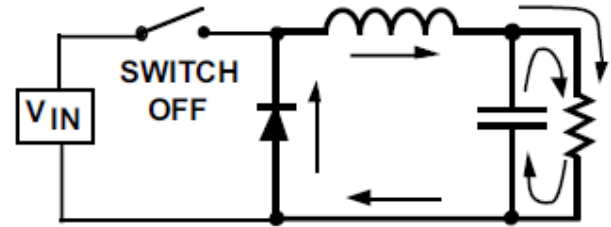


Figure 9: Buck Converter Operation (Switch OFF)

Now in order to design the buck converter, we need to calculate the values of the components used to form the buck converter [17]. Following are the mathematical equations that help design a buck converter.

$$D_{max} = \frac{V_{OUT}}{V_{IN(max)} \times \eta} \quad (5)$$

Where, $V_{IN(max)}$ = maximum input voltage

V_{OUT} = output voltage

η = efficiency of the converter

$$L = \frac{V_{OUT} \times (V_{IN} - V_{OUT})}{\Delta I_L \times f_s \times V_{IN}} \quad (6)$$

Where, f_s = minimum switching frequency of the converter

ΔI_L = estimated inductor ripple current

$$\Delta I_L = (0.1 \text{ to } 0.4) \times I_{OUT(max)} \quad (7)$$

$$C_{OUT} = \frac{\Delta I_L}{8 \times f_s \times \Delta V_{OUT}} \quad (8)$$

Where, C_{OUT} = minimum output capacitance

ΔV_{OUT} = desired output voltage ripple

$$\Delta V_{OUT} = ESR \times \Delta I_L \quad (9)$$

Where, ESR = equivalent series resistance of the used output capacitor [17].

The buck converter design is complete once we have calculated the values for the inductance and the capacitance to be used.

D. Li-Ion Battery

For the battery storage system of our PV system, we used a Li-Ion battery. The model present in Simulink is called a generic battery model, which allows you to mimic the charging and discharging characteristics of any rechargeable battery. Figure 10, shows the equivalent circuit of the rechargeable battery model [18].

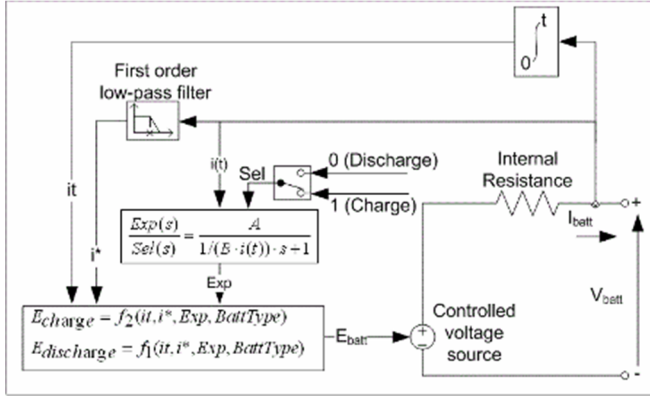


Figure 10: Equivalent circuit of a rechargeable battery model

The battery model has three observable parameters i.e. battery voltage, battery current and SOC. As evident from the name, battery voltage is the rated voltage of battery. A positive battery current represents that the battery is discharging, whereas a negative battery current represents that the battery is charging. Similarly, SOC is the battery state of charge, and gives information about the state of charge of the battery at any given moment. An increasing SOC represents battery charging and a decreasing SOC represents battery discharging.

E. AC-DC Charger

In addition to the battery being able to charge from the PV panels, the micro solar electric vehicle would also have the ability to charge from an AC source, if needed. This represents the time of the day when there is no sun, and the vehicle needs charging. In such a case, the car battery could be directly plugged into an AC wall outlet. This is where the AC-DC charging block comes in handy. Figure 11, shows the equivalent circuit of an AC-DC charging block [19].

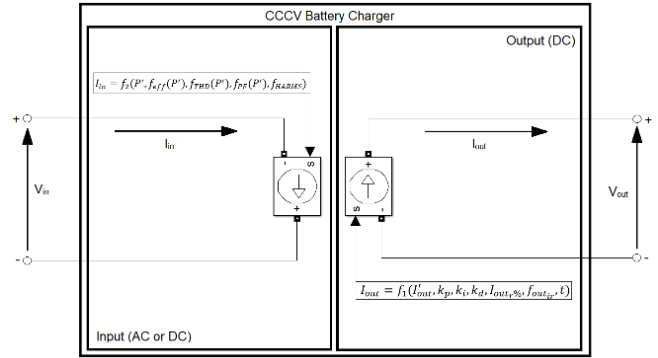


Figure 11: AC-DC Charging Block

F. DC Machine

Since our micro solar electric vehicle would be running a DC motor, in order to develop the model of such a motor, we have used the DC machine block in Simulink. The DC machine block allows us to implement both a wound-field or permanent DC machine. However, in our case, we are using a wound-field DC machine in series configuration.

The armature circuit of the DC machine consists of an inductor L_a and resistor R_a in series with a counter-electromotive force (CEMF) E , which is proportional to the machine speed.

$$E = K_E \omega \quad (10)$$

Where, K_E is the voltage constant and ω is the machine speed.

$$\text{Electromechanical Torque} = T_e = K_T I_\omega \quad (11)$$

Where, K_T is the torque constant.

There is a mechanical part of the DC machine, which is used to calculate the speed of the machine from the net torque applied to the rotor.

$$J \frac{d\omega}{dt} = T_e - T_L - B_m \omega - T_f \quad (12)$$

Where, J = inertia,

B_m = viscous friction coefficient, and

T_f = Coulomb friction torque.

G. Speed Controller

The last block of our model is a speed controller, which is being used in order to regulate the speed of the motor. The speed controller is a simple PWM Chopper circuit. Figure 12, shows the model of the circuit

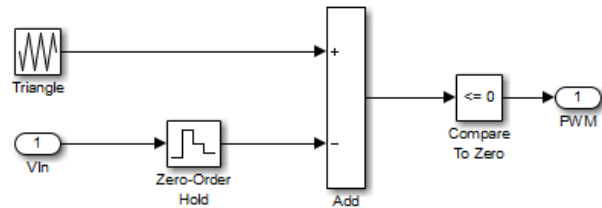


Figure 12: Voltage to PWM Signal Converter Block

By taking into account the reference voltage set on the input side of this circuit, it generates a PWM signal, which operates a high-frequency switch thus changing the average value of current supplied to the motor, which results in an increase or decrease in the motor speed as required.

V. SIMULATION RESULTS AND DISCUSSION

A dynamic model of the block diagram shown in figure 2, was implemented in Simulink, which consisted of all the major blocks discussed in the previous section. The detailed model can be seen in figure 13.

A. PV Operation

Now let us take a look at the results of each part of the simulation. First of all we have the PV array, the output of which is very much dependent on the irradiance values. We are going to introduce a variable irradiance curve, to see how PV output varies accordingly. Figure 14, shows the changing irradiance values. At these values, the PV is generating the amount of current and voltage shown in figure 15. Notice, how the voltage and current values change with the irradiance values.

B. Battery Operation

Now let us move on the battery parameters, when PV modules are in operation and actively generating power, this power is being directly transferred and deposited into the battery. This is being demonstrated in figure 16. Notice that the SOC of the battery is increasing over the course of the simulation, and the battery current is in negative, which shows the influx of current into the battery.

Similarly, when the PV is disconnected due to any reason including bad weather or absence of sun during night time, the PV panels will not be producing any power. Hence, the EV

motor will be drawing power from the battery. If we look at figure 17, we can see that the SOC of the battery is decreasing and the battery current is also positive, which shows the outflow of current from the battery.

In the last scenario, we will be charging the battery directly from the AC source, through the CCCV charging block. In this scenario, once again the battery SOC is positive and the battery current is in negative as shown in figure 18.

C. DC-Machine Operation

In this section, we are going to observe the operation of the DC machine, which represents the electric vehicle motor. The DC machine operation can be observed in two different modes.

The first one is the normal mode, in which the DC machine is drawing power from the battery and generating a normal response. This can be seen in figure 19, from the armature current, field current and electrical torque values.

The second mode of operation is when the DC machine is put in reverse, which represents the electric vehicle moving in reverse direction. In a DC series machine, in order to reverse the direction of motion, we have to change the polarity of field windings. In this model we have achieved this by using ideal relay modules, which can be switched at any instant at the start of, or during the simulation to reverse the direction of the vehicle. Figure 20 shows the vehicle direction changing at the 5 second mark during the simulation. This change is denoted by a change in the direction of field current.

D. DC-Machine Operation under variable load

In order to understand the response of our electric vehicle over uneven or steep roads, we are going to run the simulation for 60 seconds, introduce a variable load profile and see how that affects the dynamics of our electric vehicle. Figure 21, shows the dynamic load profile used in the simulation. Figure

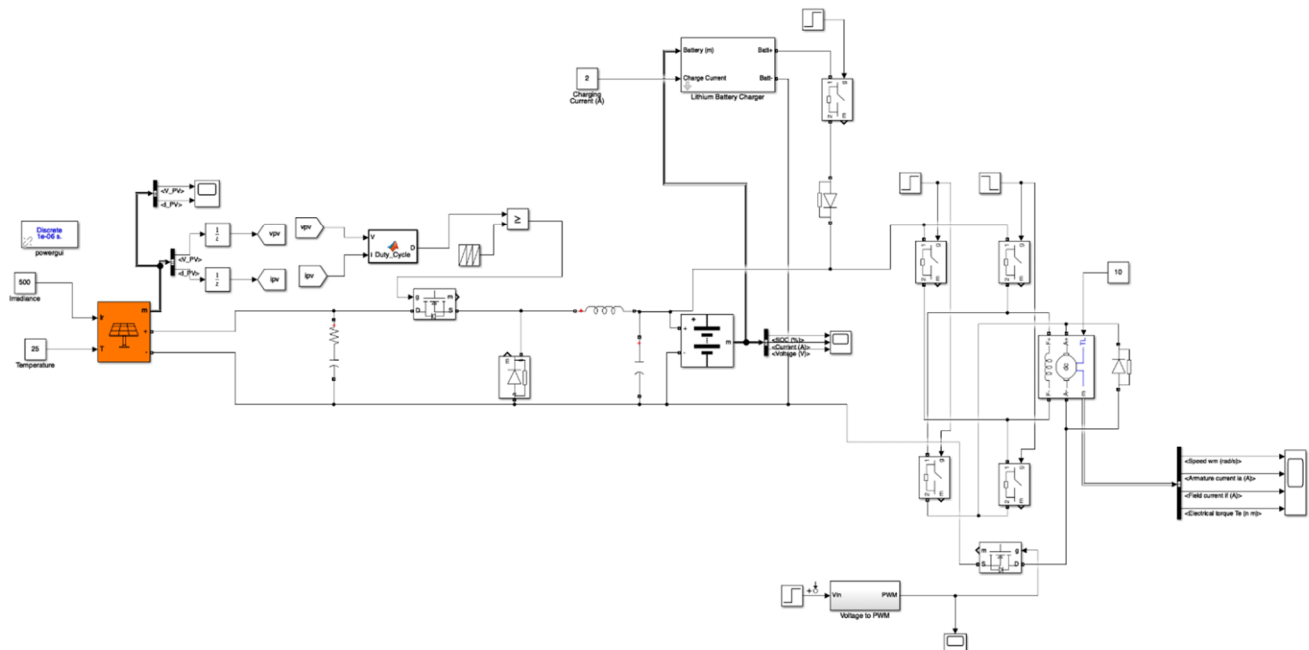


Figure 13: Simulink Model of the proposed micro solar electric vehicle

22, shows DC Machine dynamics under that changing load profile. The electrical torque and armature current of the DC-machine varies accordingly to keep up with the uneven roads.

E. Speed Control

Finally, we are going to observe how our speed control circuit works. In this case, we are running the simulation for 60 seconds, and we have introduced a signal which changes the speed reference every 10 seconds. Figure 23, shows the changing speed reference. Figure 24, shows the dynamic response of the DC-machine to keep up with the changing speed reference.

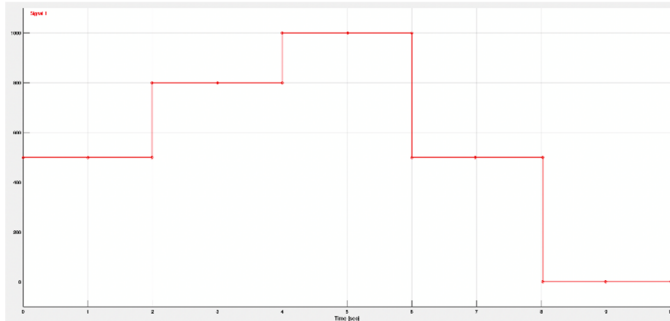


Figure 14: Variable Irradiance Values to measure PV Response

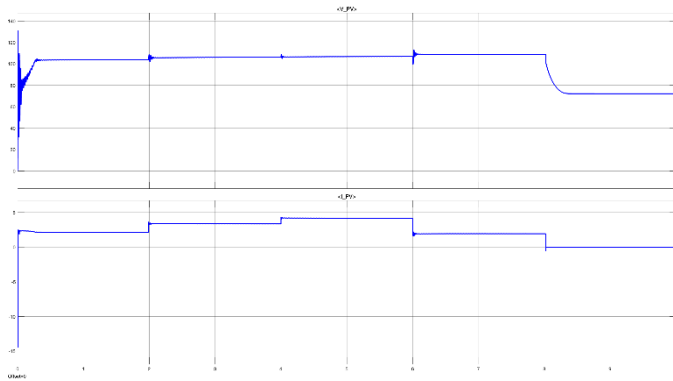


Figure 15: PV Voltage and Current variance as per changing Irradiance

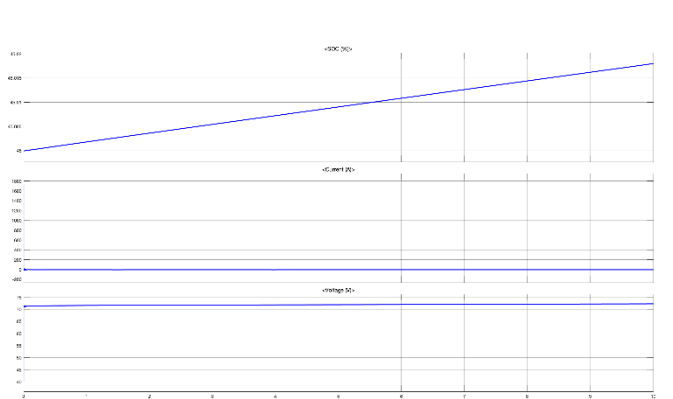


Figure 16: Battery Charging Mode (From PV Modules)

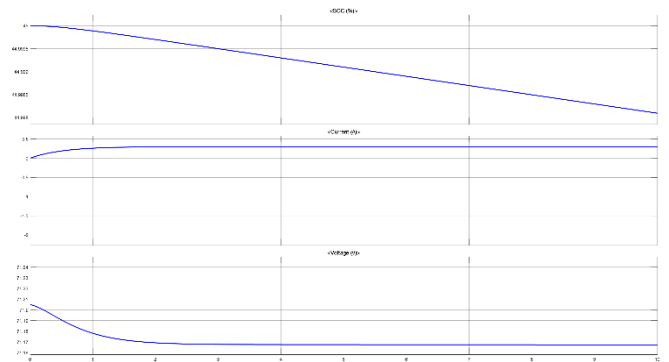


Figure 17: Battery Discharging Mode (During Motor Operation)

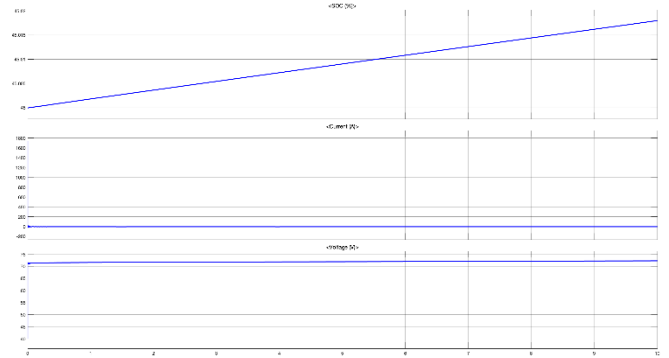


Figure 18: Battery Charging Mode (From AC Source)

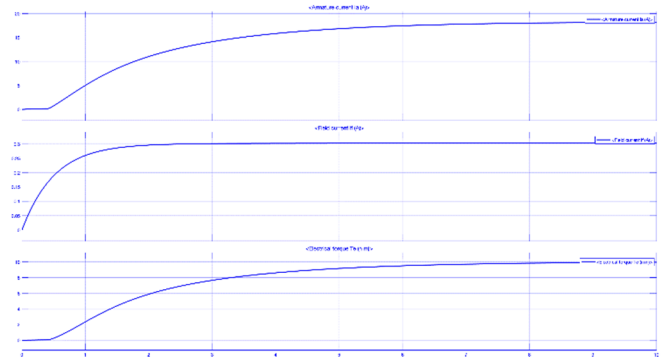


Figure 19: DC Machine Normal Operation

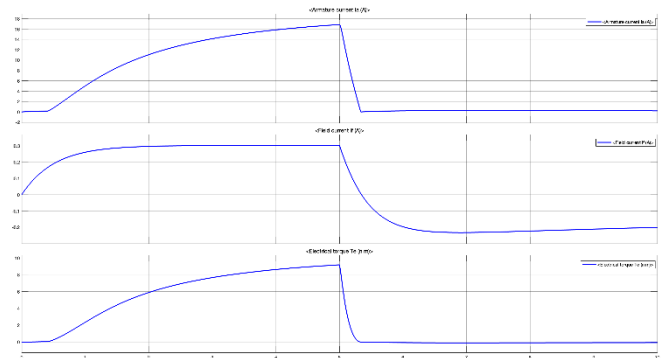


Figure 20: DC Machine Forward and Reverse Operation

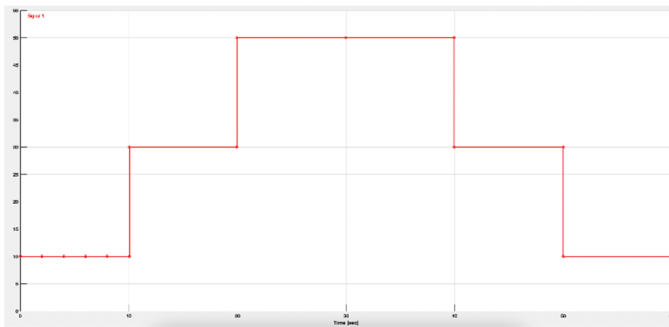


Figure 21: Variable Load Profile for DC Machine

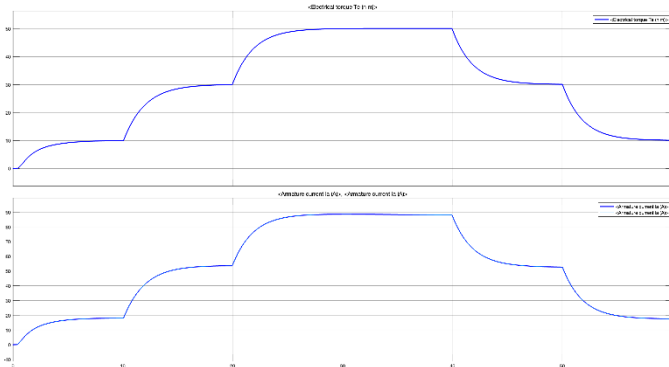


Figure 22: DC Machine dynamics under variable load

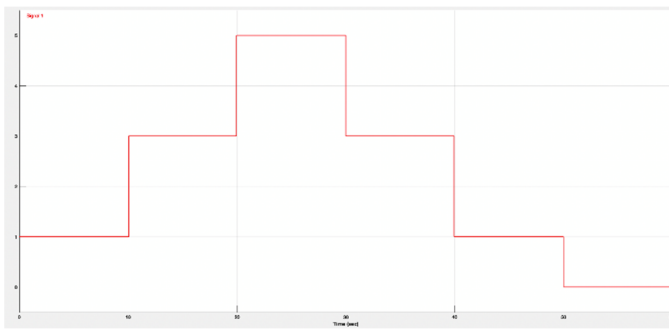


Figure 23: Variable Speed reference for DC Machine

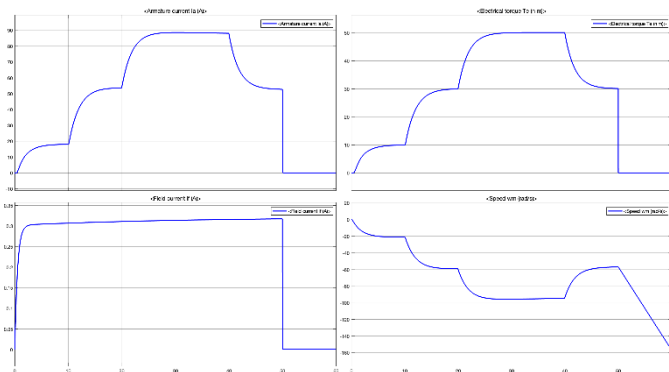


Figure 24: DC Machine dynamics under variable speed reference

VI. CONCLUSION

A novel design for a micro solar electric vehicle for application in Pakistan has been presented in this paper. Not only does this micro solar electric vehicle provide the basis for the design of a new class of cars, but it also helps solve almost

all of the problems that are currently slowing the growth of EV market share in Pakistan and other developing countries.

The said micro solar electric vehicle was designed in HOMER Pro, and a system sizing was done. According to this system sizing, the micro solar electric vehicle would require 3, 24V, 150W PV modules (Advance Solar Hydro Wind Power API – 150) to add a range of approximately 30 KMs to the pre-existing battery range of the electric vehicle.

The electric vehicle can also be charged from a 220V AC outlet with the help of the onboard AC-DC charging block, which allows the electric vehicle to charge at extremely low currents, easily supported by the current residential electricity infrastructure in Pakistan.

In addition, a dynamic model of the micro solar electric vehicle was developed to be studied in much more details. The dynamic model focused on three different aspects of the systems i.e. power generation by PV modules that are to be mounted on top of the vehicle, battery charging from PV modules as well as an external AC source, and finally DC-motor operations in forward and reverse mode. Other notable features of the model include a maximum point power tracker and a PWM chopper based speed control circuit.

The dynamic model of the system not only helped us fully understand the working of the system in different modes, conditions and situations, but also helped us understand how the system will react to any variations in the environmental factors such as irradiance and temperature etc.

REFERENCES

- [1] Q. Xie, C. H. L. Filho, G. Feng, W. Clandfield, and N. C. Kar, "Advanced vehicle dynamic model for EV emulation considering environment conditions," in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Apr. 2017, pp. 1–4. doi: 10.1109/CCECE.2017.7946794.
- [2] L. Prange, "Vehicle Dynamics Modeling for Electric Vehicles," Thesis, 2017. Accessed: May 04, 2022. [Online]. Available: <https://digital.lib.washington.edu:443/researchworks/handle/1773/40251>
- [3] "Electric Vehicle Dynamics Simulations: Modelica Methodologies," *Modelon*, May 28, 2020. <https://www.modelon.com/electric-vehicle-dynamic-simulation/> (accessed May 04, 2022).
- [4] O. Tremblay, "Experimental Validation of a Battery Dynamic Model for EV Applications," *E'cole de Technologie Supérieure*, Stavanger, Norway, 2009.
- [5] L. Bai, Y. Zhang, H. Wei, J. Dong, and W. Tian, "Digital Twin Modeling of a Solar Car Based on the Hybrid Model Method with Data-Driven and Mechanistic," *Applied Sciences*, 2021, doi: 10.3390/APP11146399.
- [6] A. Husnain, "System Design and PV Sizing of a Micro Solar Electric Vehicle for Pakistan," presented at the 9th IEEE Conference on Technologies for Sustainability (SusTech 2022), Sunny Riverside, CA, Apr. 2022.

- [7] "About | Bridgestone World Solar Challenge," Feb. 09, 2021. <https://worldsolarchallenge.org/about> (accessed Sep. 30, 2021).
- [8] "American Solar Challenge | Organizing solar car racing in North America." <https://www.americansolarchallenge.org/> (accessed Sep. 30, 2021).
- [9] "Vehicle," *Aptera*. <https://www.aptera.us/vehicle/> (accessed Sep. 30, 2021).
- [10] "Homepage | Lightyear." <https://lightyear.one/> (accessed Sep. 30, 2021).
- [11] "Sion Electric Car," *Sono Motors*. <https://sonomotors.com/en/sion/> (accessed Sep. 30, 2021).
- [12] H. Koten, M. Yilmaz, and M. Zafer Gul, "Design consideration of solar powered cars," Jul. 2011, Accessed: Oct. 01, 2021. [Online]. Available: <https://www.osti.gov/etdweb/biblio/21547192>
- [13] A. Sierra and A. Reinders, "Designing innovative solutions for solar-powered electric mobility applications," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 7, pp. 802–818, 2021, doi: 10.1002/pip.3385.
- [14] B. Jiang and M. T. Iqbal, "Dynamic Modeling and Simulation of an Isolated Hybrid Power System in a Rural Area of China," *Journal of Solar Energy*, vol. 2018, p. e5409069, Jun. 2018, doi: 10.1155/2018/5409069.
- [15] M. Mousavi and M. T. Iqbal, "Design and Dynamic Modelling of a Hybrid PV-battery System for a House with an RO Water Desalination Unit in Iran," *European Journal of Electrical Engineering and Computer Science*, vol. 5, no. 6, Art. no. 6, Dec. 2021, doi: 10.24018/ejece.2021.5.6.370.
- [16] A. Castaldo, "Switching Regulator Fundamentals." Texas Instruments, Feb. 2019.
- [17] A. Castaldo, "Basic Calculation of a Buck Converter's Power Stage." Texas Instruments, Aug. 2015.
- [18] "Generic battery model - Simulink." <https://www.mathworks.com/help/physmod/sps/powersys/ref/battery.html> (accessed May 05, 2022).
- [19] "Constant-current constant-voltage battery charger - Simulink." <https://www.mathworks.com/help/physmod/sps/powersys/ref/cccvbatterycharger.html> (accessed May 05, 2022).

An Optimum Sizing for a Hybrid Storage System in Solar Water Pumping Using ICA

Amirhossein Jahanfar
 Department of Electrical Engineering,
 Faculty of Engineering and Applied Science
 Memorial University of Newfoundland
 St. John's, NL, Canada
ajahanfar@mun.ca

M. Tariq Iqbal
 Department of Electrical Engineering,
 Faculty of Engineering and Applied Science
 Memorial University of Newfoundland
 St. John's, NL, Canada
tariq@mun.ca

Abstract— Solar water pumps must be the most optimum size to work efficiently and be at a reasonable price. The storage system can play a main role in both system reliability and the total cost of a solar water pumping project; thus, it should be designed carefully. Traditionally, only batteries or water tanks are used as primary storage system; each of them has its benefits and drawbacks. In this research, a new approach to a storage system is proposed, consisting of both batteries and water tanks at the same time. Such hybrid storage can decrease project cost and increase system reliability. To find the most optimum size for such a hybrid system an optimization algorithm named “Imperialist Competitive Algorithm (ICA)” is used to minimize the Life-Cycle Cost Analysis (LCCA) of the storage system. In this paper, a hybrid storage configuration for solar water pumping for a site in Iran is proposed, and results of the optimum size for that system using ICA are expressed. It is shown that the configuration is more feasible compared with many other configurations.

Keywords— Solar water pump, Energy storage, Hybrid storage system, Optimization, ICA, Life-cycle cost analysis

I. INTRODUCTION

These days, conventional pumping systems are replaced with solar water pumping in remote areas, especially in Middle East countries. The most significant issue with solar-based systems, including solar water pumping, is storage systems, which increase solar projects' cost [1].

In the authors' previous work [1], two configurations of storage systems, battery storage and water tank, were proposed, and the advantages and disadvantages of each system were discussed. It is wise to design a hybrid storage system consisting of batteries and water tanks to take advantage of both systems. In addition, the system's reliability will be increased because in case of failure in either water tank or battery bank, another one can cover partial needed storage.

Apart from what is said, the most motivation to have a hybrid storage system is that if only batteries are used as a storage system, it will be pricy, or if only a water tank is used as a storage system, an amount of produced solar energy will be missed because PV output should reach to a specific power to run the pump in the morning, and when the output power drops below a threshold point, the pump will be stopped. So, the produced

power in the early morning and late afternoon will be missed as shown in figure 1 below.

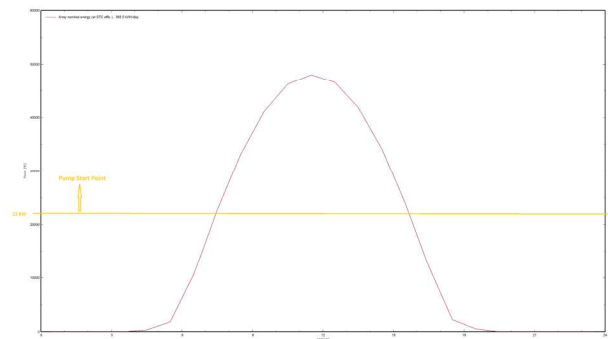


Figure 1: Nominal output array power for the site in Iran (for June 17th)

Since a hybrid system consists of both battery and water tank, conventional design software like Homer pro or PVSYS cannot be employed. Thus, an optimization method should be employed to calculate the optimum size of battery storage or water tank storage system in this hybrid system. Due to complications of this optimization problem and storage constraints, the “Imperialist Competitive Algorithm (ICA)” is used; this evolutionary algorithm is more efficient and straightforward to implement in comparison with typical optimization methods.

In this paper, first, an overview of hybrid storage systems is given; then, the optimization problem is defined in detail. After expressing the ICA, it is shown how to implement this algorithm to reach an optimum size of hybrid storage for solar water pumping for a site in Iran. In the end, a conclusion of the results of the proposed systems are given.

II. HYBRID STORAGE SYSTEM

The proposed hybrid system consists of a few strings of batteries and a water tank. The batteries can be charged during the early morning and late afternoon, save any excess energy during the day and provide power to the water pump whenever needed. The water tank is used to store excess pumped water and discharge water in case of pump failure or unexpected

water demand. A simple schematic of this hybrid storage system is depicted in figure 2.

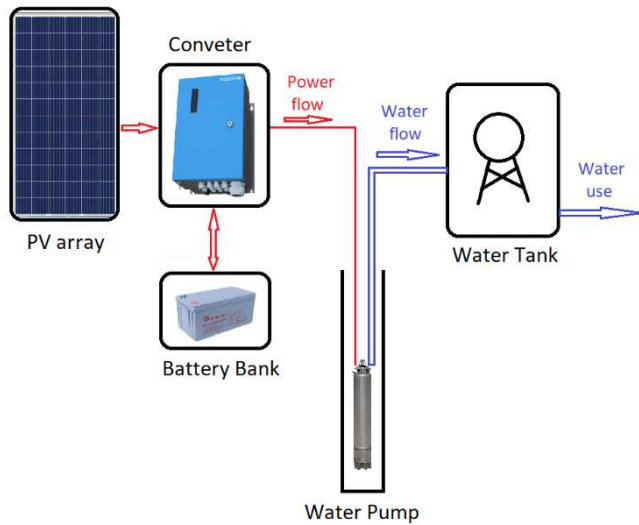


Figure 2: Schematic of a water pumping system with a hybrid storage

This system is proposed for a site near Mashhad in northeastern Iran; this site's approximate area is two hectares, which is divided into 20 cherry and apple gardens with shared water well [1]. Currently, they use diesel generators (figure 3) to provide power to the water pump which this research aims to replace it with an affordable and efficient solar water pumping.



Figure 3: Primary diesel generator which is used in this site

III. DEFINE OPTIMIZATION PROBLEM

The advantage of a hybrid storage system is expressed above. The question has still remained what is the most optimum size of the battery bank and what is the best capacity for the water tank to reduce the system cost while meeting the minimum needed back-up for solar water pumping to guarantee the system reliability; clearly, this is an optimization problem. Like any other optimization problem, first, it is necessary to determine the objective function (which can be called cost function because here is a minimization problem) and constraints that might be

linear, nonlinear, equal, or unequal. In the following, the optimization problem is expressed for a hybrid storage system of solar water pumping for the site under study.

A. Cost function

In this research, like some other work such as [2], Life-Cycle Cost Analysis (LCCA) is considered as the cost function:

$$LCCA = C_C + C_{O\&M} + C_R$$

Where:

- C_C =Capital cost
 - $C_{O\&M}$ =Operation and maintenance expenses
 - C_R =Replacement cost during the project lifetime
- *Note 1:* According to Homer pro simulation for this site in Iran, the mean battery depth of discharge is approximately %15. Also, based on the battery manufacturer company, if %15 of battery capacity is used, the battery lifetime will be more than 2000 cycles (figure 4). Since the operation period in this site is about five months per year, there is no need to replace the batteries during the project lifetime, as shown in Homer pro results in figure 5.

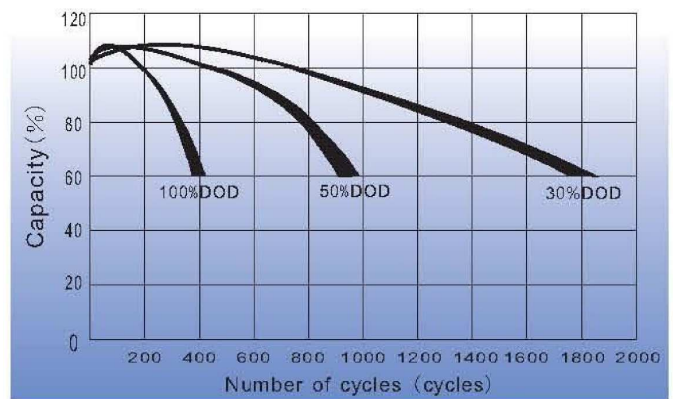


Figure 4: Life characteristics of cyclic use for Euronet Gel Battery

The water tanks are last long, and there is no need to replace them during the project lifetime. Thus, the term C_R can be omitted in the LCCA function.

Quantity	Value	Units
Autonomy	68.5	hr
Storage Wear Cost	0.133	\$/kWh
Nominal Capacity	263	kWh
Usable Nominal Capacity	210	kWh
Lifetime Throughput	228,510	kWh
Expected Life	54.7	yr

Figure 5: Results of Homer pro battery analysis

- *Note 2:* the used batteries are gel batteries which unlike the conventional lead-acid batteries, they do not need to charge after each period of use. In addition, the water tanks have no specific operation or maintenance cost. As a result, the term $C_{O\&M}$ is relatively small, so that it can be omitted as well.

For this specific site in Iran, the sum of the capital cost of water tanks and batteries is considered as a cost function:

$$\text{Cost function} = P_B + P_T ; \begin{cases} P_B = \text{Price of batteries} \\ P_T = \text{price of water tanks} \end{cases}$$

B. constraints

The only constraint in this problem is that there should be enough stored energy, whether in the form of chemical energy in batteries or potential energy of stored water in tanks, to ensure the reliability of the solar water pumping system during the operation period on the site in Iran. The best source to find this minimum needed energy is Homer pro simulation because Homer calculation is based on accurate data for the ambient condition on the site in addition to battery specifications. Based on Homer simulation, the minimum stored is calculated:

- 1) Homer suggestion for total size of the battery bank for the site in Iran is 240 KWh
- 2) Just 80% of the battery capacity is allowed to use (min SoC is 20%); also, the efficiency of this battery is 85%; so the minimum needed stored energy is:

$$\begin{aligned} \text{minimum needed stored energy} \\ &= 240 \text{ KWh} \times 0.80 \times 0.85 \\ &= \mathbf{163.2 \text{ KWh}} \end{aligned}$$

It is found that this solar water pumping in Iran needs at-least 163.2 KWh of stored energy. As a result, the constrain for this optimization problem is:

$$((E_B + E_T) - 163.2) \leq \varepsilon$$

Where:

- E_B is the stored energy in batteries
- E_T is the stored energy in water tank
- ε is a small positive number. In this paper, it is considered as five percent of the minimum needed energy to ensure daily water demand is satisfied.

$$\varepsilon = 163.2 \times 0.05$$

The following procedure is taken to calculate the stored potential energy in the battery and water tanks:

Stored energy in battery: To calculate the total stored energy in batteries, simply can multiply number of used batteries to the ampere-hour of each battery times voltage of each battery:

$$E_B (\text{Wh}) = \# \text{ of batteries} \times 100 \text{ Ah} \times 12 \text{ V}$$

Stored energy in water tank: The potential stored energy in water tank can be obtained based on injected water into the tank by water pump. To do so, first, divide tank capacity by rated water flow to see how many hours it takes to fill this water tank; then, multiply the water pump's nominal power to calculate hours to obtain the energy consumed by pump to pump sufficient water to the tank. This energy is called the *stored capacity* of the tank [3]:

$$E_T (\text{KWh}) = \text{Tank capacity} (\text{m}^3) \div 27 (\text{m}^3/\text{h}) \times 22 (\text{KW})$$

IV. IMPERIALIST COMPETITIVE ALGORITHM (ICA)

Imperialist Competitive Algorithm (ICA) is an evolutionary algorithm which is proposed by Esmail Atashpaz-Gargari in the year 2007 [4]. This optimization algorithm is inspired by imperialist competition on their properties. This algorithm starts with a random initial point called "country"; countries are divided into two groups, imperialists and colonies which each colony belongs to an imperialist. During the run of this algorithm, imperialists start a competition with other imperialists to take power over more colonies. In the end, the most powerful imperialists take control over all countries and converge them to an optimum global point. In this section, an overview of ICA is expressed, and flowchart of this algorithm is shown in figure 6.

A. Initialization

In the beginning, the algorithm generates Npop (number of total countries in the world) random initial countries which each country is defined as a $1 \times Nvar$ array where Nvar is the number of variables of intended optimization problem; each element of this array represent a property of that country like race, language, and so on [4].

$$\text{country} = [P_{\text{race}}, P_{\text{language}}, \dots, P_{Nvar}]$$

The algorithm calculates the power of each country by evaluating the objective function and then ranks the countries according to their power (a country with higher fitness is more powerful).

$$\begin{aligned} \text{fitness} &= \text{objective function}(\text{country}) \\ &= f(P_{\text{race}}, P_{\text{language}}, \dots, P_{Nvar}) \end{aligned}$$

Then Nimp (number of imperialists) of most powerful countries are selected as imperialists and the rest of the population (Ncol) as colonies. Afterward, all colonies are divided among imperialists such that an imperialist with a higher power must have more chances to get more colonies[4].

$$N_{\text{col}} = N_{\text{pop}} - N_{\text{imp}}$$

In this research, to divide colonies among imperialists, the Roulette Wheel selection is used; casinos' roulette inspires this method. This wheel is like a pie divided into different partitions, representing the normalized power of an imperialist.

$$\begin{aligned} \text{Normalized power of imperialist } k^{\text{th}} \\ &= 1 - \frac{\text{Cost}(\text{imperialist } k^{\text{th}})}{\sum_{i=1}^{N_{\text{imp}}} \text{Cost}(\text{imperialist } i^{\text{th}})} \end{aligned}$$

A random number between 0 and 100% is generated to select an imperialist and allocate a colony to that imperialist. In this way, an imperialist with a higher power has a higher chance to be selected; as a result, an imperialist with a higher power has more colonies. Figure 7 is an example that shows the second imperialist is the most powerful imperialist; as a result, it has the most share in nominalized probability.

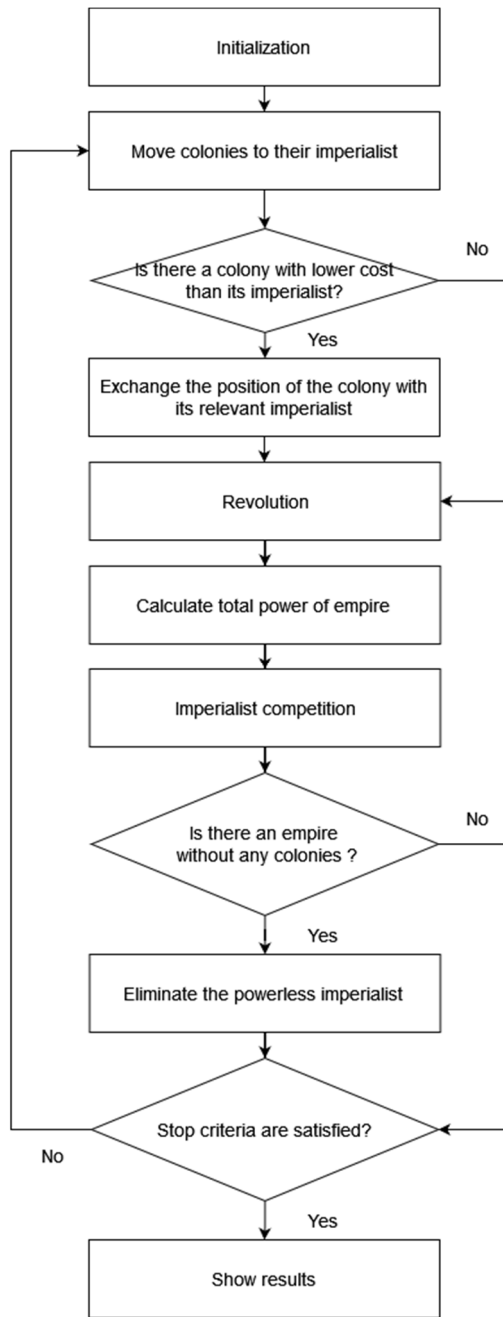


Figure 6: ICA flowchart

B. Colonies moving toward their imperialist

Imperialists try to make their empire more powerful by moving their colonies toward themselves; in this way the total power of the empire will rise, so the chance of winning that imperialist in the competition will be increased, which will result in empire expansion.

To move a colony toward its imperialist the following function is defined:

$$x = \beta \times rand \times d ;$$

- $\beta =$ moving coefficient
- $d =$ distance between imperialist k^{th} and colony h^{th}

$$new\ position\ of\ colony\ h^{th} = old\ position\ of\ colony\ h^{th} + x$$

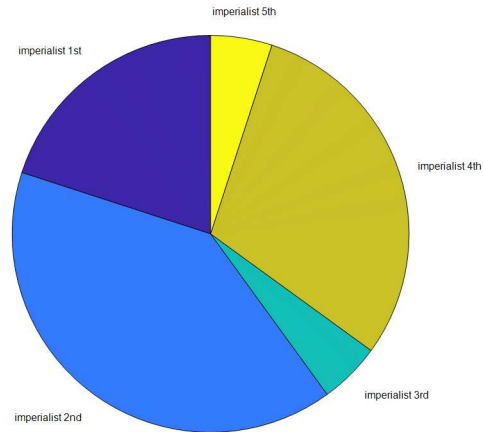


Figure 7: The nominalized probability of imperialists

C. Exchange the position of a colony with its relevant imperialist

After a colony moves toward its imperialist, it may find a better location in the search area with higher fitness than the imperialist; in this scenario, the position of the imperialist and that colony will be switched [5].

D. Total power of an empire calculation

The total power of an empire is defined as the sum of the power of the imperialist of the empire and a coefficient of the mean power of all colonies of the empire[4]:

$$\begin{aligned}
 Total\ Power\ of\ Empire\ k^{th} &= power\ imperialist\ k^{th} + \zeta \\
 &\times \left(\frac{\sum_{i=1}^{N_{col}} Fitness(colony\ i^{th})}{N_{col}} \right)
 \end{aligned}$$

ζ is a constant coefficient ($0 < \zeta < 1$) that determines the share of colonies' power in the empire's total power. Small zeta means imperialist power has the most effect on total power and large zeta means the effect of colonies' power is considered.

E. Imperialist competition

As is clear from the name of this algorithm, this step of the algorithm is the most important step. In the competition among imperialists, they try to take control over more colonies in other empires to increase their empire power. During this competition, the weakest empire has the most likelihood of losing colonies and the most powerful empire has more chance of owning more colonies.

In this algorithm, the weakest colony in the weakest empire is picked. It is given to the selected empire using the roulette wheel, which means that an empire with the most total power has more probability of owning the weakest colony [4].

F. Eliminating the powerless imperialist

During the run of the algorithm and after a couple of loops, an imperialist might lose all of its colonies; in this case, the relevant empire will be collapsed, then imperialist will be considered a colony and it will be assigned to one of the rest empires with roulette wheel selection [4].

G. Stop criteria

Stop criteria can be defined in different ways depending on the nature of the optimization problem. In this research, reaching a specific number of algorithm loop (generation) is considered a stop criterion.

V. IMPLEMENTATION OF ICA FOR OPTIMUM SIZE OF HYBRID STORAGE FOR SOLAR WATER PUMPING IN IRAN

In this section, the properties of used ICA are presented, and the result of optimum sizing is expressed.

A. ICA initial parameters

The ICA algorithm for this research is programmed and run in MATLAB R2021a. The initial parameter of used ICA is mentioned in the following table:

Table 1: The initial parameter of employed ICA in this research study

Parameter	Value	Note
Number of countries' property	2	There are two variables: battery and water tank
Number of countries	30	
Number of Imperialists	5	
Lower and Upper Bound of Battery Value	[0 200]	# of batteries
Lower and Upper Bound of water tank Value	[0 200]	In m3
Moving coefficient (β)	2	
Weight of mean cost of Colonies (ζ)	0.1	
Maximum number of iterations	10000	
Percent of Revolution	0.2	
probability of revolution operation	0.2	
probability of revolution on each colony	0.4	

The initial population is generated uniformly randomly.

B. Search area

In this research, a lookup table consisting of prices of batteries and water tanks in different capacities is given to the algorithm as input; the algorithm performs a linear interpolation method to build a continuous search area.

C. Cost function and constrain

The cost function in this research is the summation of the price of batteries and water tanks:

$$Cost = price\ of\ batteries + price\ of\ water\ tanks$$

A penalty function is defined in order to apply the constraint:

$$if \left[\left(\frac{stored\ energy}{in\ water\ tanks} + \frac{stored\ energy}{in\ batteries} \right) - \frac{min\ needed}{stored\ energy} \right] \geq \frac{min\ needed}{stored\ energy} \times 0.05$$

then

$$Cost = (price\ of\ batteries + price\ of\ water\ tanks) + penalty$$

According to section III, the minimum needed stored energy for the under-study site in Iran is 163.2 KWh. Also, the penalty should be a large number, in this research, it is CA\$9,000,000.

D. Revolution

This auxiliary operation for ICA mimics the mutation operation in the Genetic Algorithm (GA)[6], [7]. This operation takes one colony and moves that randomly in the searching area. Revolution might slow down the algorithm, but it ensures that the ICA will not be stuck in the local optimum.

E. Results

The algorithm suggestion for optimum size of storage system for this specific solar water pumping system in Iran is as follows:

Table 2: the output results of ICA

storage	Value
Number of batteries	40
Water tank (m ³)	140

Figure 8 illustrates how the best country in the world improved during the algorithm's run; at first, the algorithm found a system size with a cost of more than CA\$28,000, and after a couple of loops, it tried to optimize the size of hybrid storage. This improvement happened step by step during the algorithm run, and after the loop 1000, the algorithm reached a steady-state, which means that the most optimum size of the hybrid storage system is found.

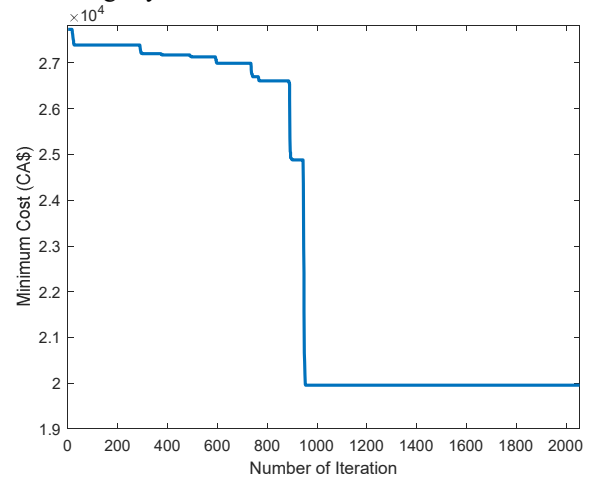


Figure 8: The cost of best imperialist during the run of algorithm

The following calculation can be done to show that the found size of the system can satisfy the constraint:

$$V_{Bus} = 12V \times 10 = 120V$$

Ah of each string is 100 Ah, so total bank Ah= 4×100= 400

Energy in batteries (Wh) = 400×120= 48000= 48 KWh

Energy in batteries (consider energy lost and depth of charge) = 48×0.8×0.85= 32.64 KWh

With 40 batteries, the pump with 22KW power can run for about 1.5 hours (32.64 /22).

This pump can pump about 40.5 m³ of water (1.5hr × 27(m³/hr)).

Also, there is a tank with 140 m³ of water storage.

As a result, this system has 180.5 m³(40.5+140) water storage This site in Iran needs about 180 m³ (188.6 ± 5%) of water as a back-up for one day. So, this system satisfies the water requirement as back-up for this site in Iran.

To justify the output result of ICA, the following table is prepared with a couple of battery and water tank combinations. As can be seen in table 3, if battery is used more, the total cost is high, as the number of batteries decreases and the capacity of water tank increases, the total cost starts to drop. At the minimum point it reaches to lowest cost and after that the cost start to increase. As a result, the optimum size for this specific site in Iran is supposed to be around this point that the ICA found it correctly.

Table 3: Some feasible size of the hybrid storage system

# of batteries	Water tank capacity (m ³)	Total Price (CA\$)
130	9	26562
120	24	25509
110	39	24483
100	55	23757
90	68	22760
80	82	21860
70	97	21366
60	112	20699
50	128	20448
40	140	20041
40	141	20342
40	142	20644
40	143	20945
30	156	22862
30	158	23464
20	170	24905
20	171	25188
10	184	26540
10	185	26740

F. Comprasion

As expressed in table 4, using a hybrid storage system with 40 batteries (each battery 100 Ah) and a 140 m³ water tank, the total cost is approximately CA\$20,000, which is cheaper than the two other configurations.

Table 4: comparison of storage methods

Storage type	capacity	Cost (CA\$)
batteries (Number of 100 Ah battery)	200	40,000
Water tank (m ³)	180	23,800
Hybrid	Batteries: 40 × 100Ah and Water tank: 140 m ³	20,041

Although the difference between hybrid storage and only water tank storage is not significant, the hybrid system boosts system reliability and can provide sufficient water during the hours of operation.

VI. CONCLUSION

In this paper, a hybrid storage system is proposed for a site in Iran. It is discussed that a hybrid storage system can take advantage of both battery and water tank at the same time, resulting in lower cost and higher system availability.

Also, ICA is employed to find an optimum size of the system. It not only guarantees the minimum needed storage, but also reduces the cost of the storage system. The algorithm suggests a hybrid storage system consisting of 40 batteries and 140 m³ water tank; this is the cheapest configuration for the site in Iran, which can satisfy the minimum needed storage.

This research shows that a hybrid storage system can be more economical in comparison to conventional configurations for storage systems in solar water pumping.

ACKNOWLEDGMENT

The authors kindly thank Roshana Gostar Shargh Barsava for providing funding for this research.

REFERENCES

- [1] A. Jahanfar and M. Tariq Iqbal, "A Comparative Study of Solar Water Pump Storage Systems," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022, pp. 1070-1075, doi: 10.1109/CCWC54503.2022.9720912.
- [2] C. Soenen et al., "Comparison of Tank and Battery Storages for Photovoltaic Water Pumping," Energies, vol. 14, no. 9, p. 2483, 2021.
- [3] U. Ashraf and M. T. Iqbal, "Optimised Design and Analysis of Solar Water Pumping Systems for Pakistani Conditions," Energy Power Eng., vol. 12, no. 10, p. 521, 2020.
- [4] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition," in 2007 IEEE Congress on Evolutionary Computation, 2007, pp. 4661–4667, doi: 10.1109/CEC.2007.4425083.
- [5] S. Nazari-Shirkouhi, H. Eivazy, R. Ghodsi, K. Rezaie, and E. Atashpaz-Gargari, "Solving the integrated product mix-outsourcing problem using the Imperialist Competitive Algorithm," Expert Syst. Appl., vol. 37, no. 12, pp. 7615–7626, 2010, doi: https://doi.org/10.1016/j.eswa.2010.04.081.
- [6] Z. Pan, D. Lei, and Q. Zhang, "A New Imperialist Competitive Algorithm for Multiobjective Low Carbon Parallel Machines Scheduling," Math. Probl. Eng., vol. 2018, p. 5914360, 2018, doi: 10.1155/2018/5914360.
- [7] R. L. Haupt and S. E. Haupt, Practical genetic algorithms. John Wiley & Sons, 2004, pp 43-44.

A Real-Time Parking Space Occupancy Detection Using Deep Learning Model

Raktim Raihan Prova, Title Shinha, Anamika Basak Pew, Rashedur M. Rahman

*Department of Electrical and Computer Engineering
North South University*

Plot-15, Block-B, Bashundhara, Dhaka1229, Bangladesh

{raktim.prova, title.shinha, anamika.pew, rashedur.rahman}@northsouth.edu

Abstract— A camera is a tool to record visual footage in the form of photographs, film or in video format. However, a smart camera can be recognized as a device to retrieve application-specific information from the recorded footage. In this paper, we have proposed a solution to detect parking lot occupancy status using deep learning model and commercially used CCTV cameras in real time. Our implemented solution is decentralized and efficient in terms of light-weight deployment to low powered devices like Raspberry Pi. Our proposed solution is compared with the existing approaches. Our deep learning model is also tested on other datasets having images taking from multiple CCTV camera implemented in different height. Along with this, we have tested our model on indoor both outdoor parking garages in low light conditions during day and evening. Result of the performed experiments shows that our model is operable in low-powered embedded devices with effective accuracy.

Keywords — Deep Learning, Machine Learning, Convolutional Neural Networks, Classification, Embedded Device

I. INTRODUCTION

Automated parking lot occupancy status detection in real time is important for managing parking lots effectively. This should be also scalable and able to reduce the time to find the parking lot occupancy status in real-time. Traditionally sensors are installed in every parking spot to detect if an individual spot is being occupied or not. The main drawback of this approach is the maintenance cost, although the detection accuracy is higher providing that the installed hardware is in working order. In addition to that, this approach needs additional cost at the hardware installation phase, and regular maintenance is required for the system up and running in normal and extreme weather conditions specially in rainy and winter season. For additional parking spot in a given parking space new hardware needs to be installed. In this scenario, the total cost of the system piled up exponentially as the number additional of parking spot is added to the system.

In recent years, there are many applications which are developed to serve the purpose to detect parking spots in a parking space at comparatively low cost and maintenance. Our proposed approach is using a decentralized system where no central server will not be used to perform computational tasks for detecting if a parking spot is occupied or not considering different lighting conditions and obstacle presence like shadow, human interference etc. We have designed our model on Deep CNN (Convolutional Neural Network). [1][2] However, the used CNN model needs to be modified in such a way so that it runs fast enough on low

powered devices like Raspberry PI. This decentralize approach solves the computational bottleneck that is normally faced in centralized systems and ensure better scalability with respect to the cost of installation and maintenance, because to add a new parking lot to the system, we need not to install additional sensors. A single camera can monitor multiple parking locations.

However, the problem to detect multiple parking spot occupancy status using video footage is not new. Several researches have been performed in this regard. The authors in [3] [4] [5], detected occupancy status using only visual information like using only image or video footage is done with case-specific contexts and scenarios. This means that model is being trained with the specific parking space context-based data. These approaches can not easily be generalized to parking lots in other parking spaces.

To make our application more generic and robust deep CNN is used for our approach. Our solution provides better accuracy in presence of shadows, other kind of obstacles and human presence. With this model training phase is less expensive in terms of using computational resources than the classification steps. For this reason, this model can be handled using low powered and embedded devices. In addition to that, this model performs even better if few context-specific training data related to the specific parking lots are provided along with generic training data. To validate the robustness of our deep CNN model in other approaches, we have used two additional parking spaces datasets. One is CNRPark-Ext, another one is PKLot [5]

These two datasets contain the images from two different and independent parking spaces, and images from different viewpoints, weather conditions and presence of obstacles. These conditions are considered as the presence of these conditions makes the detection task of the parking lot occupancy more challenging. These datasets are annotated manually. More of the datasets are discussed in Section 4. In addition to that, we trained the model on one scenario and test the model on different scenario, and then compare the results found in different approaches to find how our model performs in generalized conditions.

II. RELATED WORK

In the very first attempts to detect occupancy status one paper takes help of SVM(Support Vector Machine) classifier to separate the occupied spot from the unoccupied spot in a

parking space. In summary, this classifier just differentiates the car regions with the parking space region to detect the occupancy status [6]. This approach has lower accuracy in presence of shadow, different light and weather conditions. Later [3] another approach was made to overcome the low accuracy due to the presence of various kinds of obstacle by considering the three neighboring parking lots. The state and color histogram of the neighboring parking lots are considered as the feature in SVM classifier. In this method, situations like changes of light were not addressed properly causing lower accuracy.

Another study was conducted to address the light change [7]. The authors trained a classifier based on Bayesian to detect the occupancy status. Corners, edges and wavelet are considered as features for the classifier. Another study used Bayesian hierarchical framework to detect empty spot based on three-dimensions of the given parking space that operates day and night. Similar method was followed in another study [8] that models available parking spots as a single pack in the three-dimensional space. This model can provide higher accuracy comparatively than previous approaches addressing obstacles like light changes. Another study [9] is conducted that used customized trained neural network model that can detect occupies or not occupied based on extracted visual feature from specific parking space. For this study, 126 parking spots in a single parking space in considered. This method provides robust result than other models in considering the lighting condition.

Recently another study is conducted that addresses this problem is machine learning approaches. This study uses around 700,000 images taken using three different cameras. These images are used to train the SVM classifiers where multiple texture features are considered. Although they have used similar kind of classifier but this study provides better accuracy in terms of detection performance as then used simple aggregation function either max or min to the confidence values.

There are also other studies found based on temporal analysis of video frame. This study [4] uses this method by subtracting background with Gaussians to detect the vehicles that are parked or that are being leaving. Another study [10] focused in considering obstacles considering that many parking spaces can be hidden by another vehicle that is parked in neighboring parking lot. In this study tracking is performed considering two events: vehicles that are leaving and entering in parking lot. There are also other approaches that consider both features extracted from footage and sensors. This study [11] proposed an approach to detect parking lot occupancy status with the help of vehicle navigation system. Another study [12] proposed a solution that collects real-time occupancy status based on smartphone sensors.

In recent time this study [13] addresses our problem by using deep CNN to build generic model to detect open parking spaces. Our study is the extended version of their study to build a generic model that can detect occupancy status of a

parking lot in indoor parking spaces where obstacle in term of shadow and absence of daylight is comparatively higher.

III. DATASETS

1. *PkLot*

PKLot dataset consists of around 700,00 images of parking lots. Images are organized by the weather condition and the date of capture. Only sunny, cloudy and rainy weather conditioned in considered. PKLot dataset is comparatively larger than both CNRPark and CNRPark_EXT.

2. *CNRPark-Ext*

CNRPark-Ext is the extended version of CNRPark [14] which is a small dataset compare to CNRPark-Ext consists of around 12,000 labeled images. Images of CNRPark is captured in different situations of light including considering different kind of obstacle like shadow, human presence and trees. CNRPark_EXT dataset contains the images collected using nine cameras. The images are taken considering different perspective and angle of views. Different light conditions and the including obstacles like tree, human interference, shadow, presence of other cars in neighboring parking spaces are also being considered while selecting images for this database. This dataset also contains the images partially shadowed vehicles. This kind of dataset trains our model to perform with better accuracy as this is the real scenario in most of the cases. All of these three datasets images are being cropped with the help of masks to smaller one. Each smaller images contains only single parking lot. Those smaller images are referred as patches. The resolution of these patches is 150px*150px. Each patch is labeled as 0 if the parking spot is not occupying and 1 if it is occupied. Finally Patches of these datasets are groups together into subsets based on weather condition, camera ID, image taken date. We have trained our model on these different subsets to justify which one provides better accuracy.

CNRPark and CNRPark_EXT both datasets cover the real-life scenarios more accurately other than PKLot. So, it is expected that CNRPark will train the datasets for real-life scenarios with better accuracy in testing period.

3. *Indoor and Outdoor*

We have collected 60 minutes footage of both indoor and outdoor parking spots. Those footage are being recorded using commonly used CCTV cameras in apartments. The resolution of the captured footage is 1280px*720px at 30fps. After each 5 seconds or 150frames, each image is being extracted for testing purpose. Both footage for indoor and outdoor is captured using single camera covering three parking slots. We built masks that allow cropping the full pictures into smaller ones. Each smaller one contains single parking spaces. These smaller images are labeled for validation purpose. Scenario of our testing dataset matches with the used training dataset CNRPark and CNRPark_EXT.

In terms of obstacle presence both test datasets are matched with the extreme scenarios. For Indoor dataset (Figure 01) have higher obstacle presence in terms of presence of shadow, comparatively low presence of day light or artificial source of light. For outdoor dataset, due to high amount of camera view, the adjacent vehicle occupies a small portion of neighboring parking spaces. In both cases, detection occupancy status is quite challenging.



Fig 01: Overview of PkLot and CNRPark-EXT



Fig 02: Overview of Indoor and Outdoor Datasets.

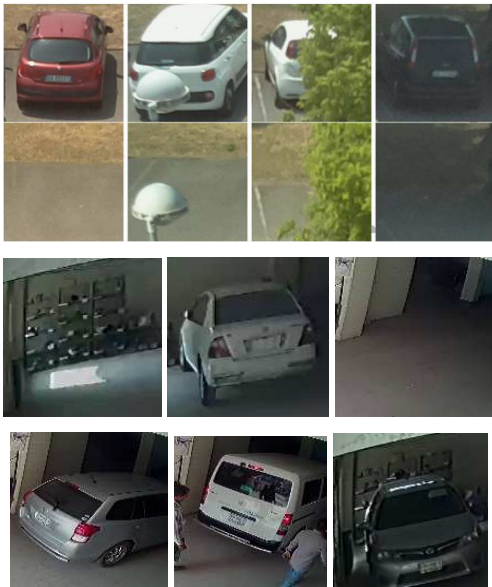


Fig 03: Overview of Segmented Images from datasets.

IV. METHODS AND APPROACH

In order to enable the model to track down the occupancy status of slots in a given parking space, we need to follow following steps. However, before deploying the model to the embedded Raspberry PI, we need to pretrain the model. Based on the pretrain model segmented images will be classified.

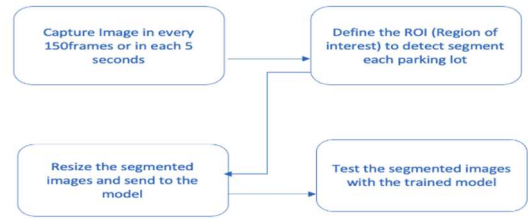


Fig 04: Work Flow of the System

1. Capture Image from Footage

The method to capture image from the video footage in every 5 seconds. In this process, image extraction depends on the frame rate of fps of the CCTV camera. Our tested CCTV camera delivers footage at 25fps. For this reason, we extract each image after closely 125frames. This Image extraction process can be changed based on the camera configuration. Each Image is extracted automatically for the next step to be segmented, and resized.



Fig 04: Extracted Image from CCTV after each ~ 5 seconds.

2. Define ROI and Prepare Patches

The captured images are filtered by Mask. The Mask identifies each and every parking spots in the given space. For our application, mask is implemented in case specific manner. This means that for every new parking space this Mask is needed to be redefined according the image resolution and location of the parking lot in the parking space.

For example, in Figure 4 the given parking space has three parking spots. The mask will segment the images into 3 parts. These segmented images are marked as patches.



Fig 05: Segmented Patches from the Image by Mask.

3. Prepare Patches for Input to Model

Our CNN Model take 224*224 RGB image as input for performing classification. In this stage we resize the image into the following format and prepare for the next step which is to feed the resized images to the pre-trained model for making a decision.

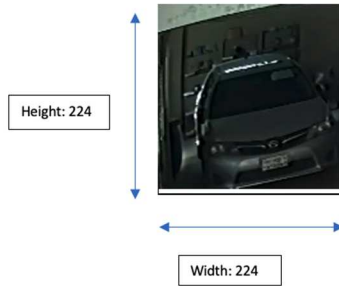


Fig 06: Resized image to 224*224.

4. Deep Convolutional Neural Networks (CNN)

We have tested deep CNN architectures. One is mini-AlexNet and another one is miniLetNet. Those are reported in [15] [16] These architectures have higher accuracy rate in extracting features using visual recognition. Both mini-AlexNet and mini-LetNet have five trainable layers. Lower number of layers ensures feasible computational capability in real-time, also ensures to be runnable on low powered embedded devices. This architecture is inspired from AlexNet. AlexNet is a large CNN architecture to decide multiple decisions. As we need to decide binary decisions, the architecture is reduced to be run on real-time. AlexNet is trained on million datasets and task was to recognize 1000 different classes.

Both mini-AlexNet and mini-LetNet takes an image of 224*224 as input. We have prepared the segmented patches for input in previous step.

Mini – LetNet has total four layers including the output layer. Out of these four layers first two are Convolutional layer. These two convolutional layers is followed by two fully connected layers including a max-pooling layer. The first layer is similar to the layer of AlexNet [15]. Thus, it is suitable for the resized segmented images in previous step. Number of neurons in last convolutional layer and last fully connected layer is reduced to meet the classification dimension for detecting two classes. Reduced number of neurons and filters does not overfit in the testing phase. In this study [16] Gaussian Radial Bias Function is used in the last layer. For mini-LetNet this is replaced by 2-way SoftMax classifier.

Mini-AlexNet CNN architecture that is defined for classifying the occupancy status for a parking spot is inspired from the AlexNet[7]. Five trainable layers are used in this architecture, among them three layers convolutional layer. Last two layer are fully connected layer. Two fully connected layer are followed the convolutional layers, and the fully connected layer is considered as the output layer. First and second convolutional layer is followed by *max – pooling* and *local rectification* (ReLU). Third and last convolutional layer follows the same architecture. Only first two convolutional layer implement *local normalization* (LRN).

The final architecture of miniAlexNet with reduced number of neuron and filter have roughly $\frac{1}{1340}$ parameters than AlexNet[7]. In the last two fully connected layer, drop out regularization method is followed. Same as mini LetNet, mini AlexNet takes 224*224 RGB image as input for classification.

In the training phase, for data augmentation purpose *random – crop* and *horizontal – flip* technique is followed. Each horizontally image is assigned with the probability of 0.5.

However, in training phase, no horizontal flipping is applied. Only resized images 224 * 224 are feed for the input to the architecture.

$$size\ of\ the\ filter = number * width * height + stride$$

$$Max - pooling = width * height + stride$$

Net	Conv 1	Conv 2	Conv 3	Fc – 4	Fc – 5
miniL Net	30 * 11 * 11 + 4 pool 5 * 5 + 5	20 * 5 * 5 + 1 pool 2 * 2 + 2	-	100 ReLU	2 Soft – max
miniA Net	16 * 11 * 11 + 4 pool 3 * 3 + 5 LRN, ReLU	20 * 5 * 5 + 1 pool 3 * 3 + 2 LRN, ReLU	30 * 3 * 3 + 1 pool 3 * 3 + 2 ReLU	48 ReLU	2 Soft – max

Table 01: CNN Architectures

Both mini – AlexNet and mini – LetNet architecture has densely connected convolutional layer. Details of the layers are represented in the above table.

V. EVALUATION

To assess the performance of proposed CNN based on existing architecture, *miniAlexnet* and *miniLetNet*, we considered two cases:

- i. Train and Test CNN model on same dataset
- ii. Compare accuracy with existing approaches and datasets

1. Train and Test CNN model on Same Dataset

Both Indoor and Outdoor dataset contains the images taken from one camera. For both datasets images are in roughly 50%~50% split. All patches are horizontally flipped and resized to 256*256 pixels. Each image is classified independently. No previous model is used for testing purpose. Each time for every dataset the model is trained again. Overfitting of the model will be limited as training and testing dataset is different and there exist sufficient amount of data for the training purpose.

Datasets	Unoccupied Slots	Occupied Slots	Total
Indoor Garage	310	230	540
Outdoor Garage	312	246	558

Table 2: Details of datasets used for training and testing purpose in (i)

We trained and tested our model on multiple splits to limit the problems of overfitting. The model is trained on Split A and tested on Split and vice versa.

Datasets	Unoccupied Slots	Occupied Slots	Total
Indoor Split A	150	110	260
Indoor Split B	160	120	280
Outdoor Split A	160	120	280
Outdoor Split B	152	126	278

Table 3: Splits of datasets used for training and testing purpose in (i)

Train Dataset	Test Dataset	Model	Base Learning Rate	Accuracy
Indoor Split A	Indoor Split B	miniAlexNet	0.01	0.861
Indoor Split B	Indoor Split A	miniAlexNet	0.01	0.833
Outdoor Split A	Outdoor Split B	miniAlexNet	0.01	0.905
Outdoor Split B	Outdoor Split A	miniAlexNet	0.01	0.911

Table 4: Accuracy in different splits for miniAlexNet

Train Dataset	Test Dataset	Model	Base Learning Rate	Accuracy
Indoor Split A	Indoor Split B	miniLetNet	0.001	0.845
Indoor Split B	Indoor Split A	miniLetNet	0.001	0.822
Outdoor Split A	Outdoor Split B	miniLetNet	0.001	0.897
Outdoor Split B	Outdoor Split A	miniLetNet	0.001	0.874

Table 5: Accuracy in different splits for miniLetNet

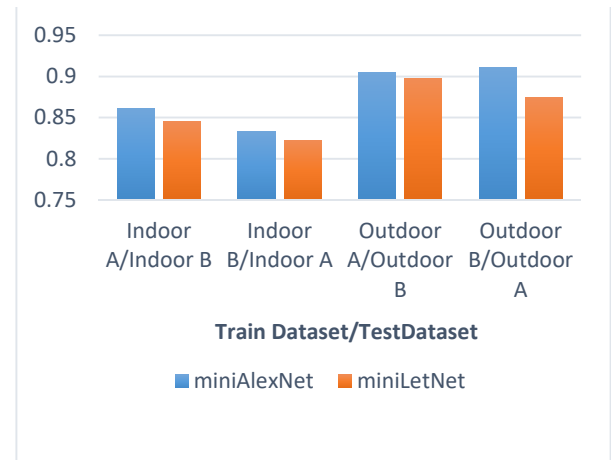


Fig 06: Representation of accuracy for various splits

2. Compare Accuracy with Other Approaches

To compare the accuracy among other approaches we have followed the techniques discussed in this paper [5]. The discussed classification techniques on this paper are based on SVM trained on textural of features. A study [17] [18] used LBP and their variations to detect occupancy status like taking

binary decisions. Our model is compared with various variations of LBP used in study [19]. In this experiment, we also CNRPark, PkLot, indoor and outdoor datasets.

CNRPark is divided into two partitions named as CNRPark-Even and CNR-Park. CNRPark-Even consists of even number of parking lots and CNRPark-ODD consists of odd number of parking lots. PkLot is divided into two subsets named as PkLot2days and PkLotNot2Days. Indoor and Outdoor datasets are in two split A and B like in previous experiment. To compare mAlexNet with study [4], extracted features of LBP-type ranges *radius* 1~8. SVMs in this study is trained using *grid – search* and cross validation on training set. With selected parameters SVM is then trained on entire training dataset. The final probabilistic output of the SVM is obtained with the help of posterior probability using sigmoid function. These same steps are followed in this study [4].

Method	Variation	Input dimension	Description
mAlexNet	-	224	224*224 resolution of inputted patch
SVM	LBP	256	Classical LBP [12]
SVM	LBPu	59	Uniform LBP [12]
SVM	LBPuri	10	Uniform and Rational invariant LBP [12]

Table 6: Accuracy in different splits for miniAlexNet

Dataset	Unoccupied	Occupied	Sub-Total
CNRPark A	2549	3622	6171
CNRPark B	1632	4781	6413
CNRPark	4181	8403	12584
CNRPark ODD	2201	3970	6171
CNRPark Even	1980	4433	6413
CNRPark	4181	8403	12584
PkLot2Days	27314	41744	69058
PkLotNot2Days	310466	316375	626841

PkLot	337780	358119	695899

Table 7: Details of CNRPark, PkLot Datasets

Train	PKLot2 Days	CNRPark_ Odd	Indoor Split A	Outdoor Split A
Test	PkLotNot2Days	CNRPark Even	Indoor Split B	Outdoor Split B
Model				
mAlexNet	0.961	0.901	0.861	0.905
LBP	0.916	0.821	0.761	0.781
LBPu	0.926	0.836	0.782	0.812
LBPuri	0.861	0.841	0.801	0.810

Table 8: Comparison between miniAlexNet and LBP



Fig 5: Representation of comparison between LBP and miniAlexNet

VI. CONCLUSION

We have presented a model for make a decision of occupancy status in parking lots of a parking space. Our model is based on deep Neural Network. As we need to only take binary decision either occupied and not occupied, our CNN architecture is robust enough to take decision on real time on embedded device.

In our study we have tested our CNN architecture on different scenarios to test the generalizability to justify how this system will perform if specific parking space related data is not provided in the training phase. To justify this, we considered both of the splits Indoor and Outdoor as training data, and test the model on our test dataset Indoor and Outdoor. Indoor and outdoor data are divided roughly 50%~50% split. In every case, the accuracy is equal and above of 80%. However, for dataset Outdoor, the accuracy is higher than Indoor as Indoor Dataset consists of segmented images

covers obstacles like light lamp, human interference, shadow, variation of light, adjacent vehicle presence, where Outdoor dataset does not address light changes. These scenarios are close to real-life scenarios.

We also tested our model with other approaches based on SVM. In those cases, our proposed CNN model performed higher accuracy than other models based on SVM. For Intra database experiment (Table 08), highest accuracy of mAlexNet is ~96%, and the lowest accuracy is ~86%, where for SVM based approaches, highest accuracy is ~92% and lowest accuracy is ~78%. This can be concluded that on generalized approaches where park space specific training data will be absent, mAlexNet based on CNN architecture will also perform better than other approaches like SVM and their variations.

This research can be extended to implement ANPD (Automatic Number Plate Detection) system. Implementing a system to detect vehicle number plate, we can more accurately pinpoint which vehicles are staying or leaving. Furthermore, this can be implemented for commercial use to replace centralized parking spots. However, this is quite challenging to perform OCR (Optical Character Recognition) on noisy and low resolution images.

REFERENCES

- [1] A. N. Belbachir, Smart Cameras, vol. 2, Springer Volume , 2010.
- [2] Y. Bengio, Learning deep architectures for AI, Foundations and trends R in Machine Learning, 2009.
- [3] Q. H. C. W. S.-y. C. W.-C. & C. T. Wu, "Robust parking space detection considering inter-space correlation," in *Multimedia and Expo*, 2007.
- [4] C. G. T. J. & M. J. M. del Postigo, "Vacant parking area estimation through background subtraction and transience map analysis.," *IET Intelligent Transport Systems*, vol. 9, pp. 835-841.
- [5] P. R. O. L. S. B. A. S. S. E. J. & K. A. L. de Almeida, "Pklot—a robust dataset for parking lot classification," *Expert System with Applications*, vol. 42, pp. 4937-4949
- [6] N. Dan, "Parking Management System and Method". US Patent App Patent 10/066,215.
- [7] L.-W. H. J.-W. & F. K.-C. Tsai, "Very Deep Convolutional Networks for Large-Scale image Recognition," in *Image Processing*, 2007.
- [8] C.-C. T. Y.-S. & W. S.-J. Huang, "Vacant parking space detection based on plane-based bayesian hierarchical framework.," in *Circuits and Systems for Video Technology*, 2013.
- [9] D. W. W. L. R. P. B. E. e. a. delibaltov, "Parking lot occupancy determination from lamp-post camera images.," in *16th International IEEE Conference on Intelligent Transportation Systems-(ITSC)*, 2013
- [10] I. W. A. J. A. & A. A. Masoudi, "Parking Spaces Modelling for inter spaces occlusion Handling," in *22nd International Conference on Computer Graphics*, 2014.
- [11] F. B. C. M. P. Caicedo, "Prediction of parking Space Availability," in *Expert Systems with Applications*, 2012.
- [12] K.-C. S. W.-Y. Lan, "An intelligent driver location system for Smart Parking," *Expert System with Applications*, vol. 41, p. 2443-2456, 2014.
- [13] X. X. S. L. C.-L. & P. C.-H. Chen, "Vehicle detection in Satellite Images by Hybrid Deep Convolutional Neural Networks," *Geoscience and Remote Sensing Letters*, vol. 11, p. 1797-1801, 2014.
- [14] G. C. F. F. F. G. C. & V. Amato, "Car parking occupancy detection using smart camera networks and Deep Learning," in *21th IEEE Symposium on Computers and Communications (ISCC)*, 2016.
- [15] I. S. a. G. E. H. A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, p. 1097-1105, 2012.
- [16] L. B. Y. B. a. P. H. Y. LeCun, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, p. 2278-2324, 1998
- [17] M. P. a. T. M. T. Ojala, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, p. 971-987, 2002
- [18] L. S. O. A. S. B. E. J. S. a. A. L. K. P. R. de Almeida, "Pklot—a robust dataset for parking lot classification," *Expert Systems with Applications*, vol. 42, no. 11, p. 4937-4949, 2015.
- [19] J. H. V. O. a. T. A. E. Rahtu, "Local phase quantization for blur-insensitive image analysis," *Image and Vision Computing*, vol. 30, no. 8, p. 501-512, 2012.

Solar PV System for Self-Consumption

Ananna Khan, Abdul Kahar Siddiki, Rashedur M. Rahman

*Department of Electrical and Computer Engineering
North South University*

Plot-15, Block-B, Bashundhara, Dhaka1229, Bangladesh
{ananna.khan1, abdul.siddiki, rashedur.rahman}@northsouth.edu

Abstract—Although Bangladesh has a total of 19 solar power projects of total 1070 MW capacity which generate 10 percent electricity from renewable energy source by 2020, PV technology is still not widely used in many parts of the country. This research focuses on using simulation tools to build and estimate the efficiency of a solar PV system for self-consumption at a region with a high solar potential that is currently underutilized. To learn more about the field and the grid local regulations in Bangladesh, a thorough review of photovoltaic search was reviewed. After that, this research supports in order to choose the best area for the installation. After that, a load profile is constructed to determine the energy consumption of the specific location, and ultimately, the technology to be implemented is chosen and characterized. An AutoCAD 3D modeling of the worksite and supporting components is put into a simulation software Pvsyst version 7.2 based on a particular premise. This enables the development of the PV scheme, the determination of the most appropriate values for the system's characteristics (tilt angle, azimuth, and row spacing), and the advancement of a shading and loss study. Finally, an economic analysis is conducted. The study finds the service to be efficient, cost-effective, and useful, as well as providing a design reference for solar PV self-consumption systems.

Keywords— Pvsyst, Solar PV panel, Inverter, Azimuth angle

I. INTRODUCTION

Energy is a crucial component of human growth. The world's population is expected to increase to more than 11.2 billion by the end of the century and world energy consumption should increase by approximately 1.7 billion tons of oil equivalent per year by 2025. This means that maintaining an oil energy share of around 33% (2015 levels) would require increasing current oil production by more than 11 million barrels per day [1]. However, fossil resources are not only limited, but they also produce greenhouse gas emissions that contribute to climate change, which can have destructive consequences for the ecosystem. The most abundant source of energy that comes straight from the Sun is solar energy. According to NASA calculations, the Sun still has 6500 million years of life left and, every hour, it throws onto Earth more of the energy necessary to supply the global demand of a whole year [2]. Solar photovoltaic (PV) panels use the photovoltaic effect to convert solar energy into electrical energy.

The advantages of photovoltaic energy is given below

- Solar energy is 100 percent renewable and will continue to radiate the earth for millions of years.
- It is environmentally friendly because it produces no greenhouse emissions or other pollutants.
- It may be utilized anywhere on the planet, reaching areas where power lines do not reach.
- Maintenance is inexpensive and straightforward, and there is no noise pollution.

The disadvantages of photovoltaic energy is given below

- Requires a high initial investment.
- Sufficient area is required for installation.
- The solar panels' efficiency is bounded.

This research considered some specific goals, such as:

- a. More efficient PV modules with cutting-edge technologies used in design consideration
- b. Reduced dependency on power utility and produce power from renewable sources
- c. Proposed system is designed with consideration of load profile analysis. So, that it can reduce purchase energy from the grid
- d. PV technology uses contribute in reducing pollution and emission to the environment
- e. capex/opex model analysis of PV system
- f. Implement simulation tools to optimized PV system design which reduce cost

Research work is divided into the following steps:

1. Design/Calculations Phase: Theoretical basic concepts considering legislation, regulations & grid-code
2. Development Phase: Data analyzed; load profile analyzed
3. Chosen Technology: Type of modules, voltage sizing, minimum and maximum number of panels per string, number of modules and inverters
4. Simulation with specialized software

II. RELATED WORKS

Very few works have been done related to this field in Bangladesh. Among them Kaptai 7.4 KWP (6.63 MW AC) grid connected solar PV power plant is the challenging one. Over 3 million house hold are now covering by solar PV system which is 8 to 20 watt per day. There some grid based project are under implementation such as Dharala 30MW and Panchagar 10 MW(IPP) which will cover the rural area’s daily need. There is another undergoing implementation project which is Rangunia’s 60 MW solar park [3].

In Bangladesh, 147 billion KW/S of solar electricity is required, which is well beyond our current projects. The Bangladesh Power Development Board (BPDB) has approved a 50 MW (AC) Solar Park by the HETAT-DITROLIC-IFDC Solar Consortium, which will cover the country's Mymensingh division. Electricity Generation Company of Bangladesh Ltd. (EGCB) is building a 50 MW solar power plant in Sonagazi, Chittagong [4].

Evacuation of power to grid will increase because of the short distance, load flows, short circuit, stability and harmonic distraction will decrease the loss of the main grid. Energon Technologies FZE and China Sunergy Co.Ltd (ESUN) are developing a 100 MW (AC) solar park, which will be sponsored by the Chinese government. The purpose of this project is to be aware of the grid voltage level and to re-project the desired capacity. These two projects, 3.28 MWs at Sharisabari Jamalpur and 25 MWk at Teknaf, are the most dependable and long-term projects [5]. Those projects necessitated the hiring of EPC contactor experts, the majority of whom came from the EU.

Another disadvantage of our geographical location is the fact that we are in the middle of nowhere. Global horizontal irradiance (GHI) is 6 KWH/M/day on average. The PPA of a 2x120 Mw peaking power plant in Siddirgonj solved this problem and used less land to achieve the best results. Among all of this labor, each power grid employs different strategies, but the unifying thread is that they are all attempting to keep their losses below 8%, which is a significant amount for every grid. The 109.77 MWp (82.5 MW AC) Solar Photovoltaic Grid Connected Power Plant at Sonagazi, Feni is now under construction. The project's goal is to reduce grid loss to 4.7 percent, which is the lowest in the field. Rather, more projects of this nature are swiftly being built in Germany and Denmark, with 346 GWh under development and another 525 GWh set to come online in 2023 [6].

III. CIRCUIT DIAGRAM AND PV POWER VARIES

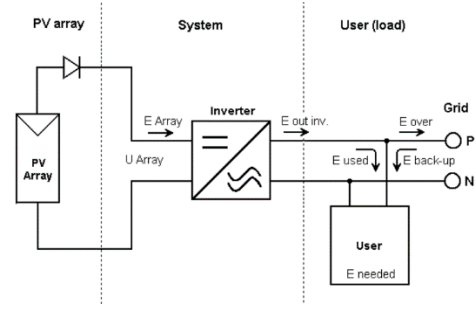


Fig. 1. Components of the System

A typical circuit diagram of PV system is given in Figure 1. In the PV installation, whenever there is an excess of energy production not used for the self- consumption needs of the household, that energy surplus will be injected into the grid. Otherwise, when there is not enough energy to cover the demand of the user, such as during night time, the energy consumed by the load will be extracted from the grid as back-up energy. For now, no storage system is going to be implemented in the first simulations. Thus, with all this explained, several simulation scenarios can be defined. To start, the inter-row spacing or pitch distance is going to be compared in different layouts to minimize the shadings. From the minimum distance calculated in section 9.2.4, with a value of 2,31m rounded to 2,3m a first scenario (Scenario A) will be built with the 35o tilt and a 0o azimuth orientation. That layout will allow the installation of 30 panels in the building rooftop in a configuration of up to 4 panels per string and 8 strings. It is represented in Section IV.

Then, the pitch distance will be increased to 3 m and to 3,3 m (Scenarios B and C), without modifying the tilt nor azimuth angles. The loss percentage is going to be evaluated at this point. Nonetheless, the number of PV modules is subsequently going to decrease as the pitch distance rises, so a whole assessment is required to find a balance point between losses and energy production. Once the best pitch distance is chosen, regarding as well the number of PV modules that fit into the rooftop, a change in tilt will be performed in the next simulation (Scenario D), in order to gauge the beneficial or detrimental effect of seasonal tilt adjustment. The PV modules in this scenario will remain oriented to the southern direction.

Afterwards, having already determined the optimal pitch distance, the number of panels, and whether the tilt should be fixed or adjustable, different tilt angles will be studied, by means of trying the initial tilt and adding and subtracting 10o from that value Scenarios E and F. Moreover, two simulation scenarios G and H with a different azimuth orientation for the modules are going to be compared as well. The modules in these scenes will be oriented south west and south east and aligned with the building orientation, maintaining the best tilt found in the previous scenarios.

IV. PARAMETERS

PVSYST 7.0.1	13/08/20	Page 1/6																										
Grid-Connected System: Simulation parameters																												
Project : LinyolaTFM																												
Geographical Site	Linyola	Country Spain																										
Situation	Latitude 41.71° N	Longitude 0.90° E																										
Time defined as	Legal Time	Time zone UT+1																										
Meteo data:	Albedo 0.20	Altitude 244 m																										
	Linyola	Meteonorm 7.2 (1997-2007), Sat=47% - Synthetic																										
Simulation variant : Scenario A																												
	Simulation date	18/06/20 21h49 (version 6.8.7)																										
Simulation parameters System type Tables on a building																												
Collector Plane Orientation	Tilt 35°	Azimuth 0°																										
Sheds configuration	Nb. of sheds 31	Identical arrays																										
	Sheds spacing 2.30 m	Collector width 1.25 m																										
Shading limit angle	Limit profile angle 29.4°	Ground Cov. Ratio (GCR) 54.4%																										
Models used	Transposition Perez	Diffuse Perez, Meteonorm																										
Horizon	Free Horizon																											
Near Shadings	Linear shadings																											
User's needs :	Ext. defined as file TFM CHR27 Sub-hourly.csv																											
	<table border="1"> <tr> <th>Jan.</th> <th>Feb.</th> <th>Mar.</th> <th>Apr.</th> <th>May</th> <th>June</th> <th>July</th> <th>Aug.</th> <th>Sep.</th> <th>Oct.</th> <th>Nov.</th> <th>Dec.</th> <th>Year</th> </tr> <tr> <td>611</td> <td>624</td> <td>639</td> <td>619</td> <td>626</td> <td>676</td> <td>703</td> <td>161</td> <td>640</td> <td>645</td> <td>623</td> <td>645</td> <td>7211 kWh</td> </tr> </table>		Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Year	611	624	639	619	626	676	703	161	640	645	623	645	7211 kWh
Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Year																
611	624	639	619	626	676	703	161	640	645	623	645	7211 kWh																
Power factor	Cos(phi) 1.000 leading	Phi 0.0°																										
PV Array Characteristics																												
PV module	CdTe Model FS-6450 Jun2019																											
Custom parameters definition	Manufacturer First Solar																											
Number of PV modules	In series 4 modules	In parallel 8 strings																										
Total number of PV modules	Nb. modules 32	Unit Nom. Power 450 Wp																										
Array global power	Nominal (STC) 14.40 kWp	At operating cond. 13.31 kWp (50°C)																										
Array operating characteristics (50°C)	U mpp 693 V	I mpp 19 A																										
Total area	Module area 79.2 m ²	Cell area 72.6 m ²																										
Inverter	Model SOFAR 12000TL-X																											
Original PVsyst database	Manufacturer SofarSolar																											
Characteristics	Unit Nom. Power 12.0 kWac	Oper. Voltage 160-960 V																										
	Max. power (= >25°C) 13.2 kWac																											
Inverter pack	Total power 12.0 kWac	Pnom ratio 1.20																										
	Nb. of inverters 1 units																											
Total	Total power 12 kWac	Pnom ratio 1.20																										
PV Array loss factors																												
Array Soiling Losses		Loss Fraction 3.0 %																										
Thermal Loss factor	Uc (const) 29.0 W/m ² K	Uv (wind) 0.0 W/m ² K / m/s																										
Wiring Ohmic Loss	Global array res. 609 mΩ	Loss Fraction 1.5 % at STC																										
Module Quality Loss		Loss Fraction -1.3 %																										
Module mismatch losses		Loss Fraction 0.8 % at MPP																										
Strings Mismatch loss		Loss Fraction -0.10 %																										

V. METHODOLOGY

The basic design of a PV solar plant is divided into several stages. Firstly, before pre-design we studied the basic theoretical concepts, regulations of grid-connected electricity systems.

In the development phase solar potential, plant site location is selected through detail data analysis. Then a specific area measured because solar panel will be mounted there. We have taken the load profile and account all the energy consuming devices during each hour of the day.

Then we take into consideration new types of modules with higher efficiency and I-V characteristics parameters. PV system electrical configuration based on: voltage sizing, minimum and maximum number of panels per string, number of modules and inverters. This will give us a gross estimation of the produced energy.

Specialized simulation software (PVsyst 7.2) simulates the electrical behavior and performance of this pre-design PV system. Different scenarios created with diverse

orientations and tilts angle. The final design derived after several iterative processes of comparing, discarding and changing parameters on the simulations. Losses due to shadings and electrical connection also examine on this step. The best layout design with application of a storage system is now optimized which produce the maximum energy. Our studied research is based on efficiency and performance.

Design/Calculation Phase

Theoretical Concepts---- At a temperature of 5778 K, the Sun is an element that behaves like a blackbody. Through fusion reactions, it is an excellent absorber and emitter of a relatively constant amount of radiation from its surface. When solar radiation reaches the atmosphere, it is influenced by a number of significant mechanisms that reduce its power and split it. These are radiation absorption (by dust and air molecules), scattering, and reflection [6]. As a result, solar radiation is broken down into three types: direct, diffuse, and reflected [7]. The Sun's azimuth, and hence its radiation, varies constantly based on the season of the year and the hour of the day. The deflection angle is the angle measured between the Earth's axis and a straight line connecting the Earth's center and the Sun's center. It is given in Figure 2.

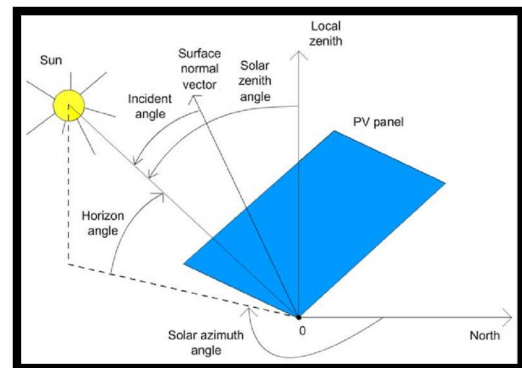


Figure 2: Angles considered for solar orientation

Regulations---- When we design the self-consumption PV system for household or commercial space, there are some object we need to think about which can make the whole system more efficient. First we have to consider the solar module type. Multiple solar cells are grouped and linked by bus bar connectors to increase their output power, resulting in a PV module. Similarly, when modules are joined in series, PV panels are formed, which are then connected to form solar strings or arrays. The majority of PV modules are stable units with a non-reflective front surface to absorb the most amount of sunlight possible. The angle between two PV panel and cells are more important for solar module which can make the system 2% more efficient. Power rating, power tolerance and temperature coefficient also can change the PV module connection more efficient and productive.

Then we have to consider the quality and durability. This is the most important part to choose a PV module system. A 250W and 150W panel produce two different amounts of voltage but if we maintain the quality then both of them will provide same kind of efficiency which is almost 16.2%. Its compulsory to know the irradiance data evolution for sunlight and locate the area where the power of the ray is higher. For this measurement we did use the latest Meteonorm 8.0 which show us the perfection location for the setup based on an area. It is also shows us the worldwide data so that we can change the location and identify the more productive region for this area. This graph shows us the maximum and minimum radiation around the globe.

Data Analyzed-----In development phase of our project, details and data are analyzed in order to find the suitable location for building that can make the most out of the solar potential of the region. Our designed PV system is located in Dhaka. We have taken the average solar potential map by Solargis tools. GHI data available over long period 1999-2018. In Dhaka region the daily and yearly GHI is 4.5 and 1643 kW/m² respectively. It is represented in Figure 3.

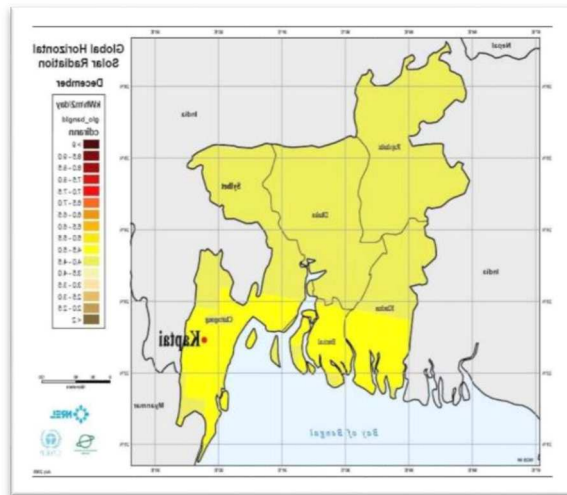


Figure 3: Global Horizontal Solar Radiation Map [7]

From the map (Fig.3) we can see the ray of sun is much higher in the north of the country which means the temperature is 1 or 2 degrees above compare to south. In that case we can consider north side more effective rather than south.

Till now Bangladesh is not much more dependable on solar power system because of the initial costing but scenario will change with time being. Almost 50-60% of the country power supplies from the coal, gas and oil those are responsible climate change and world temperature, so the world is thinking some other option like solar PV system.

For Bangladeshi perspective northern side of the country is produce more sunlight compare to other area. From the

PVsys website map we can see the effective region for solar PV installation. Recently, Bangladesh also implement a 1 MW solar PV system in Kaptai, we will try to reduce the costing such big project. Bangladesh has the longest sea beach; we can make that more productive by implementing more power plant based on nature.

We know the monthly solar irradiance, temperature. Irradiance and other meteorological data are obtained for Dhaka found in Meteonorm 8.0, a meteorological database containing worldwide weather data (1991-2010).

Site Studied-----The Site of study for this project is a Grid-Connected Solar PV Power Plant placed in Kaptai, Kaptai Upazilla. The exact location of this site is given in Fig. 4.



Figure 4: Kaptai proposed site and 11kV power evacuation line [7]

Load Profile-----The measurement of electrical demand across time is represented by load profiles. When developing a self-consumption system, it is critical to understand the user's load profiles in order to scale the system and calculate factors like performance and level of independence. Furthermore, these profiles provide information on the load's highs and lows.

The load profile (Figure 5) will always be determined by the user type, the region, local vacations, and so on. As a result, identifying the electrical load data is critical for measuring and assessing equipment requirements, servicing, and modifications, as well as determining when the device can serve the user and when it cannot.

An initial load profile will be developed for the chosen property. The site we looked at would most likely have peaks in the morning and afternoon, because everyone will be at residence during working hours, and just standby usage will be recorded on the profile.

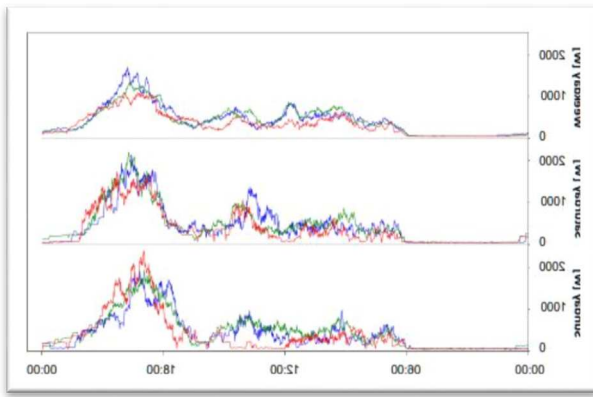


Figure 5: Load Profile of Site [10]

Chosen Technology-----Silicon-based solar cells remains the most popular type of solar panel, despite being the first generation of PV technology. Although it is not yet outdated, it has undergone additional development, and the number of technological types available today is extremely diverse. However, silicon is the most prevalent solid element on Earth [8], accounting for nearly 28% of the planet's total mass. This, combined with the fact that it is processed and manufactured, creates a powerful combination. After more than 30 years of use, technology is now quite sophisticated and widely established. Silicon-based solar cells are still the most efficient and have the lowest failure rate (less than 1%) with a market share of around 87 percent of PV technologies in 2019, it is the most appealing.

The second generation of PV technology is referred to as Thin-film technology. This technology is developed from a desire to make equipment thinner and inexpensive in order to save money, as well as to use greener and more environmentally friendly procedures. The open-circuit voltage of a solar cell increases as the thickness of the cell is reduced due to the drop in the saturation current [9]. One or more thin layers of light-absorbing substance are applied to a metal, plastic, or crystal backing support to generate these cells. They are, however, less effective. Inverters are required to change the kind of output current depending on the application. Photovoltaic systems or installations, on the other hand, generate a direct current. Distribution is also accomplished out by using DC inside the installations, but when sent to the grid, it must be in the form of alternating current, which requires the use of an inverter to convert it into oscillating compatible and synchronized with the electrical grid [10].

Simulation----- Once an estimated pre-design is done, then it can be modelled into the specialized software that will help to simulate the electrical behavior and performance of the PV system installed at the plant site. We are using simulation software PVsyst version 7.2 which allowed creating the PV layout to find the most accurate values for the parameters to the system and developing a shading and loss analysis. By using this

software various simulations will be run so that we can compare each value, hence the system design can sequentially be optimized.

Design Process-----At **first**, we have specified the desired power or available area. Then we chose the PV module from the internal database. For this we had to import PAN file to the software. Then we chose the inverter from the internal database. For this we had to import dot OND file to the software. Inverters are necessary to change the output current according to its end use. Then the software PVsyst proposed an array/system configuration that will allow us to conduct a preliminary simulation.

In the **second** stage we import the location in PVsyst because the location of the site is not on the software's database. A Meteo file for Kaptai is created using the latitude and longitude of the place, and irradiance and temperature data are imported from Meteonorm 8.0.

Data source: Meteonorm 8.0 (1991-2019), Sets=100%

	Global horizontal irradiation kWh/m ² /mth	Horizontal diffuse irradiation kWh/m ² /mth	Temperature °C	Wind Velocity m/s	Linke turbidity []	Relative humidity %
January	107.3	62.4	17.3	0.79	6.339	76.9
February	117.5	63.7	21.5	0.91	5.884	69.6
March	158.1	83.7	26.6	0.92	6.423	65.1
April	159.1	88.8	28.9	1.70	7.000	71.9
May	167.4	102.2	29.9	1.50	7.000	74.2
June	141.7	96.9	29.3	1.29	7.000	81.9
July	130.1	94.6	29.2	1.40	5.887	82.5
August	139.0	88.5	29.0	1.10	5.217	83.8
September	123.1	72.7	28.1	0.79	5.315	86.3
October	118.1	70.5	27.1	0.92	5.776	81.9
November	111.2	60.7	23.1	0.70	6.632	77.8
December	101.2	59.2	18.8	0.69	6.982	79.2
Year	1573.7	943.9	25.7	1.1	6.288	77.6

Global horizontal irradiation year-to-year variability 4.9%

Figure 6: Meteorological Data introduced on PVsyst

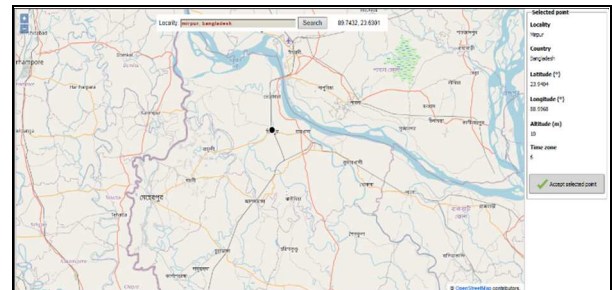


Figure 7: Interacting Map on PVsyst

The **third** stage is about module tilt and orientation. In our case we have used fix modules for economic reasons. The amount of diffusion of radiation from Sun is depending on tilt angle of the module and also the type of plane. For this we use the Albedo option into the software. Here, we have to set a common value 0.30 because our plane is concrete and the value given for concrete is 0.25 - 0.35. The installation of tracking system reduces in short period of time. Therefore, given a fixed structure for the panels, their tilt angle shall be defined. After that we must set a common value for that we have to select "copy" monthly values in copy from. MET file.

That is how we get the same value for every month of the year. We also set project site- Meteo default maximum search area 10km. So that if the software can not find the exact location it will provide temprature information for nearest 10km from our desired location. Then we set Design Conditions. Here, we input the lower temperature of Kaptai for last 5 years which is 26°C. We set winter operating temperature for Vmpp Max design is 28°C. The other 2 values are default values in this software. Those values are fixed. Then we chose IEC (1000 V) in the Array Max voltage which is the International Standard. Also, we have chosen μ Voc value from Specification, because we are not using one-diode model. Transposition Model for this project will be sophisticated. For Circumsolar treatment we chose Separate treatment because in PVsyst the circumsolar is treated the same way as the sunbeam which is coming from the direction of the Sun. This also means that the shading loss on isotropic diffuse is lower, so the shading loss on beam and the circumsolar is higher. Since our aim is to decrease the amount of loss. So, we set Limits overload loss for design is 1%. Since all places does not have same orientation so the azimuth angle varies from building to building. For that we set tilt plane angle to 25°C so that the loss with respect to optimum is 0%. If we change the plane angle the percentage of loss will increase. Also, we set azimuth angle to 0°C. If we want to tilt it to West we have to change the azimuth angle to negative value and if we want to tilt it to East we have to change the azimuth angle to positive value.

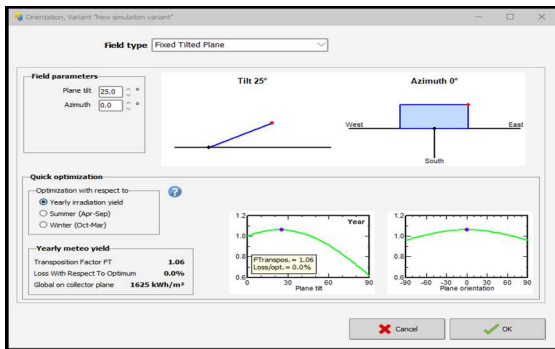


Figure 8: Orientation, variant of Azimuth angle

From the above figure we can see that the Transposition Factor FT is only 1.06. We can also see that there are 2 graphs, one is for Plane Tilt and another one is for Plane Orientation.

The **final** stage is the Row spacing. First of all, for an optimal layout of the PV modules, we have to implement the right minimum inter-row distance, so there are no desired shadows between them. The power that reaches a PV module depends not only on the power of the sunbeams but also the tilt angle between the panel and the Sun. Power density will reach a maximum peak whenever the panel is totally perpendicular to the sunlight. So the angle between the panel and the sun have to be 90°C. If

there is no losses the power density of the absorbing surface would equal the one from sunlight but due to Sun's constantly changing position and to various types of other losses this power density is always less than the incident one.

After completing all the stages, we can "Run Simulation".

VI. RESULT ANALYSIS

Three different scenario A B and C were tested 3D scene and software data and each parameter were run according the table shown as below which is our initial step for the result analysis, for first three cases the angle 0 degree to 35 tilt angle were changed. The distance was given below 2.3 m to 3.3 m continuous process. Rest parameters were constant as before to reduce the loss of the system.

Energy production represent the total effective energy from the system output. Yearly production is the total outcomes without losses from the grid which will be counted on the grid connected calculation. PR is the difference between the input and output energy ratio and finally the losses from the total system is consider as the shading loss.

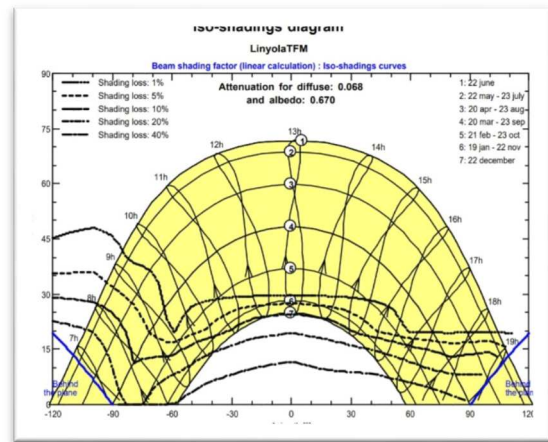


Figure 9: ISO-Shadings Diagram

According to the shading diagram different scenario create different losses from the angle. From the figure particular months shows the different shading loss. According to this more panel is installed in the area will give the more energy performance ration and global system efficiency values which means less harmful for nature.

Storage System Operating---There two different modes for different times when PV system feeds the battery and the sun goes down to back up the system and complete the user demand. Once the PV system is fully charged the system supply the users demand and it surplus the grid.

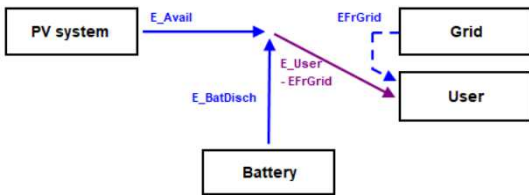


Figure 9: simplify scheme for charging mode

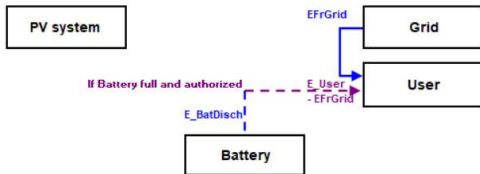


Figure 10: Storage mode during night mode

When the sun sets, there will be no other means to acquire energy from the sun, leaving the grid and batteries to satisfy the demand of the consumers. Grid storage and battery provide direct energy to users. We chose a 12.84 KW lithium battery with a storage capacity of 252 Ah and a voltage of 51.2v, and we linked the lithium batteries in parallel to ensure optimal charging capacity.

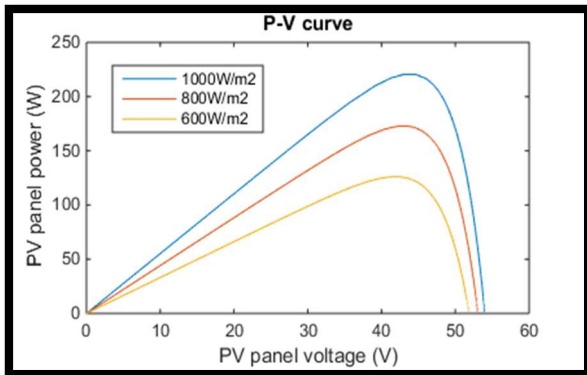


Figure 11: P-V Curve

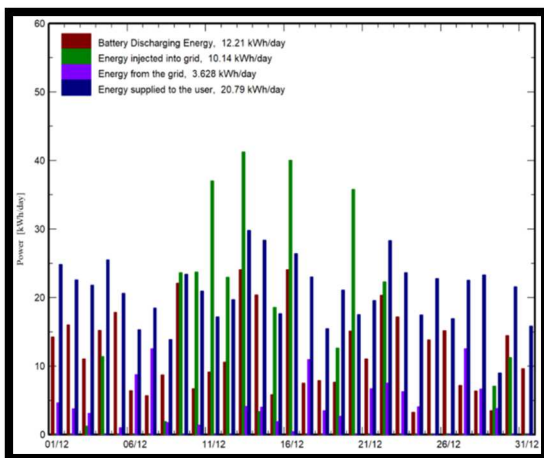


Figure 12: Comparison of charging and supplying the energy to the grid

Consider it as 20.35 MW inverter output which generate 22.7% is stored 4.36 Mw yearly to supply the family demand and rest are used to fulfill the demand of hydro based power grid in Kaptai. Only 7.6% time needs users purchase to bring outer energy. That means 346.2 Kw yearly purchase energy from outside means only 5.3% loss from the system. This outcome considers as the less possible loss from the system.

Sub-variant Result

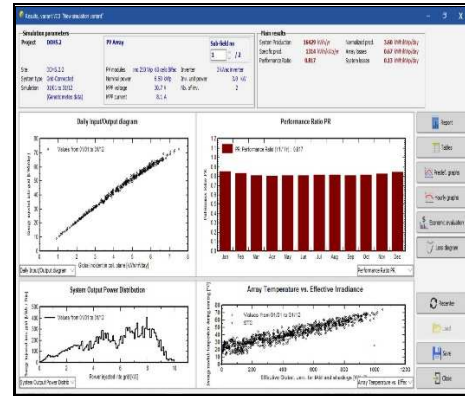


Figure 13: Simulation-1

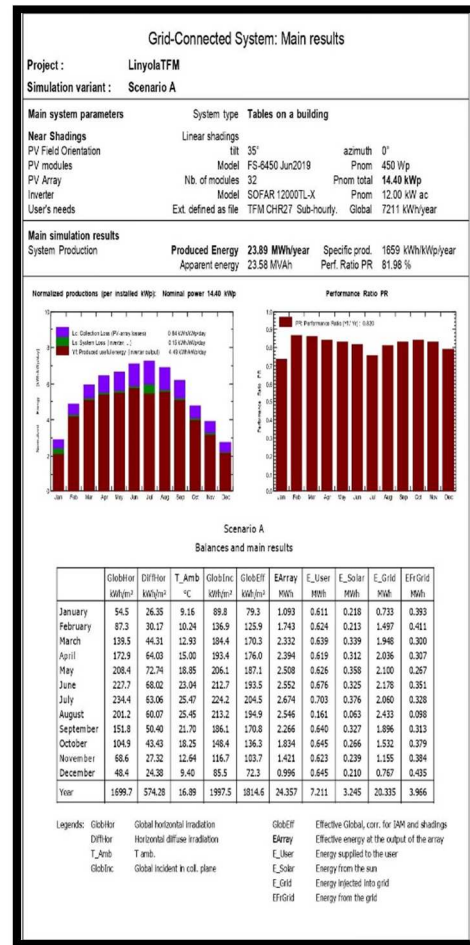


Figure 14: Main simulation results

This data chart shows us every monthly production of the year based on different variable and set the variable differently how the losses change. For example, if we choose the June which is produce the 227.7 GHI (Global Horizontal Irradiance) 68.2 DFI (Horizontal diffusion irradiation) and rapidly energy use by the consumer and energy injected in the grid and supplies from the grid. From this we can calculate the monthly and yearly losses and efficiency from the system.

Loss Diagram Report

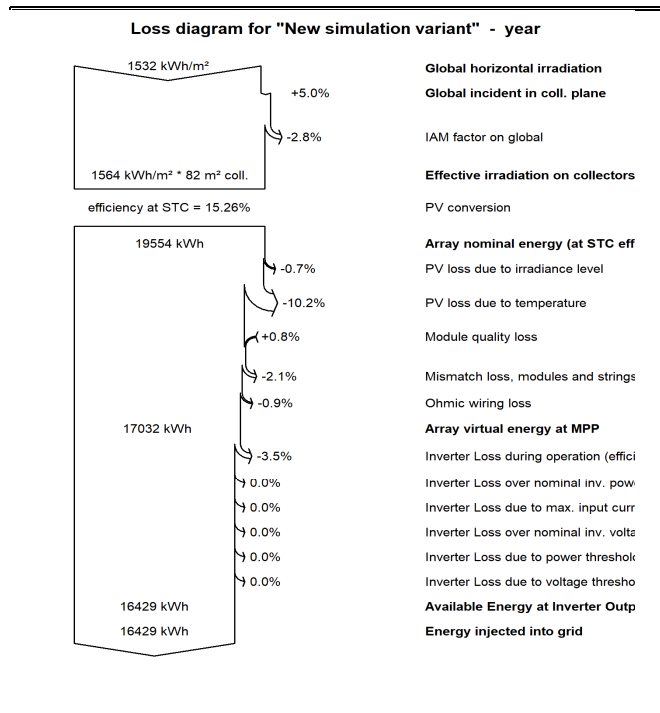


Figure 15: Loss Diagram

From the loss diagram we can observe that the PV loss due to irradiance level is -0.7%, the PV loss due to temperature is -10.2%, and the mismatch losses are fixed at the default value of a power loss at MPP of -3.5%. In terms of unavailability of the system, a constant loss of 5% is assumed. We can also see that there are Ohmic wiring loss on the DC circuit are set to a loss fraction at STC of -0.9%, while the voltage drops across the diodes and the losses on the AC circuit after the inverter are not considered.

VII. FUTURE WORK

Since the construction of the Kaptai Dam began in 1962, the site for this project has been unoccupied. As a result, there will be no loss of livelihood or agricultural land to relocate or assign as a result of this project. The system lacks a sensitive environment as well as historical or cultural significance, both of which are considered

additional value. The 7.4 MWp will be added to the national grid, resulting in only 5.3 percent system loss, which is the lowest loss from a PV system-based grid in Bangladesh [11].

We did work on the tilt angle, stability and consults time of 7.6% during the discharge period. How will the photovoltaic technology work for the solar cells such as thinner film cells can meet the silicon crystalline and reduce the system loss upto 3.6% which is close to the ideal grid system for the solar PV. Also the costing, grid connection or transportation of the energy will lot less then now. Sizing AC and DC wires, loads and installing the sun tracking system or other Nano technologies for solar cell also consider the system more affordable.

Finally, this kind of energy we can install in household building and commercials building to reduce the need of energy consumption come from non-renewable sources such as oil, gas and nuclear system.

REFERENCES

[1] F. Meneguzzo, R. Ciriminna, L. Albanese and M. Pagliaro, "The Energy-Population Conundrum and its Possible Solution," Science Magazine, 2019.

[2] Energías Renovables Info, "Energías Renovables Info," 2017. [Online]. <http://www.energias-renovables.info/>

[3] Solar Park| National Database of Renewable Energy, SREDA, <http://www.renewableenergy.gov.bd/index.php?id=1&i=1>

[4] Effective Solar Photovoltaic activities in Bangladesh. M.R. Islam, et al. / International Energy-Journal-10(2009);29-36

[5] Development of Renewable Energy Technologies by BPDB. [http:// www.bpdb.gov.bd](http://www.bpdb.gov.bd)

[7] Power System Efficiency Improvement Project Installation of 7.4 MW Solar PV Grid Connected Power Generation Project at Kaptai. https://www.adb.org/sites/default/files/project-documents/37113/37113-013-smr-en_1.pdf

[8] N. S. Lewis, "Research opportunities to advance solar energy utilization," Science Magazine, vol. 351, 2016.

[9] A. U. Taesoo D.Lee, "A review of thin film solar cell technologies and challenges," Renewable and Sustainable Energy Reviews, vol. 70, pp. 1286-1297, 2017.

[10] Irene Arranz i López; Design and simulation of a grid-connected PV system for self-consumption, Master's Thesis, Graz University of Technology Austria.

[11] Draft Final Report Barriers to Implement Renewable Energy of Power Cell.pdf

Provision of Information and Detection Systems on Two-Wheeled Motorcycle Accidents

Andi Nur Faisal
Postgraduate Student
Department of Electrical Engineering
Universitas Hasanuddin
Makassar, Indonesia
andinurfaisal93@gmail.com

Amil Ahmad Ilham
Department of Informatics
Universitas Hasanuddin
Makassar, Indonesia
amil@unhas.ac.id

Syafaruddin
Department of Electrical Engineering
Universitas Hasanuddin
Makassar, Indonesia
syafaruddin@unhas.ac.id

Abstract—This current work proposed the system of detection and information delivery on accident locations to related parties in order to get medical aids punctually. This system was based on three major components as the data source inputs, namely the sensors of tilt/slope, vibration, acceleration, and coordinate locations. The data was then processed through Arduino microcontroller by applying a fuzzy logic algorithm using Tsukamoto method as a supporting system for accident status decision. The results of the decision were then transmitted to the data center in real time using the Wi-Fi module, and then send notifications to related parties. The test results revealed that the slope sensor accuracy was about 96.1%. Additionally, the use of fuzzy logic via Tsukamoto method was successfully implemented to determine the accident status by an accuracy value of 85.18%.

Keywords—*detection system, accident detection of tilt sensor, vibration sensor, location and acceleration sensor, fuzzy logic.*

I. INTRODUCTION

The sale number of two-wheeled motorcycles year by year is almost no less than 5 million units each year. Based on information gained from the Indonesian Motorcycle Industry Association (AIS), the accumulated number of two-wheeled motorcycle sales from 2009 to 2019 reached 75 million within 10 years. The large number of vehicle sales each year has triggered problem related to the highway traffic system. Such mentioned number can affect the traffic flow on the highway. In addition to the traffic factor, the drivers' undisciplined driving attitude has also caused accidents

The data by the Traffic Center of the Indonesian National Police (locally known as *Korlantas Polri*) recorded that there were 100,028 traffic accidents over 2020 in Indonesia. This number decreased approximately 14% from the previous year's record, by 116,411 cases. The Traffic Center of the Indonesian National Police also recorded that there were 113,518 minor injuries due to traffic accidents in 2020, declining about 45% from 206,447 victims in 2019. Meanwhile, the number of seriously injured victims was 10,751 in 2020, falling 14% from 12,475 in 2019. The number of deaths due to traffic accidents reached 23,529 cases in 2020, decreasing 8% from the previous year by 25,671 victims. Therefore, the average death cases due to

traffic accidents was 1,960 people per month whilst the average death was 65 people each day, or 2-3 people per hour. Surprisingly, these accidents were dominated by two-wheeled motorcycles [1].

The accident victims are defined as not merely those who died at the time of the incidents, but also those suffering from many serious injuries and minor injuries as a result of a traffic accident, which if they do not get safety and medical assistance quickly, it will end up with death or lifelong disability [2].

The traffic accidents also caused very high costs. Such loss due to traffic accidents in Indonesia during 2002 was estimated at IDR 41.4 trillion, which was 2.91% of gross Domestic product (GDP). Notwithstanding, the number of accidents recorded in Indonesia was merely about 8%, and most of unrecorded cases were accidents without death cases [3].

To deal with the problem of traffic accidents, the Government has already established an integrated service unit in handling traffic accidents. According to the existing system, there were still fundamental weaknesses needed to be addressed. The accident information system still required eye witnesses as senders of information to the related parties, in this case the police, hospitals, and family/relatives. Following this, other factors of delays in handling accidents also often occurred due to delays in information received by the police and the nearest hospital.

This study thereby aimed to establishing an analysis system for early detection of accidents on two-wheeled motorcycles that is able to report quickly and automatically, in order to assist handling accidents more responsively, and further to minimize the impact of accidents.

II. LITERATURE REVIEW

In this given section, I compared multiple studies related to accident detection systems. Several methods had been previously used by some researchers to detect accidents. These systems included location and accident detection utilizing smartphones, VANET (Ad-hoc network), mobile applications, gyroscopes, accelerometers and deep learning.

In [12], this current work used arduino, accelerometer (ADXL336), GSM and GPS as modules to detect accidents and send accident location information automatically. The specified threshold value was $x < 70$ so that if the slope of the vehicle was less than 70 degrees, the system would detect a severe accident and would automatically send an accident location message. This detection system did not analyze the fatality rate due to accidents such as considering the acceleration and collisions that occurred. Thus, this can cause bias detection.

A. Censor MPU-6050

The MPU-6050 is a censor module which has two functions namely, an accelerometer with a micro-electromechanical system (MEMS) and a gyroscope with a micro-electromechanical system (MEMS) on a chip. There are about 16 analog pins converted first to determine the axis, so that this censor can work optimally. The values of the x-axis, y-axis, and z-axis on this censor can be taken simultaneously at one time. This censor applies an Inter Integrated Circuit (I2C-bus interface) as a connection between the censor and Arduino [4].

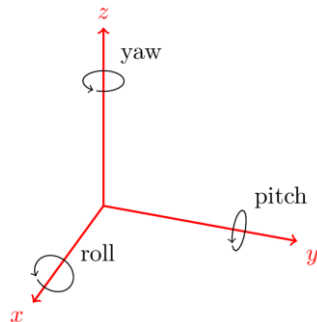


Figure 1. Titlt position or rotation of MPU-6050 module.

As seen in Figure 1, it depicts the location of the rotation or tilt of the MPU-6050 censor. Roll, pitch, and azimuth are reference points for the tilt of the system. The movement of turning upward and downward is known as roll. Then the downward and upward circular motion is called pitch. Moreover, the rotation of sideways horizontally is labelled azimuth

B. Global Positioning System (GPS)

GPS stands for Global Positioning System. It is a system for determining the position and global navigation using satellites. The system was first developed by The United States Department of Defense used for military and civilian purposes [5]. This module has a better performance as it has features of a data backup battery, a built-in electronic compass, and a strong signal catcher using the built-in antenna [6].

This existing study used the NEO-6M GPS module which could generate the data of coordinate locations and speed with an accuracy of 0.1 m/s [7].

C. Fuzzy Logic Algorithm

This kind of fuzzy logic method is a method developed based on the language of human reasoning, so that it can be utilized into several fields. Basically, this fuzzy logic method has a way to map an input space to an output space [8].

D. Fuzzy Tsukamoto

Fuzzy Tsukamoto is one of the methods of the Fuzzy Inference System. In the Tsukamoto method, every consequence of the if-then rule must be represented by a fuzzy set with a monotonous membership function [9].

- 1) *Fuzzification*: is the process carried out to change the system input that has a firm or crisp value into a fuzzy set, and determines the degree of membership in the fuzzy set
- 2) *The Formation of IF-Then rules*: is the process carried out to form rules that will be used in the form of IF-THEN stored in a fuzzy membership base.
- 3) *Inference* is the process of converting the fuzzy input into the fuzzification output of each predefined rule (IF-THEN).
- 4) *Defuzzification* is the process used to change the fuzzy output obtained from inference to a firm or crisp value. The final result is obtained by using the average weighting equation using the Average Weight Average method as shown in (1) as follows.

$$Z = \frac{\sum(\alpha_i \times z_i)}{\sum \alpha_i} \tag{1}$$

Notes : Z= Output Variable
 α_i = Value of α predicate
 z_i = Value of output Variable

III. METHODOLOGY

This accident detection system consists of two primary infrastructures, namely hardware and software interconnected with each other.

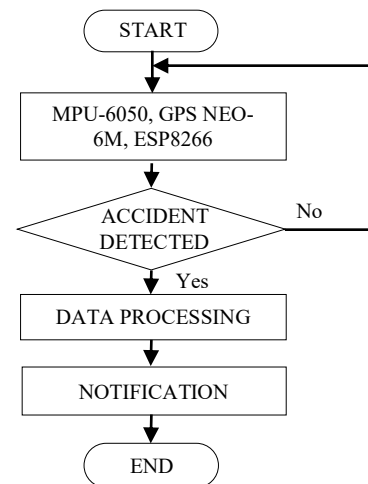


Figure 2. The Diagram Block of Accident Detection System of two-wheeled motorized vehicles.

In Figure 2, the block diagram consists of a hardware module as a data input source. The system will read the slope input data on the vehicle. If the slope exceeds the specified limit, the system will continue such process to data processing in order to find out the accident status and then send information via WhatsApp notification to related parties.

A. Hardware Design

In the hardware design, there are censors MPU-6050, Arduino Uno, GPS NEO-6M and ESP8266 as shown in Figure 3 below.

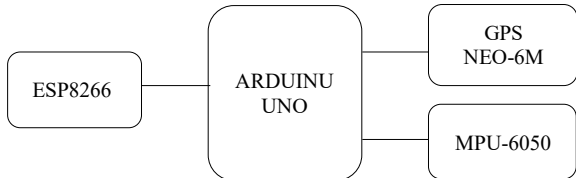


Figure 3. The design of Accident Detection System.

The MPU-6050 sensor was mainly used to detect tilt and vibration on a motorcycle. There was also a NEO-6M GPS sensor to detect the location of the coordinates and speed of the vehicle. Additionally, there was an ESP8266 module used as wifi. So it could be connected to a smartphone hotspot in order to send information.

B. Software Design

This system was designed by taking data values of slope, vibration, acceleration, and coordinate location. The slope value from the sensor was then analyzed using an Arduino Uno microcontroller with a maximum slope angle of 30°. If the slope angle exceeded this limit, the system could detect an accident [10]. Then, the system would continue the process into data processing.

Data processing used the fuzzy logic algorithm by applying the Tsukamoto method in order to get a more accurate accident status with the flow shown in the following Figure 4 as follows.

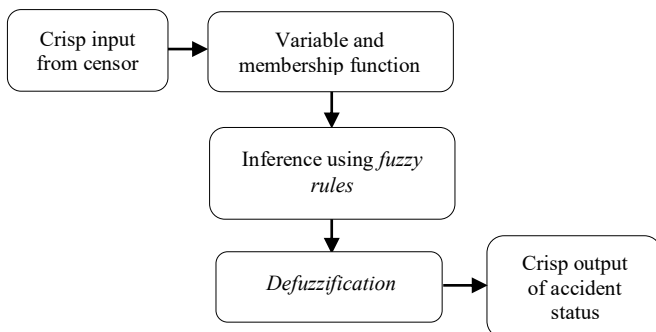


Figure 4. The System of fuzzy logic.

1) The Variable of fuzzy

In this study, the variables were divided into two, namely input variable and output variable. Each variable consisted of

three fuzzy sets. Each variable can be seen in the following Table I.

TABLE I. TABLE OF INPUT AND OUTPUT VARIABLE

Variable	Types	Values
Acceleration	Input	low, medium, high
Vibration		low, medium, high
Accident Level	Output	light, medium, heavy

2) The function of membership degree

a) Acceleration

Acceleration has three fuzzy sets, namely: $0 < \text{low} < 1.23$, $1.20 < \text{moderate} < 2.00$, $\text{high} > 1.90$ [11].

a. Low = $0 - 1.22$.

b. Medium = $1.21 - 1.90$.

c. High = $2.00 - 10.00$.

For each membership function, the degree of validity can be seen in Figure 5.

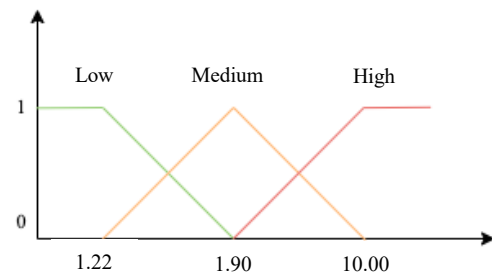


Figure 5. The function of membership acceleration.

In Figure 5, the membership of the acceleration variable indicates the membership function formulated with the following eqs. (2)-(4):

$$\mu_{PR}[x] = \left\{ \begin{array}{l} 1; \rightarrow x \leq 1.22 \\ \frac{1.90-x}{1.90-1.22}; \rightarrow 1.22 \leq x \leq 1.90 \\ 0; \rightarrow x \geq 1.90 \end{array} \right\} \quad (2)$$

$$\mu_{PS}[x] = \left\{ \begin{array}{l} 0; \rightarrow x \leq 1.22 \text{ atau } x \geq 10.00 \\ \frac{x-1.22}{1.90-1.22}; \rightarrow 1.22 \leq x \leq 1.90 \\ \frac{10.00-x}{10.00-1.90}; 1.90 \leq x \leq 10.00 \end{array} \right\} \quad (3)$$

$$\mu_{PT}[x] = \left\{ \begin{array}{l} 0; \rightarrow x \leq 1.90 \\ \frac{x-1.90}{10.00-1.90}; \rightarrow 1.90 \leq x \leq 10.00 \\ 1; \rightarrow x \geq 10.00 \end{array} \right\} \quad (4)$$

Notes : PR = Low Acceleration
PS = Medium Acceleration
PT = High Acceleration

b) Vibration

Vibration has three fuzzy sets, namely: low, medium, high.

a. Low = $15 - 30$ N.

- b. Medium = 30 – 45 N.
- c. High = 45 – 60 N.

For each membership function, the degree of reliability can be seen in the following Figure 6 as follows.

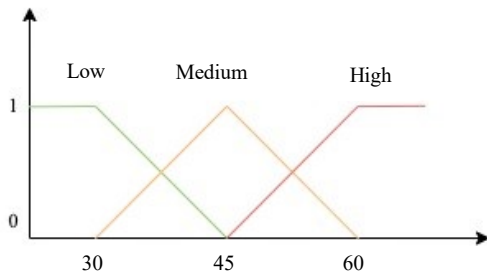


Figure 6. The function of vibration membership.

As seen in Figure 6, the membership of the vibration variable showed the membership function formulated with the following eqs. (5)-(7):

$$\mu_{GR}[x] = \begin{cases} 1; \rightarrow x \leq 30 \\ \frac{45-x}{45-30}; \rightarrow 30 \leq x \leq 45 \\ 0; \rightarrow x \geq 45 \end{cases} \quad (5)$$

$$\mu_{GS}[x] = \begin{cases} 0; \rightarrow x \leq 30 \text{ atau } x \geq 60 \\ \frac{x-30}{45-30}; \rightarrow 30 \leq x \leq 45 \\ \frac{60-x}{60-45}; 45 \leq x \leq 60 \end{cases} \quad (6)$$

$$\mu_{GT}[x] = \begin{cases} 0; \rightarrow x \leq 45 \\ \frac{x-45}{60-45}; \rightarrow 45 \leq x \leq 60 \\ 1; \rightarrow x \geq 60 \end{cases} \quad (7)$$

Notes : GR = Low vibration
GS = Medium vibration
GT = High vibration

c) Accident Status

The accident status output has three fuzzy sets, namely: light, medium, and heavy.

- a. Light = 15 – 30 N.
- b. Medium = 30 – 45 N.
- c. Heavy = 45 – 60 N.

For each membership function, the degree of reliability can be seen in the following Figure 7.

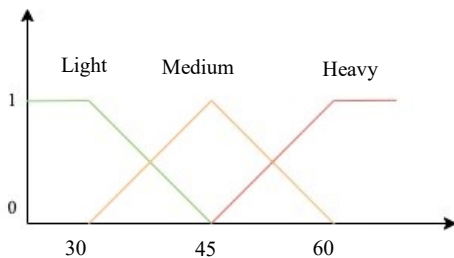


Figure 7. The membership function of accident status

As in Figure 7, the membership of the accident status variable showed the membership function formulated with the following eqs. (8)-(9):

$$\mu_{SKR}[x] = \begin{cases} 1; \rightarrow x \leq 30 \\ \frac{45-x}{45-30}; \rightarrow 30 \leq x \leq 45 \\ 0; \rightarrow x \geq 45 \end{cases} \quad (8)$$

$$\mu_{SKS}[x] = \begin{cases} 0; \rightarrow x \leq 30 \text{ atau } x \geq 60 \\ \frac{x-30}{45-30}; \rightarrow 30 \leq x \leq 45 \\ \frac{60-x}{60-45}; 45 \leq x \leq 60 \end{cases} \quad (9)$$

$$\mu_{SKT}[x] = \begin{cases} 0; \rightarrow x \leq 45 \\ \frac{x-45}{60-45}; \rightarrow 45 \leq x \leq 60 \\ 1; \rightarrow x \geq 60 \end{cases} \quad (10)$$

Notes : SKR = Low Accident Status
SKS = Medium Accident Status
SKT = High Accident Status

3) Fuzzy Rule

This stage contains rules generated based on two input variables, with the maximum number of rules is 9. Such rules state the input and output relations. The operator connecting both inputs is called AND. The implemented rules can be seen in Table II as follows.

TABLE II. TABLE OF FUZZY RULES

No	Variable		Accident Status
	Vibration	Acceleration	
1	Low	Low	Light
2	Low	Medium	Light
3	Low	High	Light
4	Medium	Low	Light
5	Medium	Medium	Medium
6	Medium	High	Medium
7	High	Low	Light
8	High	Medium	Medium
9	High	High	Heavy

C. The Testing of Detection System

The system testing was carried in order to obtaining data samples. Moreover, it as to measure the sensitivity level of each sensor. The sample data of research can be used as a reference to show the level of stability, accuracy, and feasibility of using the tool. This test included accelerometer testing, slope testing, GPS module testing, fuzzy logic algorithm testing, and information delivery testing respectively.

1) Vibration testing

Vibration is an input variable in the fuzzy logic analysis process that determines the accident status. This vibration testing was done by converting the accelerometer value into vibration, in order to get the vibration threshold value of the vehicles. The following Table III is the data generated from the accelerometer test when there was no accident which resulted in the resultant value working on each axis without a significant change.

TABLE III. THE TABLE OF ACCELEROMETER DATA WITHOUT ACCIDENT CASES

Time (s)	Acc-x	Acc-y	Acc-z	Resultant
1	0.08	1.62	9.08	9.22
2	0.32	1.82	9.95	10.12
3	0.45	1.74	9.88	10.04
4	0.12	1.23	10.87	10.94
5	0.15	1.45	9.89	10.00
6	0.22	1.34	9.88	9.97
7	0.34	1.88	10.89	11.06
8	0.45	1.98	11.89	12.06
9	0.55	1.99	10.89	11.08
10	0.43	1.78	9.88	10.05

Following this, the testing of the accelerometer was also carried out as an accident occurred. Table IV below revealed result data when an accident occurred causing the values in the 4th and 5th seconds of the y-axis and z-axis of the accelerometer. It experienced a significant change from the average value of each axis. Such change in value caused the resultant value becoming high.

TABLE IV. TABLE OF ACCELEROMETER DATA WITH ACCIDENT CASES

Time (s)	Acc-x	Acc-y	Acc-z	Resultant
1	0.08	1.62	9.08	9.22
2	0.32	1.82	9.95	10.12
3	0.45	1.72	9.88	10.04
4	0.12	3.23	18.76	19.04
5	-1.15	2.45	17.89	18.09
6	1.22	1.34	11.88	12.02
7	0.34	1.88	10.89	11.06
8	0.45	1.98	9.89	10.10
9	0.55	1.99	9.89	10.10
10	0.43	1.78	9.88	10.05

2) Slope/Tilt Testing

The slope data was obtained from the MPU-6050 module consisting of 3 axes, namely the x-axis, y-axis, and z-axis. To get slope data, the sensor was placed in the trunk of the motorbike, and then tilted the vehicle. The system detected the slope before further proceeding to the data processing in order to determine the accident status.

TABLE V. TABLE OF DATA AS MOTORBIKE FELL TO RIGHT SIDE

Time (s)	Pitch	Roll
1	-0.4°	-6.4°
2	-0.5°	-6.5°
3	-0.6°	-9.5°
4	-0.4°	-9.5°
5	-0.5°	-35.5°
6	-0.6°	-40.4°
7	-0.4°	-48.4°
8	-0.5°	-55.4°
9	-0.6°	-55.4°
10	-0.4°	-56.4°

The slope test on the vehicle in Table V shows a negative roll. Based on the data above, it can be concluded that the negative value on the roll surely indicated that the vehicle was falling to the right side.

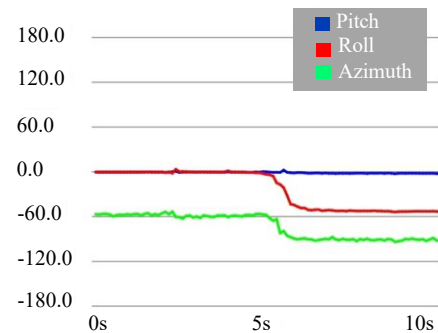


Figure 8. The slope change of X-axis, Y-axis, and Z-axis as motorbike was falling to right side .

As seen in previous Figure 8 above, the graph as the vehicle had an accident and tilted to the right side beyond the maximum slope limit. In addition, the testing was also carried out when the vehicle fell to the left as shown in table 6 below. When it fell to the left, the roll value showed a positive value. Moreover In Figure 9, the graph indicates the slope of the vehicle when it fell to the left side.

TABLE VI. TABLE OF DATA AS FALLING TO LEFT SIDE

Time (s)	Pitch	Roll
1	-0.4°	8.4°
2	-0.5°	6.5°
3	-0.6°	9.5°
4	-0.4°	10.5°
5	-0.5°	35.5°
6	-0.6°	40.4°
7	-0.4°	48.4°
8	-0.5°	55.4°
9	-0.6°	55.4°
10	-0.4°	56.4°

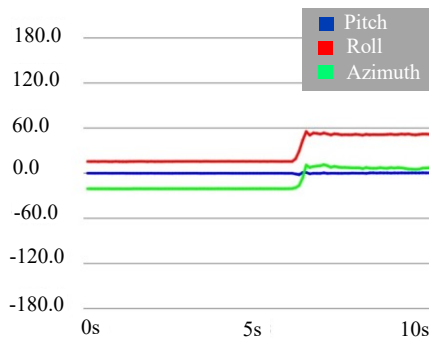


Figure 9. The slope change of X-axis, Y-axis, and Z-axis as motorbike was falling to left side .

3) GPS Testing

GPS testing aimed to determine the accuracy of GPS showing the location coordinates of the accident victim as compared to google maps. In addition, it was used to determine the speed of the vehicle.

4) Testing of Information delivery

Information delivery was the end point of the accident detection system. This test was carried out to ensure that the system could send information after an accident had been detected. Information was sent/transmitted in the form of short messages via WhatsApp.

IV. RESULTS AND DISCUSSION

After having tested the accident detection system for 26 trials, this system is regarded to be able to detect accidents with the results of slope testing by an accuracy of 96.1%. Furthermore, testing was also carried out towards the algorithm, namely fuzzy logic with the Tsukamoto method in order to determine the accident status with an accuracy value of 85.18%. Finally, the accident status information was successfully sent to the relevant parties.

V. CONCLUSION

In this current research, the detection and delivery system a two-wheeled motorcycle accidents had been carried out using Arduino Uno hardware, MPU-6050, GPS NEO-6M, ESP8266 hardware and a fuzzy logic system by using the Tsukamoto method to support the system of the accident status decision-making. Several stages of this research consisted of hardware design, software design and system testing. According to the test results, the slope sensor accuracy was by 96.1%, and the use of fuzzy logic with the Tsukamoto method was successfully applied to determine the accident status with an accuracy value of 85.18%. The information of both coordinate location and accident status were successfully transmitted using the ESP8266 module which was connected to a smartphone hotspot via a WhatsApp notification to related parties for quick response.

Some challenge faced in this research was, for instance, the limitation of testing on vehicles with higher speeds to get a better level of detection accuracy. In addition, this study used the wi-fi module to transmit accident information. So, in

such certain locations with poor internet connections there would be delays or failures in sending information.

VI. REFERENCES

- [1] National Police of Republic of Indonesia (locally known Polri), 6th November 2021. <https://korlantas.polri.go.id/>. (references)
- [2] N. Fathurrahman, A. Hendriawan, and S. Wasista, "Rancang Bangun Smart Vehicle untuk Mendeteksi dini Kecelakaan dan Keadaan Darurat," Electrical Engineering Study Program, State Electronic Polytechnics of Surabaya Campus PENS-ITS Sukolilo, Surabaya, Jan. 2011.
- [3] Directorate of Land Transportation, "Manajemen Keselamatan Transportasi Jalan, Naskah Workshop Manajemen Keselamatan Transportasi Darat," Directorate of Land Transportation, Department of Transportation of Indonesia, Jakarta, 2007.
- [4] InvenSense, "MPU-6000 and MPU-6050 Product Specification Revision 3.4," California, Aug. 2013. [Online]. Available: <https://invensense.tdk.com/wp-content/uploads/2015/02/MPU-6000-Datasheet1.pdf>
- [5] Winardi and S. Abdullah, "Pengenalan GPS dan Penggunaannya," Coral Reef Rehabilitation and Management Program (COREMAP), Jakarta, 2006. [Online]. Available: <https://docplayer.info/storage/64/51521504/1646902871/XeE60IIW M2UU0t8WqB61A/51521504.pdf>
- [6] Y. N. Rizalldhi, "Pelacakan Lokasi Sepeda Motor Menggunakan Modul GPS UBLOX NEO 6M DAN GSM SIM800L," *PROGRAM STUDI Tek. ELEKTRO Fak. Tek. Univ. MUHAMMADIYAH Surak.*, p. 14, 2019.
- [7] "NEO-6 series Versatile u-blox 6 GPS Datasheet," *u-blox*, 2011. [https://www.u-blox.com/sites/default/files/products/documents/NEO-6_DataSheet_\(GPS.G6-HW-09005\).pdf](https://www.u-blox.com/sites/default/files/products/documents/NEO-6_DataSheet_(GPS.G6-HW-09005).pdf) (accessed Mar. 13, 2022).
- [8] R. Munir, "Pengantar Logika Fuzzy," *Informati Engineering. - STEI ITB*, p. 95, 2007
- [9] S. Kusumadewi and H. Purnomo, *Aplikasi Logika Fuzzy Untuk Mendukung Keputusan*. Yogyakarta: Graha Ilmu, 2004.
- [10] A. Suprayogi, H. Fitriyah, and T. Tiyani, "Sistem Pendeteksi Kecelakaan Pada Sepeda Motor Berdasarkan Kemiringan Menggunakan Gyroscope Berbasis Arduino.," vol. Vol 3 No 3 (2019), pp. 3079–3085, Jan. 2019.
- [11] A. Ali and M. Eid, "An automated system for Accident Detection," in *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, May 2015, pp. 1608–1612. doi: 10.1109/I2MTC.2015.7151519.
- [12] Rajvardhan Rishi, Sofiya Yade and Keshhav Kunal, "Automatic Messaging System for Vehicle Tracking and Accident Detection" in *2020 Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)*, May 2020, pp. 831-834. doi: 978-1-7281-4108-4/20/\$31.00.

A Data-Centric Machine Learning Approach for Controlling Exploration in Estimation of Distribution Algorithms

Antonio Bolufé-Röhler

*Mathematical and Computational Sciences
University of Prince Edward Island
Charlottetown, Canada
aboluferohler@upe.ca
ORCID 0000-0002-3181-1864*

Jordan Luke

*Mathematical and Computational Sciences
University of Prince Edward Island
Charlottetown, Canada
jluke@upe.ca*

Abstract—Exploration plays a key role in the performance of metaheuristics. An algorithm should perform more exploration when reaching the “ideal search scale”; this happens when solutions are regularly sampled from different attraction basins. The moment this search scale is reached depends on the topological features of the objective function and the inherent randomness of the heuristic optimization process. Previous work on adjusting exploration have mostly used fixed rules based on fitness improvement, in this paper, we model it as a supervised machine learning problem. We apply a data-centric approach to understand whether variations in the data are more relevant than variations in the classification models. For our study we use the Estimation Multivariate Normal Algorithm with Threshold Convergence, which provides an ideal framework as it allows us to directly control exploration through the γ parameter. Optimization results show that the machine learning hybrid significantly outperforms the baseline algorithm.

Index Terms—metaheuristics, machine learning, data centric, threshold convergence, estimation multivariate normal algorithm

I. INTRODUCTION

The optimization of multimodal problems involves two distinct tasks: identifying promising attraction basins (exploration) and finding the local optima in these basins (exploitation). However, many metaheuristics were initially conceptualized for unimodal search spaces. For example, Particle Swarm Optimization (PSO) begins with a cornfield vector (see Section 3.2 in [1]) and Differential Evolution (DE) builds its foundation from a simple unimodal cost function (see Figure 1 in [2]). As a consequence, the primary search mechanisms of these metaheuristics are biased towards converging on local optima rather than searching for the best attraction basins.

Estimation of Distribution Algorithms (EDAs) are faced with a similar challenge [3]. In EDAs, the best solutions are used to estimate the parameters of the distribution, thus rapidly focusing on the best found attraction basins. The efficacy of exploration in EDAs relies strongly on the diversity of their populations. However, once the algorithm starts to converge, the variance of the distribution functions becomes smaller. With a smaller variance, new candidate solutions

are generated closer to each other, which in turn, leads to even smaller variance, which leads to a rapid convergence. As a consequence Estimation Distribution Algorithms may converge prematurely [4]. A variety of techniques have been developed to avoid premature convergence, including mechanisms such as diversification, speciation and sub-population formation [5], restarts [6], Threshold Convergence [7] and niching [8].

Threshold Convergence (TC) has been used to avoid premature convergence in an Estimation Multivariate Normal Algorithm (EMNA) [4]. The proposed algorithm, named Estimation Multivariate Normal Algorithm with Threshold Convergence (EMNA-TC), led to an improvement of approximately 50% in performance compared to the original EMNA. One of its key contributions was the design of an adaptive threshold function for controlling the transition from exploration towards exploitation, based on the results gathered by the algorithm during the optimization. In this paper we take the design of adaptive threshold functions for TC a step further, by designing a machine learning based adaptive threshold function. Such a function uses a classification model to determine how to adjust the threshold given the topological features observed during the optimization process.

The combination of metaheuristics with machine learning is currently one of the most successful trends in optimization [9]; in this paper, we continue this trend by designing a machine learning based hybrid of EMNA-TC that leads to a significant improvement in performance. However, the key contributions of this paper go beyond the mere improvement in performance, on one hand, we present a novel way of modelling the problem of controlling exploration as a supervised learning problem and we use a trained model for adjusting exploration directly during the optimization process. On the other hand, we pursue a data-centric approach when solving the classification problem. Such an approach has been promoted by Andrew Ng, the cofounder of DeepLearning.AI, Coursera, and LandingAI, who recently launched a campaign to support data-centric Artificial Intelligence [10].

This paper begins with a background on EMNA, Threshold Convergence and the hybridization of machine learning with metaheuristics. The problem of controlling exploration is then defined as a classification problem in Section III. Section IV presents the classification results for a variety of models on the different datasets. Section V presents the machine learning based optimization hybrid. Optimization results, including an analysis of how the models influence the hybrids, are presented in Section VI. The paper then concludes with a discussion of current results and future work in Section VII.

II. BACKGROUND

Estimation of distribution algorithms (EDAs) are metaheuristics that guide the optimization by building and sampling explicit probabilistic models of promising candidate solutions. Thus, EDAs transform the optimization problem into a search over probability distributions; in the specific case of the Estimation Multivariate Normal Algorithm (EMNA) a multivariate normal distribution is used for sampling the search space. In every iteration the mean μ and co-variance matrix Σ are calculated for the subset of best solutions in the current population. The next population is then calculated by sampling an inverse multivariate Gaussian function using μ and Σ [3].

A. Estimation Multivariate Normal Algorithm with Threshold Convergence

In globally convex search spaces, promising solutions tend to cluster around the best attraction basins and, as a result, with every iteration the variance of the estimated distribution will become smaller. New solutions sampled from a distribution with a smaller deviation will be closer to each other. Such close solutions will have an even smaller variance, which subsequently leads to even closer solutions in the next iteration. This auto-catalytic process is an important mechanism which allows EMNA to converge; however, it can also lead to a ‘‘cascading convergence’’ and cause premature convergence [4].

Premature convergence can be limited through the use of the Threshold Convergence (TC) technique. TC modifies the sampling strategy so that new solutions are created at least a ‘threshold’ distance away from its reference solution(s). This controlled sampling strategy allows management of the transition from exploration to exploitation, convergence is thus ‘held’ back until the last stages of the search process [7]. The name *Threshold* is a combination of the words threshold and held.

Threshold Convergence has been incorporated into multiple metaheuristics, significantly improving results when optimizing multi-modal functions. In most metaheuristics, TC has been used to control the distance among existing and new solutions directly in the search space (e.g. Differential Evolution [11], Particle Swarm Optimization [12] and Evolution Strategies [13]). However, in the case of EMNA, Threshold Convergence was used to control the Σ parameter of the distribution function. The resulting algorithm, EMNA-TC,

was presented in [4]. The application of TC in the parameter space did not only report state of the art performance, but also opened a new line of research into the application of TC to combinatorial optimization problems [4].

$$\Sigma_{norm} = init_ \Sigma_{norm} \times \left(\frac{totalFEs - FEs}{totalFEs} \right)^\gamma \quad (1)$$

In EMNA-TC the value provided by the threshold function is used to set the (Euclidian) norm of the Σ matrix. This is done by normalizing Σ and then multiplying it by the threshold value. Initially the norm is set to be equal to the norm of Σ calculated from initial population of randomly sampled solutions ($init_ \Sigma_{norm}$), and it is updated over the execution of a metaheuristic by following the decay rule shown in Equation 1. In this equation, $totalFEs$ is the total number of function evaluations, and FEs is the amount of evaluations performed so far. The parameter γ controls the decay rate, and thus controls the convergence rate in EMNA-TC.

Algorithm 1 EMNA-TC

```

Randomly initialize population  $P$ 
while  $FEs \leq maxFEs$  do
     $best\_solutions \leftarrow select\_best(P)$ 
     $[\mu, \Sigma] \leftarrow estimate\_parameters(best\_solutions)$ 
    if first iteration then
         $\Sigma_{norm} \leftarrow init\_ \Sigma_{norm} \leftarrow norm(\Sigma)$ 
    else
        if best solution has improved then
             $\gamma \leftarrow 0.95 \times \gamma$ 
        else
             $\gamma \leftarrow 1.05 \times \gamma$ 
        end if
         $\Sigma_{norm} \leftarrow init\_ \Sigma_{norm} \times \left( \frac{maxFEs - FEs}{maxFEs} \right)^\gamma$ 
    end if
     $\Sigma \leftarrow \Sigma_{norm} \times \frac{\Sigma}{norm(\Sigma)}$ 
     $P \leftarrow sample\_normal\_multivariate(\mu, \Sigma)$ 
end while
return  $best\_solution$ 

```

B. Controlling Exploration in EMNA-TC

An advantage of controlling the convergence rate is that convergence can be held to a pause when an ‘ideal search scale’ is reached. Arguably, this ideal search scale happens when solutions are regularly sampled from different attraction basins. If the distance between different attraction basins could be known *a priori*, then the norm of Σ could be held longer at this scale, allocating more function evaluations to this threshold level and thus increasing the chances of finding the best regions before performing local search [14].

Previous work has shown the challenges of determining the ideal search scale, focusing on the number of solutions that are kept from one iteration to the next, as an indicator of whether this ideal search scale has been reached [11]. However, because in EMNA the population is entirely replaced by the new solutions, a different strategy was used in EMNA-TC. This strategy used as a criterion whether the

best new solution improves over the best solution from the previous population [4]. Thus, if the best solution is improved, then γ is decreased by multiplying it by 0.95 (which slows convergence); otherwise, γ is increased by 1.05. A general outline of EMNA-TC algorithm with adaptive threshold is presented in Algorithm 1.

C. Machine Learning for Improving Metaheuristics

In this paper we present a novel way for adjusting γ using supervised machine learning models. The combination of metaheuristics with machine learning is currently one of the most successful trends in optimization [15]. There are a variety of approaches that use machine learning techniques for improving metaheuristics, e.g. machine learning methods can be used to cleverly generate the initial population [16], for building approximate models of the fitness function and replacing function evaluations [17], for setting up parameters [18] or managing the population of evolutionary algorithms [19]; among other applications [9].

In this paper we focus on a broader approach, where optimization can be understood as a decision-making process. In such an approach, the best search strategy needs to be chosen for a particular search region and moment during the optimization [20]. Such decisions must be made as the optimization proceeds, taking into account information gathered during the search. As discussed in [21], two distinctive factors influence these decisions:

- **The topology of the objective function:** different characteristics such as unimodality, multimodality, separability, deceptiveness, etc. are better optimized by different search strategies.
- **The optimization process:** given the stochastic nature of the heuristic optimization process, the same decision may be optimal at different moments, even for the same objective function.

The first step towards applying machine learning techniques in this approach is to model the decision problem as a machine learning problem, either supervised [20], unsupervised [22] or reinforcement problem [23]. Then, a model can be trained to identify the function's topological features, learn to generate a sampling sequence associated with a given search strategy, determine which strategies are better fitted for a given topology, and deciding when to apply them. In the case of this paper, a strategy refers to either decreasing γ to promote more exploration or increasing γ for a faster convergence.

III. CONTROLLING EXPLORATION AS A SUPERVISED LEARNING PROBLEM

In EMNA-TC the parameter γ determines the norm of Σ (Equation 1), thus by adjusting γ during the optimization process it is possible to indirectly control the exploration of EMNA-TC. The problem of finding the best value for γ at any given point during the optimization can be modeled as a supervised learning problem, in this paper, we will specifically focus on modelling it as a classification problem.

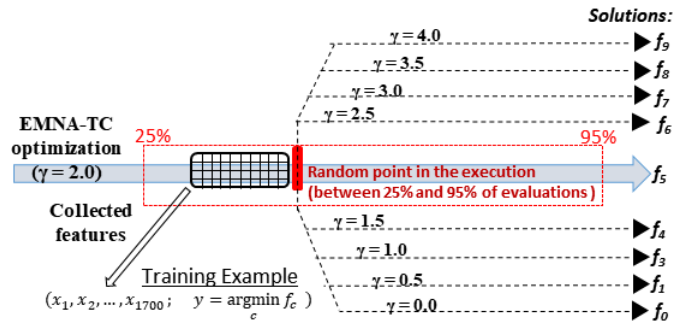


Fig. 1. Procedure for creating a training example for the V_0 dataset.

We modeled the problem as a multi-class classification problem, with 9 different output classes associated to different values for the γ parameter. As shown in Figure 1 classes 0, 1, 2, ..., 7, 8 correspond to the values of $\gamma = 0, 0.5, 1.0, \dots, 3.5, 4.0$, respectively. To create one training example we execute EMNA-TC starting with $\gamma = 2$, then at a random point of the execution between 25% and 95% of the total budget of function evaluations, EMNA-TC is branched out with all the 9 possible values for γ . The final result for each branch is obtained by completing the optimization with the remaining evaluations.

Information regarding the optimization status of the algorithm is collected before branching EMNA-TC. The collected features are intended to capture knowledge about the function topology and the optimization process itself. Seventeen features are collected in every other iteration of the 200 iterations prior to the branching point, thus leading to a total of $100 \times 17 = 1700$ features; the collected features are:

- Feature 1: The function evaluations used so far.
- Feature 2: The function evaluations available.
- Feature 3: The current value for the gamma parameter.
- Feature 4: The current threshold value.
- Feature 5: The standard deviation of the fitness in the current population.
- Feature 6: The average fitness of the current population.
- Features 7-17: The 0%, 10%, 20%, ..., 100% percentiles of the fitness of the current population.

As illustrated in Figure 1, we create a training example by associating the vector of collected features to the class that led to the best final result. As part of our data-driven approach, we decided to create different variations of the original dataset in order to assess the effect on the classification and optimization process.

A. Dataset variations

In the original dataset, which we denoted as V_0 , training examples that had multiple γ values producing the exact same final result were excluded. However, V_0 does include training examples which have a very small difference in the final results among the γ branches (i.e. between f_0, \dots, f_8 in Figure 1). Such training examples may mislead the classifier into learning decisions that are not relevant in practice.

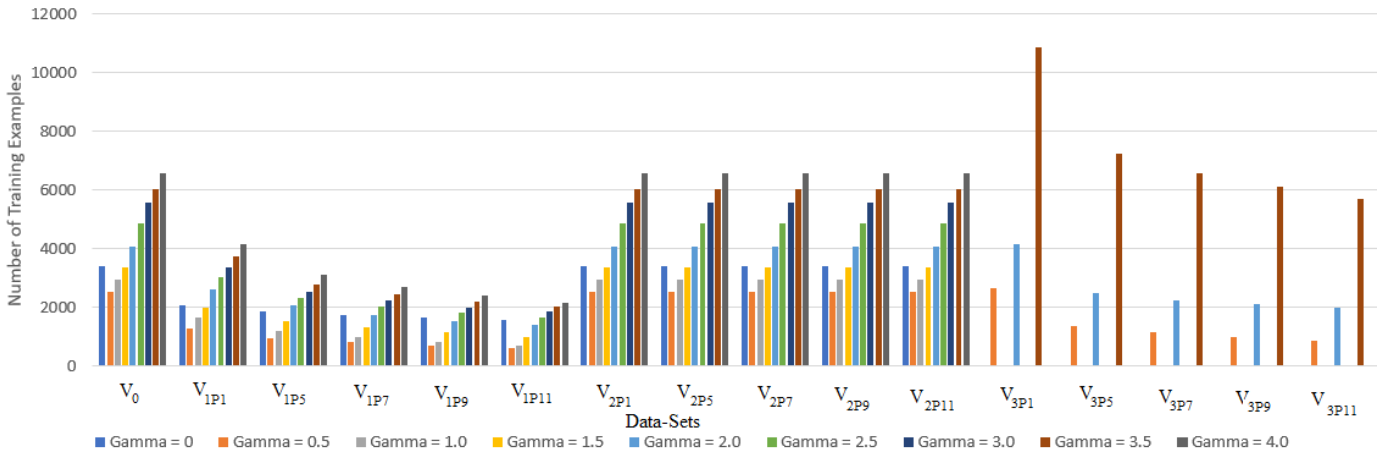


Fig. 2. Distribution of classes among training examples in the different datasets.

To tackle this issue the following variants used a “gamma index strategy” based on the difference between the best and worst final result achieved after branching. The strategy varies depending on the variant, but the decisions were based on whether the relative difference between the best and worst branching results were larger than 1%, 3%, 5%, 7%, 9% or 11%.

- Variant 1 (V_1): This variant uses the strategy of not including a training example if the difference between the best and worse results was less than the previously mentioned percentages. We denote each of these subtypes as $V_{1p1}, V_{1p3}, \dots, V_{1p9}$ and V_{1p11} .
- Variant 2 (V_2): A disadvantage of V_1 is that it removes training examples, thus reducing the size of the dataset. V_2 uses a different strategy, instead of removing the training example, it keeps the training example but indicating that no relevant action is required. This is done by setting the prediction class to 5, i.e. $\gamma = 2.5$, which is the median value among all the collected γ values in the dataset. Similarly to V_1 , variants subtypes are denoted as $V_{2p1}, V_{2p3}, \dots, V_{2p9}$ and V_{2p11} .
- Variant 3 (V_3): Variant 3 uses a similar gamma strategy as V_1 , but it only uses three classes, for $\gamma = 0.5, 2$ and 3.5 , selecting the best option (of the three) which had the best average result between it and its two nearest neighbours (eg. Average of results from $\gamma = 0, 0.5, 1$ for $\gamma = 5$).

The training data was gathered using the 28 functions from the IEEE CEC’13 benchmark [24]. This benchmark consists of a set of 28 unimodal and multi-modal functions with various characteristics. The process of creating one training examples is computationally expensive, as it requires completing a full execution of EMNA-TC plus the 8 branches. We were able to collect slightly over 1400 training examples per function, resulting in a total of 39382 examples in the V_0 dataset.

The number of training examples and the class predictions varied depending on the dataset variant. Figure 2 shows how

the number of training examples distributes along each class on each dataset. It can be noticed how the overall number of examples is reduced in datasets V_1 and V_3 as the percentages for the gamma strategy are increased. A surge of examples towards larger γ values can also be observed, which indicates a bias towards faster convergence. In multi-modal functions training examples are balanced, but in multi-modal functions training examples are skewed into increasing γ , and thus convergence. Future data collection should focus more on multi-modal functions.

IV. CLASSIFICATION RESULTS

We tested over 20 different classification models, more data and the project code can be found in the Github repository [25]. In this paper we focus on the most relevant and best performing models; six of those models are classic machine learning models from the Sklearn library, and two are deep neural networks implemented using TensorFlow Keras. These models are:

- Decision Trees (DT): Sklearn implementation with default settings.
- K Nearest Neighbour (KNN): Sklearn implementation with the number of neighbours set at 20, the weights set to distance, otherwise default settings.
- Radial Nearest Neighbour (RNN): Sklearn implementation with default parameters, the radius was set to 2600.
- Linear Discriminant Analysis (LDA): Sklearn implementation with solver set to ‘svd’ and otherwise default settings.
- Quadratic Discriminant Analysis (QDA): Sklearn implementation with default settings.
- Logistic Regression (LR): Sklearn implementation with the solver set to ‘sag’, the maximum number of iterations set at 100000, and otherwise default settings.
- Deep Neural Network (TF3): a Sequential Keras model with five layers. Output dimensions 64, 32, and 9 for the first, third and fifth layers. The second and fourth layer being Dropout Layers with a 0.1 dropout rate. The

first layer has a *relu* activation function the last having a *sigmoid* activation function.

- Deep Neural Network (TF4): a Sequential Keras model with three layers. Output dimensions 256, 128, and 9; the first layer has a *relu* activation function, the last layer has a *softmax* activation function. Both TF networks use Adam optimizer with a learning rate of 0.01 and a categorical crossentropy loss function.

For testing the classifiers we used a holdout-validation set with 25% of the dataset. Table I shows the accuracy of the different models for the best dataset subtypes; we choose the subtype based on the overall performance when we applied the hybrid to the optimization benchmark (see next section). Accuracy for Variant 3 is higher because it has only 3 classes, while Variants 0, 1 and 2 have 9 classes. It is noticeable that the average accuracy for all the models doesn't vary much among the different datasets, with V_1 and V_2 having an almost identical average. The average per model shows that RNN, DT and LOG perform slightly better than the rest.

When looking at individual models it can be appreciated that some models such as RNN and LOG show a good performance for all the dataset variants. Other models, such as DT and QDA, seem to be more sensitive to the different datasets. Despite being one of the best performing models overall, DT performs rather poorly in V_3 , while QDA (the worst performing model), performs very poorly in V_0 and V_3 . It is also noticeable that the Deep Network models don't perform as well as other models, this could be due to the relatively small size of the datasets. Although the TF3 model does perform reasonably well and with a stable performance among all the datasets. In the following sections, we will analyze how these accuracies correlate to the optimization performance.

TABLE I
CLASSIFICATION ACCURACY FOR VARIANT TYPES WITH BEST OPTIMIZATION PERFORMANCE

Model	V_0	V_1p_5	V_2p_3	V_3p_9	Model Average
RNN	0.182	0.188	0.182	0.656	0.302
TF4	0.151	0.167	0.164	0.660	0.286
TF3	0.167	0.167	0.164	0.660	0.290
DT	0.184	0.201	0.187	0.638	0.302
KNN	0.153	0.172	0.161	0.630	0.279
LOG	0.185	0.186	0.185	0.659	0.304
LDA	0.180	0.191	0.179	0.609	0.290
QDA	0.116	0.181	0.131	0.466	0.224
Dataset Average	0.165	0.182	0.169	0.622	

V. THE MACHINE LEARNING HYBRID

The EMNA-TC machine learning hybrid starts the execution of EMNA-TC with $\gamma = 2.0$. At every other iteration the algorithm collects the features described in Section III; this information is then used to build the input features required by the classifier. Once the algorithm reaches the 10% of function evaluations the trained classifier is called regularly every 10 iterations. The classifier is used to make a prediction

of what the optimum value for γ is at that given time in the optimization; then γ is set to this value and the optimization resumes.

These regular updates of the γ parameter allow EMNA-TC to adjust its exploration directly from information gathered during the optimization process. Ideally, γ would be decreased (which means the threshold will decay more slowly) when the algorithm reaches a threshold size that promotes the sampling of solutions from different attraction basins. To guarantee a final convergence, the hybrid stops calling the classifier once 70% of the function evaluations are reached, and γ is set to 4 to promote a fast convergence during the final iterations. A pseudocode of the machine learning hybrid is shown in Algorithm 2.

Algorithm 2 EMNA-TC-ML(classifier)

```

Randomly initialize population  $P$ 
while  $FES \leq maxFES$  do
     $best\_solutions \leftarrow select\_best(P)$ 
     $[\mu, \Sigma] \leftarrow estimate\_parameters(best\_solutions)$ 
    if first iteration then
         $\Sigma_{norm} \leftarrow init\_Sigma_{norm} \leftarrow norm(\Sigma)$ 
    else
         $input\_data \leftarrow collect\_features()$ 
        if  $0.1 \times maxFES < FES < 0.7 \times maxFES$  then
            if every 10 iterations then
                 $\gamma \leftarrow classifier.predict(input\_data)$ 
            end if
        else
            if  $FES > 0.7 \times maxFES$  then
                 $\gamma \leftarrow 4.0$ 
            end if
        end if
    end if
     $\Sigma_{norm} \leftarrow init\_Sigma_{norm} \times \left( \frac{maxFES - FES}{maxFES} \right)^\gamma$ 
     $\Sigma \leftarrow \Sigma_{norm} \times \frac{\Sigma}{norm(\Sigma)}$ 
     $P \leftarrow sample\_normal\_multivariate(\mu, \Sigma)$ 
end while
return  $best\_solution$ 

```

A. Machine Learning Based Adaptive Threshold Function

As stated in [7] "The ideal case for Threshold Convergence is to have one sample solution from each attraction basin, and for each sample solution to have the same relative fitness with respect to its local optimum". To achieve this ideal case, an adaptive threshold function should flatten once the algorithm reaches a scale that allows sampling random solutions from different attraction basins. This scale is directly related to the size and shapes of attraction basins.

In the Rastrigin function, where local optima are located at the integer values of a regular grid of size 1, we know that the minimum distance between adjacent local optima is 1 and the maximum distance is \sqrt{n} (n being the number of dimension) [14]. In such case, the threshold function should have a smaller γ and thus a slower decrease, when the threshold value is in the $[1, \sqrt{n}]$ interval. Because in EMNA-TC the threshold value is set over the parameter space and not over the search space, it becomes more difficult to determine

when the average sampling distance of new solutions is within a given search interval. However, it is still possible to illustrate how different models vary the threshold decay.

Figure 3 shows the threshold values for the best two models (RNN and TF4, both shown here for V_0) and for the worst performing model (QDA) on the Function 12 of the CEC'13 benchmark, i.e. the Rotated Rastrigin function. The figure also shows the fixed threshold decay as a baseline comparison. Because the Rastrigin function is a sinusoid super-imposed on a quadratic base function, a fixed threshold decay with $\gamma = 2$ provides a good convergence schedule. However, delaying convergence when solutions are sampled in the $[1, \sqrt{n}]$ interval can still lead to a slight improvement in performance. This logic is captured differently by the different models, RNN for instance, deviates little from the very effective quadratic decrease, it converges slightly faster up to the iteration 150, and at that point reduces the convergence speed. The TF4 model delays convergence earlier with a faster decrease in the second half of the optimization. The QDA model delays convergence far too much and convergence is only achieved once the model stops being used and γ is set to 4 in the final iterations.

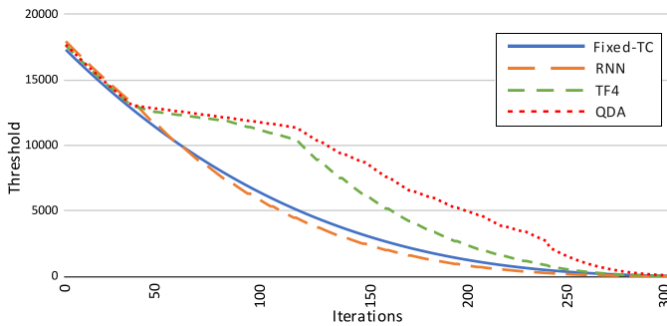


Fig. 3. Adaptive threshold functions.

Even though the accuracy among some models doesn't seem to vary much, there is a clear difference among the adaptive threshold functions each model has learned. This difference is the underlying reason behind the significant variations in optimization performance observed among the machine learning hybrids in the next section.

VI. OPTIMIZATION RESULTS

In this section we present and analyze the optimization performance of the EMNA-TC machine learning hybrids. Results are presented using the 28 functions from the IEEE CEC'13 benchmark suite [24]. This benchmark consists of a set of 28; these are divided into three sets: unimodal functions (1 to 5), basic multi-modal functions (6 to 20) and composite multi-modal functions (21 to 28).

The experimental setup consists of 30 independent runs on each function on 30 dimensions with a maximum of 300,000 function evaluations, as recommended in the benchmark description [24]. Performance is measured by comparing the mean error. The relative difference in performance between

two given algorithms a and b is also presented using equation: $100(b - a) / \max(a, b)$. These values indicate by what amount (percentage) algorithm a outperforms algorithm b .

A. Best Optimization Performance on the Dataset Variants

Table II shows the improvement (relative difference) of the machine learning hybrids when compared to EMNA-TC over the entire benchmark. As in Table I, results are presented for the best subtype of each dataset. It is noticeable that the RNN-hybrid shows a steady performance, clearly outperforming EMNA-TC on every dataset. These results are consistent with RNN's strong accuracy performance in Table I. Conversely, the TensorFlow models, which weren't among the best classifiers, achieve the second and third best average optimization results. It is less surprising to notice that models such as QDA and LDA, which performed poorly as classifiers, also perform poorly when integrated into the optimization hybrid.

TABLE II
RELATIVE IMPROVEMENT OVER EMNA-TC FOR BEST VARIANT TYPES

Model	V_0	V_{1p5}	V_{2p3}	V_{3p9}	Model Average
RNN	21.67%	18.04%	13.06%	15.32%	17.02%
TF4	2.79%	18.58%	13.66%	22.70%	14.43%
TF3	7.07%	5.87%	7.53%	6.79%	6.82%
DT	4.08%	3.82%	4.84%	6.03%	4.69%
KNN	-0.09%	4.35%	0.76%	13.27%	4.57%
LOG	2.35%	0.32%	4.44%	4.63%	2.94%
LDA	3.69%	-9.53%	5.50%	2.97%	0.66%
QDA	2.26%	-10.02%	5.48%	-20.68%	-5.74%
Dataset Average	5.48%	3.93%	6.91%	6.38%	

Perhaps the most noticeable result comes from looking at the individual differences of each model for specific dataset variants. Table I didn't show much variation in classification accuracy across the datasets, however, the optimization performances vary more significantly. For instance, the RNN-hybrid performs very well across all the dataset variants and achieves the overall second best performance in the V_0 dataset with an improvement of 21.67%. The TF4-hybrid, on the contrary, varies its performance significantly depending on the dataset. This hybrid achieves the best overall improvement of 22.70% when trained on the V_{3p9} dataset, while performing quite poorly (2.79%) when trained on the original V_0 dataset. In a similar note, some hybrids perform very badly at specific datasets, e.g. the LDA-hybrid when trained with the V_2 variant and the QDA-hybrid for V_3 .

B. Effect of the Classification Accuracy on the Optimization Performance

To better illustrate how the accuracy of the classifiers relates to the optimization performance we have plotted both values in Figure 4 for the whole rank of dataset variations. The accuracy plot for the RNN model is very similar to that of other models, with a stable accuracy around 0.2 for all the variants and subtypes with 9 classes, and then an increase to approximately 0.65 for the V_3 subtypes. The

relative difference for the RNN-hybrid also shows a very stable performance, outperforming EMNA-TC in the range of 10% to 20% in all of the dataset variants.

Because RNN achieves a relatively high accuracy, and the highest average optimization improvement among all the models, we could hypothesize that there is a direct correlation between both measures. However, when looking at the plot for Logistic Regression (LOG) we can see that even though the accuracy plot is very similar (just slightly lower values), the optimization performance is clearly worse for the LOG-hybrid; the DT plot is similar to the LOG plot (not included for space reasons). This observation may lead us to the conclusion that there is either a weak correlation between classification accuracy and optimization performance, or that the optimization performance is very sensitive to small variations in accuracy.

The plots for the two TensorFlow models (TF3 and TF4) favour the hypothesis of a weak correlation between accuracy and relative difference. It can be noticed that while the accuracy plot is similar to that of most models, the optimization performance varies significantly: from performing slightly worse than EMNA-TC for some datasets (e.g. TF3 for V_1p_1 and V_1p_9) to achieving the overall best result (TF4 for V_3p_9). Meanwhile, the plots for QDA and KNN show a different story. Both these plots suggest a correlation between accuracy and performance, however, they suggest opposite relationships. In the case of QDA the plot seems to suggest an inverse correlation: the relative performance worsens when the accuracy improves. The KNN plot shows what would have been the most natural result to expect: optimization performance improving as classification accuracy improves.

Figure 5 presents a scatter plot for all the models, correlating the relative difference in optimization vs the classification accuracy. This figure confirms that there is no relationship between both measures for the V_0 , V_1 and V_2 datasets. But, it also shows that the relative performance for V_3 is overall better when compared to the other datasets. It also confirms that results of the QDA classifier are outliers with respect to the other classification models. It can also be noticed that 4 hybrids achieved an improvement of over 20% when compared to EMNA-TC, with the TF4-hybrid as the best performing one.

C. Full Optimization Results for the TF4-hybrid

Table III presents the optimization results for the TF4-hybrid on the entire CEC'13 benchmark. Results are compared against the EMNA-TC algorithm, for determining the effectiveness of the machine learning based adaptive threshold function. Table III also includes some well-known population-based metaheuristics such as PSO, DE and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES). CMA-ES is a state of the art algorithm, which uses a multivariate normal distribution for generating new solutions, default parameters were used as recommended in [26]. The PSO implementation is a standard version with a ring topology, zero initial velocities, "Reflect-Z" for particles that exceed

the boundaries of the search space and a population size of 50 individuals [27]. The DE algorithm is based on a DE/rand/1/bin version with a population size of 50, crossover $Cr = 0.9$, and scale factor $F = 0.8$ [2]. Results include the mean error over 51 runs and the relative difference (%-diff) in performance between a given algorithms vs. the TF4-hybrid; positive percentages indicate by what amount the algorithms outperform the hybrid, negative values indicate the hybrid performs better.

Results in Table III show that compared to EMNA-TC the hybrid achieves an overall improvement of 22.70% over the entire benchmark. It is noticeable that much of that improvement comes from the unimodal functions, this is due to the classifier's ability to promote convergence in this set of functions (as mentioned in Section III-A). However, because EMNA-TC is originally designed for optimizing multi-modal functions, it is no surprise that the set of unimodal functions is where the hybrid performs the worst when compared to CMA-ES, DE and PSO.

In the set of classic multimodal functions is where the TF4-hybrid performs the best when compared to other metaheuristics, achieving improvements of 61.31%, 66.09% and 61.54% when compared to CMA-ES, DE and PSO, respectively. When compared to EMNA-TC, the hybrid achieves an improvement of 16.88%, confirming the effectiveness of the adaptive threshold for controlling exploration in multimodal search spaces. In the set of composition multi-modal function, the TF4-hybrid achieves the highest improvement when compared to CMA-ES (37.65%) but almost no improvement (0.79%) when compared against EMNA-TC. This result indicates that most of the improvement comes from applying TC for controlling convergence, rather than from an adaptive threshold function. Also suggesting that for this type of functions, the quadratic threshold decrease from the baseline EMNA-TC algorithm is the most appropriate convergence schedule; probably due to the globally convex structure that many of these function have.

Overall, the TF4-hybrid's adaptive threshold function leads to a significant improvement over EMNA-TC, with most of this improvement coming from the unimodal functions. It also shows a significant improvement over CMA-ES, DE and PSO, with most of that improvement coming from both sets of multi-modal functions. It is also relevant to notice that there are 8 multi-modal functions, for which the TF4-hybrid achieves results that are at least 10% better than all of the other algorithms (bold results in Table III)

VII. CONCLUSIONS

This paper presented a novel way of controlling exploration by modeling it as a supervised learning problem that outputs an adaptive threshold function. Computational results (Table III) confirm the effectiveness of the approach, which can be easily adapted to other metaheuristics that use Threshold Convergence.

An analysis of the classification models and the corresponding hybrids (Tables I and II; Figures 4 and 5) suggests that

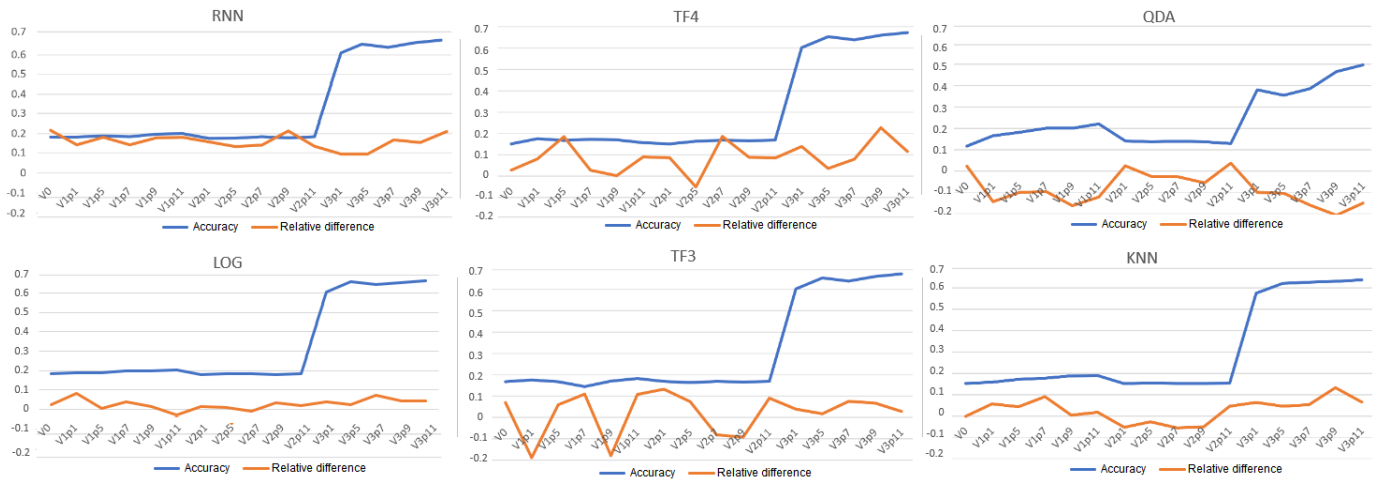


Fig. 4. Comparison between the classification accuracy of the models and the relative improvement of the corresponding hybrids vs. EMNA-TC.

TABLE III
COMPARISON OF EMNA-TC-ML(TF4) vs. EMNA-TC, CMA-ES, DE AND PSO .

Function	TF4-hybrid	EMNA-TC		CMA-ES		DE		PSO	
No.	Mean	Mean	%-diff	Mean	%-diff	Mean	%-diff	Mean	%-diff
1	1.79E-04	1.96E-03	-90.87%	0.00E+00	100.00%	4.17E-07	99.77%	0.00E+00	100.00%
2	1.27E+06	2.40E+06	-47.06%	0.00E+00	100.00%	3.91E+06	-67.45%	1.87E+06	-31.94%
3	1.64E+09	5.09E+09	-67.84%	2.07E+00	100.00%	2.14E+06	99.87%	9.02E+07	94.49%
4	2.25E-04	2.38E-03	-90.51%	2.57E+03	-100.00%	2.36E+04	-100.00%	1.67E+04	-100.00%
5	8.09E-04	3.98E-03	-79.68%	0.00E+00	100.00%	7.79E-05	90.37%	0.00E+00	100.00%
F1-F5			-75.19%		60.00%		24.51%		32.51%
6	2.40E+01	2.31E+01	3.54%	7.68E+00	67.98%	1.38E+01	42.46%	1.53E+01	36.20%
7	1.14E+00	4.74E+00	-75.85%	3.14E+01	-96.36%	2.13E+01	-94.63%	6.49E+01	-98.24%
8	2.09E+01	2.09E+01	-0.08%	2.15E+01	-2.75%	2.10E+01	-0.44%	2.09E+01	0.04%
9	2.03E+00	1.83E+00	9.63%	2.01E+01	-89.91%	3.80E+01	-94.66%	2.84E+01	-92.86%
10	1.92E-03	1.77E-02	-89.16%	1.38E-02	-86.09%	5.89E-01	-99.67%	1.19E-01	-98.39%
11	2.09E+00	2.65E+00	-21.24%	5.43E+01	-96.15%	6.12E+01	-96.59%	6.32E+01	-96.69%
12	1.63E+00	1.53E+00	5.81%	5.28E+01	-96.92%	2.19E+02	-99.26%	7.97E+01	-97.96%
13	1.39E+00	1.84E+00	-24.79%	1.24E+02	-98.88%	2.17E+02	-99.36%	1.36E+02	-98.98%
14	3.89E+02	4.36E+02	-10.71%	3.80E+03	-89.76%	2.68E+03	-85.49%	2.60E+03	-85.04%
15	2.79E+02	2.78E+02	0.46%	4.38E+03	-93.63%	7.26E+03	-96.15%	4.12E+03	-93.22%
16	3.09E-01	3.55E-01	-13.12%	8.91E+00	-96.53%	2.46E+00	-87.44%	1.63E+00	-81.05%
17	5.21E+01	6.42E+01	-18.91%	1.67E+02	-68.82%	1.13E+02	-53.91%	9.82E+01	-46.97%
18	9.26E+01	1.13E+02	-17.81%	2.85E+02	-67.52%	2.46E+02	-62.38%	1.72E+02	-46.19%
19	3.15E+00	3.15E+00	0.06%	3.25E+00	-3.17%	1.36E+01	-76.86%	5.72E+00	-44.98%
20	1.48E+01	1.50E+01	-1.07%	1.50E+01	-1.07%	1.29E+01	13.07%	1.17E+01	21.16%
F6-F20			-16.88%		-61.31%		-66.09%		-61.54%
21	2.97E+02	3.05E+02	-2.81%	3.19E+02	-6.93%	2.95E+02	0.64%	2.31E+02	22.20%
22	3.48E+02	3.67E+02	-5.18%	4.16E+03	-91.63%	2.27E+03	-84.66%	3.06E+03	-88.62%
23	4.82E+02	4.64E+02	3.76%	4.81E+03	-89.98%	7.30E+03	-93.40%	4.65E+03	-89.64%
24	2.08E+02	2.06E+02	0.69%	2.41E+02	-13.75%	2.38E+02	-12.66%	2.76E+02	-24.68%
25	2.05E+02	2.04E+02	0.26%	2.66E+02	-23.04%	2.50E+02	-18.11%	2.91E+02	-29.65%
26	2.97E+02	3.00E+02	-1.15%	3.11E+02	-4.61%	2.03E+02	31.57%	2.10E+02	29.21%
27	3.67E+02	3.72E+02	-1.49%	7.36E+02	-50.17%	9.77E+02	-62.46%	1.04E+03	-64.73%
28	3.01E+02	3.02E+02	-0.40%	3.81E+02	-21.11%	3.00E+02	0.19%	2.92E+02	2.86%
F21-F28			-0.79%		-37.65%		-29.86%		-30.38%
F1-F28			-22.70%		-32.89%		-39.56%		-35.85%

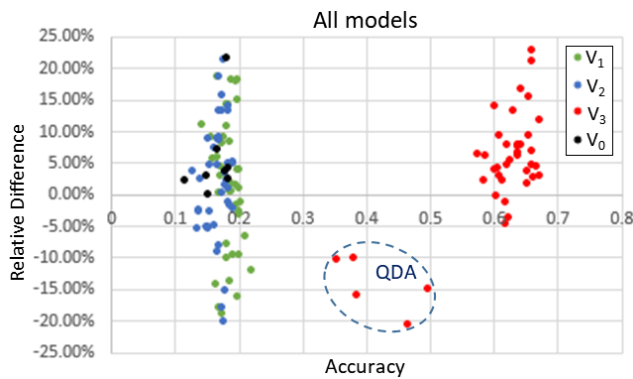


Fig. 5. Scatter plot of accuracy vs. optimization performance for all the models.

the correlation between the classification accuracy and the optimization performance is weaker than expected. However, the adaptive threshold functions produced by each model illustrate the contrast among the different models (Figure 3). These differences are reflected in the optimization results of the corresponding hybrids, even though they are not captured by the classification accuracy. Future work should dig deeper into this analysis and potentially suggest alternative metrics that could better correlate the classification and optimization results.

The data-centric analysis suggests that the difference among the machine learning models is more relevant than the variations in the data. Although slightly superior results for V_3 indicate that reducing the number of output classes leads to a better performance of by the hybrids. Future work should focus on further varying the datasets, e.g. by reducing dimensions through Principal Component Analysis or modelling γ output as a regression problem. Future work should also focus on compiling a larger dataset, which could significantly improve the performance of the Deep Learning models. Another promising line for future research is modelling the machine learning problems as a reinforcement learning problem. This would remove the need to collect a dataset, and would allow the metaheuristic to learn as it optimizes.

REFERENCES

[1] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[2] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[3] P. Larrañaga and J. A. Lozano, *Estimation of distribution algorithms: A new tool for evolutionary computation*. Springer Science & Business Media, 2001, vol. 2.

[4] D. Tamayo-Vera, A. Bolufé-Röhler, and S. Chen, "Estimation multivariate normal algorithm with threshold convergence," in *2016 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2016, pp. 3425–3432.

[5] M. Lozano and C. García-Martínez, "Hybrid metaheuristics with evolutionary algorithms specializing in intensification and diversification: Overview and progress report," *Computers & Operations Research*, vol. 37, no. 3, pp. 481–497, 2010.

[6] M. Kaucic, "A multi-start opposition-based particle swarm optimization algorithm with adaptive velocity for bound constrained global optimization," *Journal of Global Optimization*, vol. 55, no. 1, pp. 165–188, 2013.

[7] S. Chen, J. Montgomery, A. Bolufé-Röhler, and Y. Gonzalez-Fernandez, "A review of threshold convergence," *GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología*, vol. 3, no. 1, 2015.

[8] W. Sheng, S. Swift, L. Zhang, and X. Liu, "A weighted sum validity function for clustering with a hybrid niching genetic algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 6, pp. 1156–1167, 2005.

[9] M. Karimi-Mamaghan, M. Mohammadi, P. Meyer, A. M. Karimi-Mamaghan, and E.-G. Talbi, "Machine learning at the service of metaheuristics for solving combinatorial optimization problems: A state-of-the-art," *European Journal of Operational Research*, vol. 296, no. 2, pp. 393–422, 2022.

[10] "Andrew ng launches a campaign for data-centric ai," <https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/>, accessed: 2021-09-23.

[11] A. Bolufé-Röhler, S. Estévez-Velarde, A. Piad-Morffis, S. Chen, and J. Montgomery, "Differential evolution with threshold convergence," in *2013 IEEE Congress on Evolutionary Computation*, pp. 40–47.

[12] J. Chen, S. y Montgomery, "Particle swarm optimization with threshold convergence," in *Evolutionary Computation (CEC), 2013 IEEE Congress on*. IEEE, 2013, pp. 510–516.

[13] A. Piad-Morffis, S. Estévez-Velarde, A. Bolufé-Röhler, J. Montgomery, and S. Chen, "Evolution strategies with threshold convergence," in *Evolutionary Computation (CEC), 2015 IEEE Congress on*. IEEE, 2015, pp. 2097–2104.

[14] Y. Gonzalez-Fernandez and S. Chen, "Identifying and exploiting the scale of a search space in particle swarm optimization," in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, 2014, pp. 17–24.

[15] E.-G. Talbi, "Machine learning into metaheuristics: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–32, 2021.

[16] V. V. de Melo and A. C. B. Delbem, "Investigating smart sampling as a population initialization method for differential evolution in continuous problems," *Information Sciences*, vol. 193, pp. 36–53, 2012.

[17] S. F. Adra, A. I. Hamody, I. Griffin, and P. J. Fleming, "A hybrid multi-objective evolutionary algorithm using an inverse neural network for aircraft control system design," in *2005 IEEE Congress on Evolutionary Computation*, vol. 1. IEEE, 2005, pp. 1–8.

[18] I. Pereira, A. Madureira, E. Costa e Silva, and A. Abraham, "A hybrid metaheuristics parameter tuning approach for scheduling through racing and case-based reasoning," *Applied Sciences*, vol. 11, no. 8, p. 3325, 2021.

[19] L. Calvet, J. de Armas, D. Masip, and A. A. Juan, "Learnheuristics: hybridizing metaheuristics with machine learning for optimization with dynamic inputs," *Open Mathematics*, vol. 15, no. 1, pp. 261–280, 2017.

[20] A. Bolufé-Röhler and Y. Yuan, "Machine learning for determining the transition point in hybrid metaheuristics," in *2021 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2021, pp. 1115–1122.

[21] A. Bolufé-Röhler and D. Tamayo-Vera, "Machine learning based metaheuristic hybrids for s-box optimization," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14, 2020.

[22] J. García, B. Crawford, R. Soto, and G. Astorga, "A clustering algorithm applied to the binarization of swarm intelligence continuous metaheuristics," *Swarm and evolutionary computation*, vol. 44, pp. 646–664, 2019.

[23] A. Arin and G. Rabadi, "Integrating estimation of distribution algorithms versus q-learning into meta-raps for solving the 0-1 multidimensional knapsack problem," *Computers & Industrial Engineering*, vol. 112, pp. 706–720, 2017.

[24] J. Liang, B. Qu, P. Suganthan, and A. G. Hernández-Díaz, "Problem definitions and evaluation criteria for the cec 2013 special session on real-parameter optimization," *Computational Intelligence Laboratory, Zhengzhou University, Technical Report*, vol. 201212, no. 34, pp. 281–295, 2013.

[25] J. Luke and A. Bolufé-Röhler, "Emna-tc-ml-hybrid," <https://github.com/jmpluke/EMNA-TC-ML-Hybrid>, 2022.

[26] N. Hansen, "The cma evolution strategy: A tutorial," 2016.

[27] A. Engelbrecht, "Particle swarm optimization: Velocity initialization," in *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pp. 1–8.

Aggregated Modeling of Synchronous Generators Using Transfer Matrices

Arash Safavizadeh, Erfan Mostajeran, Seyyedmilad Ebrahimi, Taleb Vahabzadeh, and Juri Jatskevich

*Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada*

{arash.safavizadeh, mostajeran, ebrahimi, talebv, juriij}@ece.ubc.ca

Abstract— In power system analysis, sometimes the combined influence of a part of the network is desirable/investigated instead of the dynamics of the individual components therein. For such studies, aggregated models of system components would become beneficial when obtaining the equivalent circuits for the subject part of the network. This paper presents an aggregation method for modeling a cluster of synchronous generators (SGs) connected to a common bus in a power grid. First, it is shown how the so-called transfer matrices can be used for equivalent dynamic modeling of the system components, e.g., SGs with field and damper windings. Then, the transfer matrix-based models of SGs are aggregated to obtain an equivalent dynamic model for the cluster of SGs. The proposed aggregated model of SGs is then validated compared to their full-order qd models using computer simulations. It is verified that the proposed aggregation technique can preserve the dynamic behavior of parallel SGs very accurately for low- and high-frequency electromagnetic oscillations while being computationally more efficient than their classical full-order qd model counterparts.

Keywords— *Aggregated model, equivalent dynamic modeling, impedance-based modeling, synchronous generator (SG), transfer matrices.*

I. INTRODUCTION

Modeling large-scale power systems considering the details of all the components to analyze their dynamic behavior would be cumbersome and computationally demanding, especially in pace with the growth of switching converters [1]. As a matter of fact, the detailed dynamic behavior of all the internal states of some parts of the grid might be unnecessary in some studies [1]. In addition, with the rapid growth of distributed energy resources (DERs) in power grids, new dynamic behaviors and stability issues are emerging [2]. Thus, new and computationally efficient modeling methods are appreciated by planners and operators when conducting analyses to improve the stability and resiliency of the power grids considering the new dynamic behaviors [2]. Traditionally, the state-space modeling [3], [4] and impedance-based modeling [5], [6] were the two main approaches for considering the dynamic behavior of the system components. In the state-space modeling technique, all the system states are required to model the dynamic behavior of the whole system [6]. However, some manufactures may only provide the black-box model of their products [where enough information may not available for obtaining the details of all the states], or the parameters of some components cannot be measured using field tests [4], [5]. In such cases, the impedance-

(and/or admittance-) based modeling approach can be utilized in the form of a transfer function, from input voltages/currents to output currents/voltages [6], [7]. This modeling approach only includes the states corresponding to the input/output voltages/currents to the equivalent impedances/admittances. Consequently, such methods can predict only the dynamic behavior of the states at the interfacing points/connections [6].

One advantage of the impedance/admittance-based modeling technique is that it can provide the possibility of aggregation of the system components [5], [8]. Recently, various aggregation and reduction techniques have been introduced for DERs [1], [7], [9], and wind turbine generators (WTGs) [10], [11], due to their distributed nature. Even with the fast-growing DERs, synchronous generators (SGs) are still pivotal components of power systems as they constitute the dominant share of power generation. Therefore, leveraging aggregated models for SGs can alleviate the computational burden in dynamic simulations of large-scale power systems [5], [12].

Conventionally, modeling SGs in the rotating rotor reference frame (RRF) is more convenient as the varying rotor-position-dependent machine parameters become constant and their implementation for computer simulations would be more efficient and straightforward [3]. However, using RRF in SG modeling would encompass the inevitable dynamics of the reference frame, aka frame dynamics [5]. This introduces a significant challenge in using the impedance-based modeling techniques to aggregate SGs [5], [13]. Particularly, in wind farm studies where the same issue exists, two aggregation approaches have been introduced [11]–[14]. The first method [11], [14] assumes that the rotor speeds of all the SGs are equal, e.g., for those groups of wind turbines that receive equal wind speed. The second method [11]–[13] considers the differences in the rotor speeds of various SGs, which requires appropriate modeling of their mechanical system individually. However, for WTGs the investigations in the literature were mainly focused on the aggregation of doubly-fed induction generators (DFIGs) or permanent-magnet SGs (PMSGs), and the aggregation of wound-field SGs with damper windings has not been explored [11]–[14]. Furthermore, a reduced single equivalent machine model has been reported in [15] for aggregating coherent groups of SGs (which show similar rotor angle swings following a fault) for system-level studies. However, this reduced/aggregated model [15] cannot accurately predict the exact low- and high-frequency electromagnetic oscillations of the original system.

This paper proposes an aggregation technique for incorporating the dynamic models of wound-field SGs with damper windings (which are dominantly used in power systems) into a single transfer matrix-based equivalent dynamic model. With the assumption of a constant synchronous rotational speed, the transfer matrices are employed first for modeling each SG separately. Then, the parallel-connected SGs are transformed into an intermediary reference frame between individual machines where they are aggregated using simple mathematical operations. It is verified that the proposed method results in an aggregated transfer matrix model for a cluster of SGs which not only preserves the exact dynamic information of all the SGs for a wide range of frequencies, but also improves the computational performance in dynamic simulations compared to using multiple classical full-order qd models for SGs.

II. PROPOSED AGGREGATED MODELING METHOD FOR SGs

To present the transfer matrix modeling of an individual SG, and later the aggregation of transfer matrices for a cluster of parallel SGs, the full-order qd model of an SG is considered first. Here, RRF is used and it is assumed that the rotor variables are referred on to the stator side. Also, for simplicity but without loss of generality, only the field winding is considered on the d -axis (denoted by fd) and one damper winding is considered on the q -axis (denoted by kq). The procedure of calculating the equivalent parameters of the field and the one damper winding based on the machine time constants is described in Appendix A.

A. State-Space Model of the SGs in qd Reference Frame

The voltage equations of SGs can be expressed in the qd reference frame as [3, Ch. 5]

$$v_{qs} = r_s i_{qs} + \omega_r \lambda_{ds} + \frac{d\lambda_{qs}}{dt}, \quad (1)$$

$$v_{ds} = r_s i_{ds} - \omega_r \lambda_{qs} + \frac{d\lambda_{ds}}{dt}, \quad (2)$$

for stator windings, and as

$$v_{fd} = r_{fd} i_{fd} + \frac{d\lambda_{fd}}{dt}, \quad (3)$$

$$v_{kq} = r_{kq} i_{kq} + \frac{d\lambda_{kq}}{dt}, \quad (4)$$

for the field and damper windings. The flux linkages can also be expressed as

$$\lambda_{qs} = L_q i_{qs} + L_{mq} i_{kq}, \quad (5)$$

$$\lambda_{ds} = L_d i_{ds} + L_{md} i_{fd}, \quad (6)$$

$$\lambda_{fd} = L_{fd} i_{fd} + L_{md} i_{ds}, \quad (7)$$

$$\lambda_{kq} = L_{kq} i_{kq} + L_{mq} i_{qs}, \quad (8)$$

where

$$L_{mq} = L_q - L_{ls}, \quad L_{md} = L_d - L_{ls}, \quad (9)$$

$$L_{kq} = L_{mq} + L_{lkq}, \quad L_{fd} = L_{md} + L_{lfd}, \quad (10)$$

ω_r : rotor electrical rotational speed,

v_{qs}, v_{ds} : q - and d -axes stator voltages,

i_{qs}, i_{ds} : q - and d -axes stator currents,

$\lambda_{qs}, \lambda_{ds}$: q - and d -axes stator flux linkages,

r_s, L_{ls} : stator resistance and leakage inductance,

L_q, L_d : q - and d -axes stator inductances,

L_{mq}, L_{md} : q - and d -axes magnetizing inductances,

v_{fd}, i_{fd} : field winding voltage and current,

λ_{fd}, L_{fd} : field winding flux linkage and self-inductance,

r_{fd}, L_{lfd} : field winding resistance and leakage inductance,

v_{kq}, i_{kq} : q -axis damper winding voltage and current,

λ_{kq}, L_{kq} : q -axis damper winding flux linkage and self-inductance,

r_{kq}, L_{lkq} : q -axis damper winding resistance and leakage inductance.

B. Transfer Matrix Model of a Single SG

Here, it is considered (as a commonly used assumption) that all parallel SGs have a constant and equal rotational speed (i.e., $\omega_r = \omega_b$ for all SGs where ω_b is the nominal/base rotational speed). This can be a fair assumption considering the slower dynamics of the mechanical subsystems compared to the electromagnetic phenomena of interest. Yet, it facilitates the derivation of transfer matrices for SGs in the Laplace domain. Specifically, considering constant rotational speed for SGs avoids the convolution operation occurring in the frequency domain from the multiplication of two varying state variables in the time domain (i.e., the products of $\omega_r \lambda_{qs}$ and $\omega_r \lambda_{ds}$ in the qd model of SGs). Consequently, (1), (2) become linear ordinary differential equations with constant parameters, and the formulation of the SGs in the Laplace domain would be straightforward.

The field and the damper winding currents can be expressed in Laplace domain based on (3), (4), (7), (8) as

$$i_{fd}(s) = \frac{1}{r_{fd} + sL_{fd}} v_{fd} - \frac{sL_{md}}{r_{fd} + sL_{fd}} i_{ds}(s), \quad (11)$$

$$i_{kq}(s) = -\frac{sL_{mq}}{r_{kq} + sL_{kq}} i_{qs}(s), \quad (12)$$

where v_{fd} is assumed to be constant for SGs in this paper, and v_{kq} is equal to zero for the damper winding. Using (11), (12), the q - and d -axes stator voltages can be expressed in Laplace domain based on (1), (2), (5), (6) as

$$v_{qs}(s) = A(s) \cdot i_{qs}(s) + B(s) \cdot i_{ds}(s) + C(s) \cdot v_{fd}, \quad (13)$$

$$v_{ds}(s) = D(s) \cdot i_{qs}(s) + E(s) \cdot i_{ds}(s) + F(s) \cdot v_{fd}, \quad (14)$$

where

$$A(s) = \left(r_s + sL_q - \frac{s^2 L_{mq}^2}{r_{kq} + sL_{kq}} \right), \quad (15)$$

$$B(s) = \left(\omega_b L_d - \frac{s\omega_b L_{md}^2}{r_{fd} + sL_{fd}} \right), \quad (16)$$

$$C(s) = \left(\frac{\omega_b L_{md}}{r_{fd} + sL_{fd}} \right), \quad (17)$$

$$D(s) = \left(\frac{s\omega_b L_{mq}^2}{r_{kq} + sL_{kq}} - \omega_b L_q \right), \quad (18)$$

$$E(s) = \left(r_s + sL_d - \frac{s^2 L_{md}^2}{r_{fd} + sL_{fd}} \right), \quad (19)$$

$$F(s) = \left(\frac{sL_{md}}{r_{fd} + sL_{fd}} \right). \quad (20)$$

Based on (13), (14) and considering the voltages as inputs and currents as outputs, the SG model in the Laplace domain can be expressed in transfer matrix form as

$$\begin{bmatrix} i_{qs}(s) \\ i_{ds}(s) \end{bmatrix} = \overbrace{\begin{bmatrix} Y_{qq}(s) & Y_{qd}(s) \\ Y_{dq}(s) & Y_{dd}(s) \end{bmatrix}}^{\mathbf{Y}_{qd}} \begin{bmatrix} v_{qs}(s) \\ v_{ds}(s) \end{bmatrix} - \overbrace{\begin{bmatrix} G_{qfd}(s) \\ G_{dfd}(s) \end{bmatrix}}^{\mathbf{G}} v_{fd}, \quad (21)$$

where

$$\begin{bmatrix} Y_{qq}(s) & Y_{qd}(s) \\ Y_{dq}(s) & Y_{dd}(s) \end{bmatrix} = \begin{bmatrix} A(s) & B(s) \\ D(s) & E(s) \end{bmatrix}^{-1}, \quad (22)$$

$$\begin{bmatrix} G_{qfd}(s) \\ G_{dfd}(s) \end{bmatrix} = \begin{bmatrix} A(s) & B(s) \\ D(s) & E(s) \end{bmatrix}^{-1} \begin{bmatrix} C(s) \\ F(s) \end{bmatrix}. \quad (23)$$

Here, the \mathbf{Y}_{qd} matrix with Y_{ij} ($i,j=q,d$) entries is the admittance matrix specifying the relationship between q - and d -axes stator currents and voltages, and the \mathbf{G} matrix with G_{ik} ($i=q,d$, and $k=fd$) entries specifies the relationship between q - and d -axes stator currents and the field voltage.

C. Aggregation of SGs with Transfer Matrices

To demonstrate the general concept of aggregation of the transfer matrix-based models of SGs, it is assumed that N parallel SGs are connected to a power grid, as depicted in Fig. 1. It is assumed that the rotor angle (δ) of each SG can be obtained from a power flow analysis. Knowing the rotor angle of SGs, the variables can be transformed between different qd reference frames as

$$\begin{bmatrix} u'_q \\ u'_d \end{bmatrix} = \overbrace{\begin{bmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{bmatrix}}^{\mathbf{T}_{qd}} \begin{bmatrix} u_q \\ u_d \end{bmatrix}, \quad (24)$$

where u is a system quantity, e.g., voltage or current, and φ is the angle difference between the two qd frames which is assumed to remain constant through the system dynamics [5]. It should be noted that transforming between qd frames using the \mathbf{T}_{qd} matrix leads to an alteration of transfer matrices of (21) as

$$\begin{bmatrix} Y'_{qq}(s) & Y'_{qd}(s) \\ Y'_{dq}(s) & Y'_{dd}(s) \end{bmatrix} = \mathbf{T}_{qd} \begin{bmatrix} Y_{qq}(s) & Y_{qd}(s) \\ Y_{dq}(s) & Y_{dd}(s) \end{bmatrix} \mathbf{T}_{qd}^{-1}, \quad (25)$$

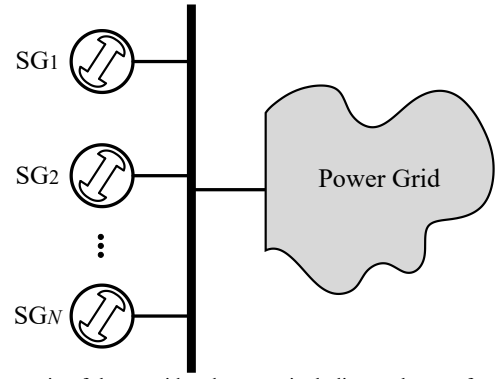


Fig. 1. Schematic of the considered system including a cluster of parallel SGs connected to a power grid.

$$\begin{bmatrix} G'_{qfd}(s) \\ G'_{dfd}(s) \end{bmatrix} = \mathbf{T}_{qd} \begin{bmatrix} G_{qfd}(s) \\ G_{dfd}(s) \end{bmatrix}. \quad (26)$$

Without loss of generality, it is assumed that SG1 remains in its qd reference frame (i.e., RRF1), and the qd equations of all other SGs (i.e., SG n with $n \in \{2, \dots, N\}$) are transformed from their qd reference frame (i.e., RRF n) to RRF1. Therefore, the transfer matrices of the first and n -th SGs can be expressed as

$$\begin{bmatrix} i_{qs1}(s) \\ i_{ds1}(s) \end{bmatrix} = \begin{bmatrix} Y_{qq1}(s) & Y_{qd1}(s) \\ Y_{dq1}(s) & Y_{dd1}(s) \end{bmatrix} \begin{bmatrix} v_{qs1}(s) \\ v_{ds1}(s) \end{bmatrix} - \begin{bmatrix} G_{qfd1}(s) \\ G_{dfd1}(s) \end{bmatrix} v_{fd1}, \quad (27)$$

$$\begin{bmatrix} i'_{qsn}(s) \\ i'_{dsn}(s) \end{bmatrix} = \begin{bmatrix} Y'_{qqn}(s) & Y'_{qdn}(s) \\ Y'_{dqn}(s) & Y'_{ddn}(s) \end{bmatrix} \begin{bmatrix} v'_{qsn}(s) \\ v'_{dsn}(s) \end{bmatrix} - \begin{bmatrix} G'_{qfdn}(s) \\ G'_{dfd n}(s) \end{bmatrix} v_{fdn}. \quad (28)$$

where $v_{qs1} = v_{qsn} = v_{qs}$ and $v_{ds1} = v_{dsn} = v_{ds}$ due to their parallel connection.

Finally, the transfer matrices of all the SGs, which are referred to as RRF1 can be aggregated by adding their stator currents [e.g., adding (27) to (28)] and obtaining the combined currents as

$$\begin{bmatrix} i_{qs}(s) \\ i_{ds}(s) \end{bmatrix} = \begin{bmatrix} i_{qs1}(s) \\ i_{ds1}(s) \end{bmatrix} + \sum_{n=2}^N \begin{bmatrix} i'_{qsn}(s) \\ i'_{dsn}(s) \end{bmatrix} \\ = \begin{bmatrix} Y'_{qq}(s) & Y'_{qd}(s) \\ Y'_{dq}(s) & Y'_{dd}(s) \end{bmatrix} \begin{bmatrix} v_{qs}(s) \\ v_{ds}(s) \end{bmatrix} - \begin{bmatrix} G'_{qfd1}(s) & \dots & G'_{qfdN}(s) \\ G'_{dfd1}(s) & \dots & G'_{dfdN}(s) \end{bmatrix} \begin{bmatrix} v_{fd1} \\ \vdots \\ v_{fdN} \end{bmatrix}, \quad (29)$$

where

$$\begin{bmatrix} Y'_{qq}(s) & Y'_{qd}(s) \\ Y'_{dq}(s) & Y'_{dd}(s) \end{bmatrix} = \begin{bmatrix} Y_{qq1}(s) & Y_{qd1}(s) \\ Y_{dq1}(s) & Y_{dd1}(s) \end{bmatrix} + \sum_{n=2}^N \begin{bmatrix} Y'_{qqn}(s) & Y'_{qdn}(s) \\ Y'_{dq n}(s) & Y'_{ddn}(s) \end{bmatrix}, \quad (30)$$

$$\begin{bmatrix} G'_{qfd1}(s) & \dots & G'_{qfdN}(s) \\ G'_{dfd1}(s) & \dots & G'_{dfdN}(s) \end{bmatrix} = \begin{bmatrix} G_{qfd1}(s) & \dots & G_{qfdN}(s) \\ G_{dfd1}(s) & \dots & G_{dfdN}(s) \end{bmatrix}. \quad (31)$$

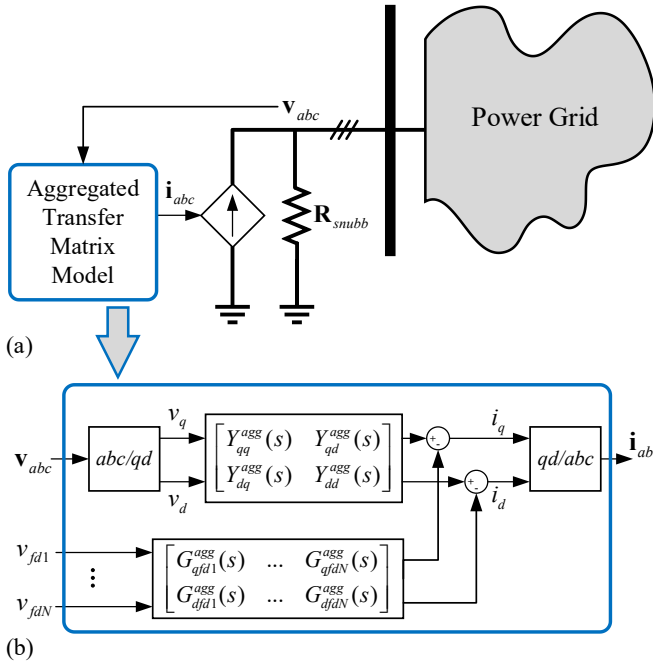


Fig. 2. Implementation of the proposed aggregated transfer matrix model of the parallel SGs: (a) interfacing with external network using voltage-controlled current sources, (b) implementation of the transfer matrices.

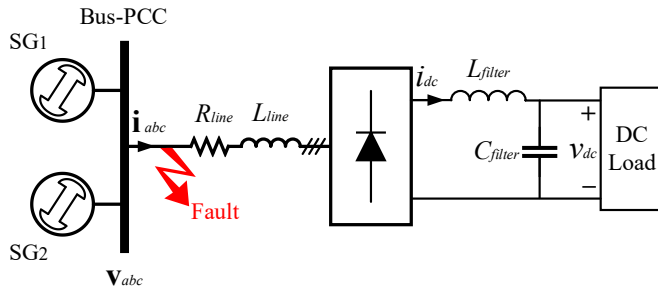


Fig. 3. Schematic of the case-study system consisting of two parallel SGs supplying a dc system through a rectifier.

Based on (29), the cluster of SGs can be modeled as an aggregated transfer matrix, as illustrated in Fig. 2(a). As seen, the inputs of the aggregated transfer matrix model of the SGs are the terminal and field voltages and the outputs are the injected stator currents of SGs. Therefore, similar to the qd models of SGs, a three-phase dependent current source (with possibly a three-phase snubber resistance (R_{snubb}) in case of connection to an inductive or a power-electronic-based network [16]) can be used for circuit implementation of the proposed aggregated transfer matrix model of the SGs.

III. COMPUTER STUDIES

Here, the accuracy and computational performance of the proposed aggregated transfer matrix model of the SGs, formulated in Section II, are investigated for a representative case study shown in Fig. 3. Therein, two SGs are connected in parallel and supply a dc load through a six-pulse diode rectifier. The SGs are connected to the rectifier through a transmission line (represented by R_{line} , L_{line}). A low-pass filter (represented by L_{filter} , C_{filter}) is also used to smoothen the dc voltage supplied to the dc load. Without loss of generality, it is also assumed that

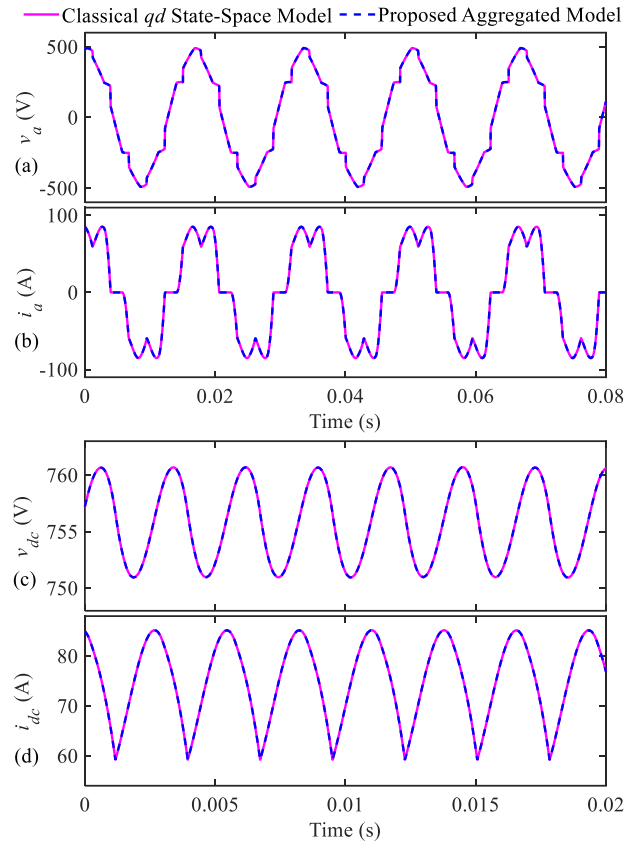


Fig. 4. Waveforms of the system variables in steady-state obtained by the classical qd state-space model and the proposed aggregated transfer matrix-based model for: (a) phase a voltage at Bus-PCC, (b) total injected phase a current to Bus-PCC by the two SGs, (c) dc load voltage, and (d) rectifier dc current.

the dc load can be represented by a resistor consuming the required power. The parameters of the case-study system are summarized in Appendix B.

To verify the proposed transfer matrix-based modeling technique, two different approaches are considered for modeling the parallel SGs: First, as the reference model, the two SGs are implemented individually using their classical qd state-space models in MATLAB/Simulink. Secondly, the proposed aggregated model of the two SGs has been implemented using the transfer matrices following the procedure explained in Section II. Both the reference and the proposed aggregated models are interfaced with the external network using 100 pu (with base $S_b = 105$ MVA and $V_b = 560$ V) artificial snubbers connected across the terminals of the SGs [see Fig. 2(a)].

A. Steady-State Operation with High-Frequency Harmonics

Here, it is assumed that the field voltages (E_{xfdn}) of both SGs [i.e., $E_{xfdn} = (X_{mnd}/r_{fdn}) v_{fdn}$ with $n \in \{1, 2\}$], are held constant at 2.30 pu. As a result, at the rated line-to-line voltage (i.e., 560 V), SG1 delivers 52 kW with rotor angle $\delta_1 = 32.94^\circ$ and SG2 delivers 22 kW with rotor angle $\delta_2 = 26.02^\circ$. This corresponds to 74 kW (neglecting the losses in the transmission line) consumed by the dc network, which is modeled as a resistive load of 10 ohms. The system variables obtained by the subject models in steady-state are shown in Fig. 4. As it can be observed in Figs. 4(a) and 4(b), the proposed aggregated model provides

consistent results with the qd state-space model for the ac-bus (Bus-PCC) voltage and the combined current of the two SGs, including their harmonics. Also, as seen in Fig. 4(c) and 4(d), the proposed aggregated model provides very accurate results for the dc-side variables, compared to the reference model, including the switching ripples.

B. Transients Following a Short-Circuit Fault

Here, it is assumed that the system initially operates in a steady-state condition as specified in Section III-A. Then, at $t = 1$ s, a three-phase short-circuit fault occurs on the transmission line and is very close to Bus-PCC (as indicated in Fig. 3). The fault is removed at $t = 1.1$ s. The system variables obtained by the subject models for the duration of the transient are depicted in Fig. 5. As it can be observed in Figs. 5(a) and 5(b), the voltage at the ac-bus becomes zero after the short-circuit fault, resulting in a very high current due to the high voltage at the terminals of the SGs (the field voltages were assumed to be constant). It can also be seen in Fig. 5(d) that the dc current becomes zero following the fault. This is due to the fact that the diodes in the rectifier become reverse biased after the fault. Consequently, the dc voltage starts to decrease (with the time constant of the dc subsystem). It is also seen in Figs. 5(a)–(d) that the system starts to recover after removing the fault.

In the meanwhile, it is verified in Figs. 5(a)–(d) that the proposed aggregated model is very accurate (compared to the qd state-space model as the reference) in predicting the system variables even in such large-signal transients.

C. Predicting the Operational Impedances

The so-called operational impedances [3, Ch. 7] can be analyzed to verify the accuracy achieved by the proposed aggregated model in steady-state and transient studies in Sections III-A and B. For this purpose, the q - and d -axes stator flux linkages per second can be written in the frequency domain as [3, Ch. 7]

$$\psi_{qs}(s) = X_{qs}(s)i_{qs}(s), \quad (32)$$

$$\psi_{ds}(s) = X_{ds}(s)i_{ds}(s) + K_{fd}(s)v_{fd}. \quad (33)$$

Here, $X_{qs}(s)$ and $X_{ds}(s)$ are the operational impedances [3, Ch. 7], and $K_{fd}(s)$ is a transfer function that expresses the relationship between the d -axis stator flux linkage per second and the field voltage. Comparing (5), (6) with (32), (33), and using (11), (12), the operational impedances for an SG can be expressed as

$$X_{qs}(s) = \omega_b \left(L_q - \frac{sL_{mq}^2}{r_{kq} + sL_{kq}} \right), \quad (34)$$

$$X_{ds}(s) = \omega_b \left(L_d - \frac{sL_{md}^2}{r_{fd} + sL_{fd}} \right), \quad (35)$$

and $K_{fd}(s)$ is defined as

$$K_{fd}(s) = \frac{L_{md}}{r_{fd} + sL_{fd}}. \quad (36)$$

The operational impedances can also be calculated using the proposed transfer matrix-based model of the individual SGs based on the Y_{qd} transfer matrix in (21). This is achieved by

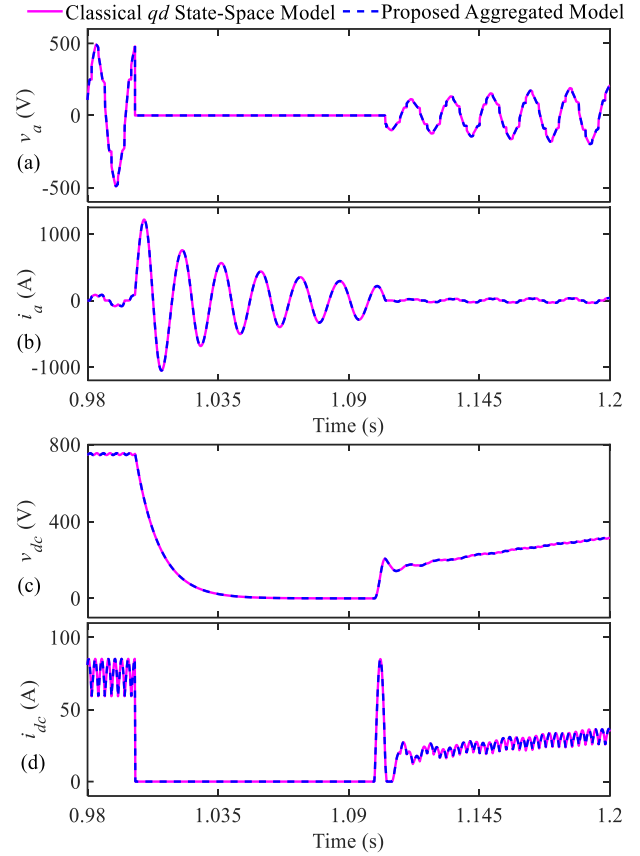


Fig. 5. Waveforms of the system variables when a short circuit occurs at $t=1$ s (and cleared at $t=1.1$ s) as obtained by the classical qd state-space model and the proposed aggregated transfer matrix-based model for: (a) phase a voltage at Bus-PCC, (b) total injected phase a current to Bus-PCC by the two SGs, (c) dc load voltage, and (d) rectifier dc current.

setting r_s and ω_r equal to zero [3, Ch. 7] for computing the impedances as

$$X_{qs}(s) = \frac{\omega_b}{s} \left. \frac{v_{qs}(s)}{i_{qs}(s)} \right|_{r_s=0, \omega_r=0} = \frac{\omega_b}{s} \left. \frac{1}{Y_{qq}(s)} \right|_{r_s=0, \omega_r=0}, \quad (37)$$

$$X_{ds}(s) = \frac{\omega_b}{s} \left. \frac{v_{ds}(s)}{i_{ds}(s)} \right|_{r_s=0, \omega_r=0} = \frac{\omega_b}{s} \left. \frac{1}{Y_{dd}(s)} \right|_{r_s=0, \omega_r=0}. \quad (38)$$

Similarly, aggregated operational impedances can be calculated for the cluster of SGs using the proposed aggregated model based on the transfer matrix in (29) and following the procedure in (37), (38).

The bode plots for the operational impedances of the two parallel SGs in Fig. 3 are obtained using their proposed aggregated model [based on (37), (38)] and compared with their exact (reference) values [i.e., $X_{qs}(s) = X_{qs1}(s) \parallel X_{qs2}(s)$ for the q -axis and $X_{ds}(s) = X_{ds1}(s) \parallel X_{ds2}(s)$ for the d -axis, based on (34), (35)] in Figs. 6 and 7. As observed, the proposed aggregated model can be effectively utilized for predicting the magnitude and phase of the d - and q -axes operational impedances of the parallel SGs with very high accuracy over a wide range of frequency spectrum. This also verifies and explains the high accuracy observed in Figs. 4 and 5 for the steady-state and transient response of the proposed model.

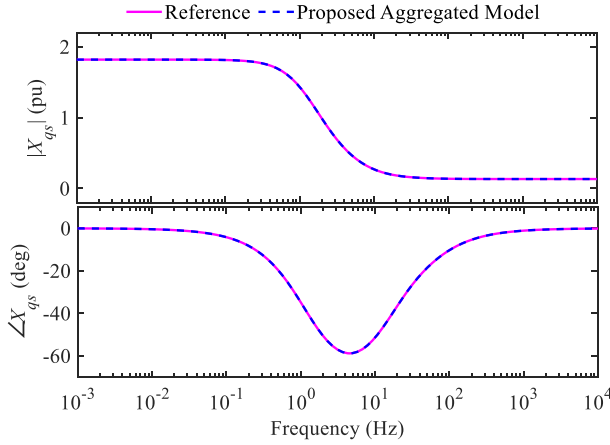


Fig. 6. Magnitude and phase of the q -axis operational impedance of the two parallel SGs over a wide range of frequency as predicted by the proposed aggregated model.

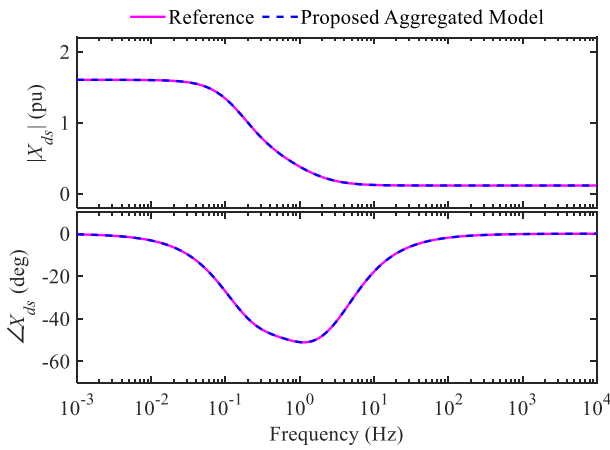


Fig. 7. Magnitude and phase of the d -axis operational impedance of the two parallel SGs over a wide range of frequency as predicted by the proposed aggregated model.

D. Comparison of Computational Performance

Finally, the computational performance of the proposed aggregated model of the parallel SGs is benchmarked against their counterpart qd state-space model. For consistency, both models are simulated in MATLAB/Simulink [17] using the stiff solver *ode23tb* as well as the non-stiff solver *ode45* with maximum and minimum time-step of 0.1 ms and 0.01 μ s, respectively. The absolute and relative errors are chosen to be 10^{-4} . The simulations are executed on a personal computer with Intel® Core™ i7-10750H CPU @ 2.60 GHz processor.

Here, the same short-circuit fault transient study explained in Section III-B is conducted on the case-study system of Fig. 3 and simulations are continued for 5 seconds. The computational performance of the two subject models is tabulated in Table I and Table II with the stiff solver *ode23tb* and non-stiff solver *ode45*, respectively.

As can be seen in Tables I and II, the two models require approximately the same number of time steps. Therefore, one can deduce that the two models have essentially similar eigenvalue structures. However, with the *ode23tb* solver, the proposed aggregated model shows a marginal improvement of 14.7% in the CPU time compared to the qd state-space model (i.e., 25.37 s vs. 29.77 s), as seen in Table I. Meanwhile, when

both models are run with the *ode45* solver, the proposed aggregated model outperforms the qd state-space model by decreasing the CPU time by 39.9% (i.e., 300.25 s vs. 499.85 s), as seen in Table II. This verifies that simulating the system using the proposed aggregated transfer matrix-based model is computationally more efficient (i.e., less complex) for the solver than computing the system variables using multiple qd state-space models for SGs.

TABLE I. COMPUTATIONAL PERFORMANCE OF THE SUBJECT MODELS FOR THE CONSIDERED 5-SECOND TRANSIENT STUDY WITH THE STIFF SOLVER *ode23tb*

Model	Number of time steps	CPU time
Classical qd State-Space Model	97,520	29.77 s
Proposed Aggregated Model	99,465	25.37 s

TABLE II. COMPUTATIONAL PERFORMANCE OF THE SUBJECT MODELS FOR THE CONSIDERED 5-SECOND TRANSIENT STUDY WITH THE NON-STIFF SOLVER *ode45*

Model	Number of time steps	CPU time
Classical qd State-Space Model	2,500,643	499.85 s
Proposed Aggregated Model	2,500,501	300.25 s

IV. CONCLUSION

This paper presents an aggregation technique for equivalent dynamic modeling of a cluster of synchronous generators (SGs). The proposed method exploits the so-called transfer matrices in the Laplace domain, where the equivalent transfer-based models are aggregated. It has been shown for a cluster of wound-field synchronous generators (SGs) with damper windings that the proposed aggregated model can follow their classical full-order qd models very accurately for a wide range of electromagnetic transients, e.g., for high-order harmonics and/or short-circuit dynamics. Furthermore, it was verified that the proposed aggregated model of the SGs is computationally more efficient than their classical full-order qd state-space models.

APPENDIX A

The magnetizing and subtransient reactances of an SG with a field and one damper winding on its d -axis as well as two damper windings on its q -axis are expressed as [3, Ch. 7]

$$X_{mq} = X_q - X_{ls}, \quad (\text{A.1})$$

$$X_{md} = X_d - X_{ls}, \quad (\text{A.2})$$

$$X_q'' = X_{ls} + \frac{1}{\frac{1}{X_{mq}} + \frac{1}{X_{lkq1}} + \frac{1}{X_{lkq2}}}, \quad (\text{A.3})$$

$$X_d'' = X_{ls} + \frac{1}{\frac{1}{X_{md}} + \frac{1}{X_{lfd}} + \frac{1}{X_{lkd}}}. \quad (\text{A.4})$$

Also, the SG's open-circuit time constants can be expressed as [3, Ch. 7]

$$\tau_{qo}'' = \frac{1}{\omega_b r_{kq2}} \left(X_{lkq2} + \frac{X_{mq} X_{lkq1}}{X_{mq} + X_{lkq1}} \right), \quad (\text{A.5})$$

$$\tau'_{do} = \frac{1}{\omega_b r'_{fd}} (X_{jfd} + X_{md}). \quad (A.6)$$

To calculate the parameters of an equivalent SG with only the field winding on the d -axis and only one damper winding on the q -axis, it is desired that the subtransient reactances of the equivalent machine remain unchanged (to keep the short-circuit subtransient peak current magnitude of the equivalent machine in the same level) [3, Ch. 7]. Therefore, using the magnetizing and subtransient reactances in (A.1)–(A.4), the leakage reactances of the equivalent field winding and one damper winding for the equivalent SG can be obtained as

$$X_{jfd}^{equ} = \frac{1}{\frac{1}{X_d'' - X_{ls}} - \frac{1}{X_{md}}}, \quad (A.7)$$

$$X_{lkq}^{equ} = \frac{1}{\frac{1}{X_q'' - X_{ls}} - \frac{1}{X_{mq}}}. \quad (A.8)$$

Then, to maintain the same open-circuit time-constants as the original SG, the transient d -axis open-circuit time-constant (τ'_{do}) is used as the field winding time-constant (τ_{fd}), and the subtransient q -axis open-circuit time-constant (τ''_{qo}) as the q -axis damper winding time-constant (τ_{kq}) for the calculation of machine winding parameters [18]. As a result, the field and the one damper winding resistances are calculated based on (A.5), (A.6) as

$$r'_{fd} = \frac{X_{jfd}^{equ} + X_{md}}{\omega_b \tau'_{do}}, \quad (A.9)$$

$$r_{kq} = \frac{X_{lkq}^{equ} + X_{mq}}{\omega_b \tau''_{qo}}. \quad (A.10)$$

In this paper, the superscript “ equ ” denoting the parameters of the equivalent SG has been removed from notations for simplicity.

APPENDIX B

Parameters of the Case-Study System in Fig. 3:

TABLE III. PARAMETERS OF SG 1 [19]

Parameter	Value	Parameter	Value
Rating	105 kVA	Voltage	560 V
r_s	0.137 Ω	X_{ls}	0.338 Ω
X_d	16.624 Ω	X_q	8.180 Ω
r_{fd}	0.0266 Ω	X_{jfd}	1.270 Ω
r_{kd}	0.120 Ω	X_{lkd}	0.062 Ω
r_{kq}	0.120 Ω	X_{lkq}	0.131 Ω

TABLE IV. PARAMETERS OF SG 2 [20]

Parameter	Value	Parameter	Value
Rating	59 kVA	Voltage	560 V
r_s	0.09 Ω	X_{ls}	1.276 Ω
X_d	29.507 Ω	X_q	17.226 Ω
r_{fd}	0.0195 Ω	X_{jfd}	2.1852 Ω
r_{kd}	0.42 Ω	X_{lkd}	5.33 Ω
r_{kq1}	25.26 Ω	X_{lkq1}	2.3287 Ω
r_{kq2}	1.07 Ω	X_{lkq2}	5.2635 Ω

TABLE V. PARAMETERS OF THE AC TRANSMISSION LINE AND DC FILTER

Transmission Line			
R_{line}	0.001 Ω	L_{line}	0.1 mH
Filter			
L_{filter}	1 mH	C_{filter}	1 mF

REFERENCES

- [1] M. G. Taul, X. Wang, P. Davari, and F. Blaabjerg, "Reduced-order and aggregated modeling of large-signal synchronization stability for multiconverter systems," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 3, p. 3150–3165, Jun. 2021.
- [2] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbić, "Foundations and challenges of low-inertia systems," in *Proc. Power Syst. Computation Conf. (PSCC)*, 2018, pp. 1–25.
- [3] P. C. Krause, O. Wasynczuk, S. D. Sudhoff, and S. D. Pekarek, *Analysis of electric machinery and drive systems*. 3rd Edition, John Wiley & Sons, 2013.
- [4] M. Shirinzad, "Frequency scan based stability analysis of power electronic systems," M.Sc. Thesis, The University of Manitoba, 2021.
- [5] Y. Gu, Y. Li, Y. Zhu, and T. C. Green, "Impedance-based whole-system modeling for a composite grid via embedding of frame dynamics," *IEEE Trans. Power Syst.*, vol. 36, no. 1, pp. 336–345, Jan. 2021.
- [6] M. Amin and M. Molinas, "Small-signal stability assessment of power electronics based power systems: A discussion of impedance-and eigenvalue-based methods," *IEEE Trans. Ind. Appl.*, vol. 53, no. 5, pp. 5014–5030, Oct. 2017.
- [7] S.-F. Chou, X. Wang, and F. Blaabjerg, "Two-port network modeling and stability analysis of grid-connected current-controlled VSCs," *IEEE Trans. Power Electron.*, vol. 35, no. 4, pp. 3519–3529, Apr. 2020.
- [8] S.-F. Chou, X. Wang, and F. Blaabjerg, "An aggregated model for power electronic system based on multi-port network reduction method," in *Proc. IEEE 21st Workshop on Control & Modeling for Power Electron. (COMPEL)*, 2020, pp. 1–6.
- [9] Y. Gu, N. Bottrell, and T. C. Green, "Reduced-order models for representing converters in power system studies," *IEEE Trans. Power Electron.*, vol. 33, no. 4, pp. 3644–3654, Apr. 2018.
- [10] D.-E. Kim and M. A. El-Sharkawi, "Dynamic equivalent model of wind power plant using an aggregation technique," *IEEE Trans. Energy Convers.*, vol. 30, no. 4, pp. 1639–1649, Dec. 2015.
- [11] M. Mercado-Vargas, D. Gomez-Lorente, O. Rabaza, and E. Alameda-Hernandez, "Aggregated models of permanent magnet synchronous generators wind farms," *Renew. Energy*, vol. 83, pp. 1287–1298, Nov. 2015.
- [12] V. Akhmatov and H. Knudsen, "An aggregate model of a grid-connected, large-scale, offshore wind farm for power stability investigations—importance of windmill mechanical system," *Int. Journal Elect. Power Energy Syst.*, vol. 24, no. 9, pp. 709–717, Nov. 2002.
- [13] M. Pöller and S. Achilles, "Aggregated wind park models for analyzing power system dynamics," in *Proc. 4th International Workshop on Large-Scale Integration of Wind Power and Transmission Networks for Offshore Wind Farms*, 2003, pp. 1–10.
- [14] L. Fernandez, C. Garcia, J. Saenz, and F. Jurado, "Equivalent models of wind farms by using aggregated wind turbines and equivalent winds," *Energy Convers. and Manage.*, vol. 50, no. 3, pp. 691–704, Mar. 2009.
- [15] J. H. Chow, *Power system coherency and model reduction*. Springer, 2013.
- [16] L. Wang *et al.*, "Methods of interfacing rotating machine models in transient simulation programs," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 891–903, Apr. 2010.
- [17] *Simulink Dynamic System Simulation Software, User's Manual*, MathWorks Inc., Natick, MA, USA, 2022. [Online]. Available: <http://www.mathworks.com>.
- [18] E. Johansson, "Detailed description of synchronous machine models used in Simpow," *Technical Report TR H 01-160*, 2001.
- [19] I. Jadric, D. Borojevic, and M. Jadric, "Modeling and control of a synchronous generator with an active DC load," *IEEE Trans. Power Electron.*, vol. 15, no. 2, pp. 303–311, Mar. 2000.
- [20] A. Davoudi, H. Behjati, and J. Jatskevich, "Simulation-based dynamic characterization of transformer-isolated machine-rectifier systems," in *Proc. IEEE Transport. Electrification. Conf. and Expo (ITEC)*, 2012, pp. 1–7.

Systematic Analysis and Proposed AI-based Technique for Attenuating Inductive and Capacitive Parasitics in Low and Very Low Frequency Antennas

Kate G. Francisco¹, R-Jay S. Relano¹, Mike Louie C. Enriquez¹, Ronnie S. Concepcion II¹, Jonah Jahara G. Baun², Adrian Genevie G. Janairo², Ryan Rhay P. Vicerra¹, Argel A. Bandala², Elmer P. Dadios¹, Jonathan R. Dungca³

¹Department of Manufacturing Engineering and Management, De La Salle University, Manila, Philippines

²Department of Electronics and Computer Engineering, De La Salle University, Manila, Philippines

³Department of Civil Engineering, De La Salle University, Manila, Philippines

{kate_g_francisco, r-jay_relano, mike.enriquez, ronnie.concepcion, jonah_baun, adrian_janairo, ryan.vicerra, argel.bandala, elmer.dadios, jonathan.dungca}@dlsu.edu.ph

Abstract—Non-destructive mapping of underground utilities is one of the fundamental concepts of subsurface imaging technology that has a great contribution to the improvement of many infrastructure concerns. It is incorporated with electrical resistivity measurement of various ground conditions through electrodes with functional geometric configuration. Due to the presence of an electrical field, electromagnetic noise and interference will likely occur and might cause inaccuracy of data. On that note, it is vital to understand the different factors affecting the system and its impact as the initial step in the development of an effective filtering and shielding mechanism. Thus, this paper discusses the possible impacts of parasitic inductance and capacitance affecting the performance of low and very low-frequency antennas, and the collection of various optimization methods as well as the tools and software used in the mitigation of parasitic elements in an electronics system found in different research publications and journals. Furthermore, an AI-based framework was also provided as an initial step in the development of a parasitic antenna filter that performs well for underground imaging single antenna array. Genetic algorithm is the AI technique proposed for the optimization of the antenna filter by providing the best combination of material by considering its conductivity and thickness.

Keywords—*Electromagnetic Interference, Genetic Algorithm, Low Frequency, Mitigation, Stray Capacitance, Stray Inductance, Very Low Frequency*

I. INTRODUCTION

Subsurface imaging technology has been known for its functional detection of utilities underground [1]. In the Philippines, it has a huge potential in solving infrastructure problems through the non-destructive mapping of underground utilities as there are more investments in infrastructure projects since 2017 [2]. In putting up an underground imaging system, there are transmitter and receiver circuits as well as capacitive electrodes that have functional geometrical arrangements that are capacitively coupled to the ground. They are used in the recording of electrical resistivity measurement which depends on varying ground conditions. Thus, it is inevitable to encounter challenges such as close levels of electrical and electromagnetic noise and interference due to the presence of the electrical field [3].

It is eminent that electromagnetic interference (EMI) brings disruption to the electrical circuit because of the presence of electromagnetic radiation coming from an outer source [3]. Relatedly, the noise all around the EMI frequency spectrum is caused by impedances and current/voltage sources [4]. The increasing use of electromagnetic transmission systems for information technology applications in the surroundings caused the electromagnetic spectrum to be congested that resulting in EMI which affects the normal function of a certain system [3]. In addition, this disturbance was caused by electromagnetic energy called environmental interference (EI) which is categorized according to its sources such as environmental, incidental, or intentional [3]. Also, the spectrum is crowded in low-frequency bands [5]. Hence, an effective EMI shielding mechanism is vital for the protection of a sensitive frequency-operated system such as in the mapping of underground utilities.

In power electronics, electrical circuits usually contain parasitic elements due to the coupling of electromagnetic and electrostatic between different components of the circuit [6]. Truly, it is difficult to obtain a unified and accurate model to characterize noise emissions due to varying speed, configuration, and information of a particular system in power electronics [4]. With that, this research addresses the conventional sources of electromagnetic interference and the impacts of parasitic inductance and parasitic capacitance affecting the performance of very low and low frequency antennas. It aims to discuss how parasitics may possibly influence and contribute to the disturbance of systems. Furthermore, this study contributes to the: (1) discussion and emphasizing of the tools and software used by different studies in this field and the optimization methods done to establish an effective mitigation mechanism for parasitic elements in an electronics system, and (2) formulation of an innovation framework for the use of AI-based techniques in mitigating parasitic elements that could be used for further development of a shield or filter mechanism for VLF-LF antenna operation for subsurface imaging.

II. ACTIVE PARASITIC ELEMENTS IN ANTENNA

Self and mutual parasitics are the two types of parasitic parameters in a particular electronic circuit. The self-parasitics basically covers the parasitic components of inductors while the

mutual parasitics usually occur between the layout of printed circuit board (PCB) and its traces [7]. Additionally, pertaining in the analysis of electromagnetic compatibility, there are factors to contemplate, such as distinguishing the EMI sources, locating the crucial signal paths and coupling loops, and applying EMI mitigation techniques, especially in designing EMI filters [8].

Undeniably, parasitic elements are present in any frequency spectrum, and it just varies on how it affects the circuit system. It is interesting to unveil how parasitic elements deal with low and very low-frequency ranges. It is known that underground and underwater are two difficult environments to deal with when it comes to radio frequency communications since the incorporation of high-frequency electromagnetic waves is rejected in concentrated media [9]. Hence, stray inductance and stray capacitance in antennas with low and very low frequencies were attempted to tackle. This will help in establishing an EMI filter that could potentially be used in an underground imaging application operating in a very low and low-frequency range.

III. STRAY INDUCTANCE AND CAPACITANCE IN ANTENNA OPERATIONS

Stray inductance can be defined as unavoidable inductance present within a circuit which disrupts the normal flow of the current. The basic concept of stray inductance is when wires or any other electronic components conceived magnetic fields around them producing an inductive effect that must be avoided. The goal of minimizing the parasitic inductance is a great challenge to engineers and circuit designers as it contributes to difficulties in the field of power electronics. Some problems related to it are the rise of unwanted overshoots and ringing after switching transients [10]. Likewise, stray capacitance is incorporated into unnecessary capacitance in an electronic system. Capacitance can be produced if there are any two elements at different electric potentials that are near to each other. Those elements could generate an electric field and become like capacitors [11]. It must be considered also that any wire used in the circuit can contain limited capacitance concerning the ground. Likewise, a sensor that acts as a capacitor could also contribute to the additional capacitances in the circuit due to its obtained wires [12]. Changes in capacitance such as capacitance fluctuations are considered one source of error due to the said circumstances [12]. Hence, one important way of minimizing the stray capacitance is through the separation of sensor elements and the rest of the circuit [12].

A. Impacts of Stray Inductance and Capacitance

The inductance is not of great concern in low-frequency circuits as well as with the lower HF band circuits [13]. Strays become crucial when passing frequencies from upper HF to the VHF region. At those frequency ranges, the stray inductance becomes a significant component of the total circuit inductance [13]. In regards to the VLF range, the authors have difficulty gathering research papers about the effects of stray inductance and capacitance in VLF. To the best of the researchers' knowledge, most of the studies found that are related to the VLF antennas are very outdated [14-16]. Thus, it can be concluded that most of the authors studying the field of parasitics are not

exploring too much VLF compared to a high-frequency range which has a huge impact on this matter.

On the other hand, stray capacitance has a notable impact on the common-mode noise in conductive electromagnetic interference (EMI) [17]. However, stray capacitance should not give special attention to dc and low ac frequencies [13]. It is mentioned in [18] that in dealing with high-frequency, the effects of such unwanted capacitance are amplified. Capacitive effects are more dynamic at higher voltages in low frequency than inductive effects [18]. However, it should also be noted that as the frequency increases, those parasitic elements obtained a massive proportion of their totality [13].

B. Factors Affecting Stray Inductance and Capacitance

Proper circuit layout is an important factor in RF circuits as it can reduce the effects of parasitic elements [13]. Stray inductance is mainly incorporated into metal connecting wire where the distance, diameter, type of material, shape, and linking methods are factors to be considered for a change in electrical characteristics of a certain power module [19]. For stray capacitance, it is also noteworthy to consider the use of broad printed circuit tracks for connection rather than wires [13]. Air as an excellent dielectric is a most renowned contributor to stray capacitance [20]. Evidently, one or more dielectrics can be incorporated into stray capacitances. Hence, its magnitude will vary depending on the location of the main capacitor [20]. Resistors also exhibit a few amounts of parasitic capacitance due to their connecting leads and certain parts that have the capability to store charge [18].

IV. MITIGATION STRATEGIES FOR PARASITIC ELEMENTS

There are stray couplings between components that are more challenging to classify, so the development of a mitigation model that incorporates the identification of magnetic and electric fields around certain parts components were employed to control its parasitic elements [21]. Efficient EMI mitigation strategies have gained increasing interest as EMI designing and simulation enable the estimation of the emf performance of a certain circuit system before prototyping [22]. It is known that the sole objective of noise filters is to reduce the unnecessary electrical waves to the electronically powered equipment and propagates via alternating current lines and to the nearby devices [22]. Thus, EMI filters must be properly designed according to the need and complexity of an electronic system [22].

A. Tools and Software

In developing an EMI filter, one important thing to consider is the tools and software used in measuring current flows in components and other parasitic elements. For an antenna model estimation, a study used an LCR meter that functions in frequencies between 12 Hz and 187 Hz to measure the electrical circuit of a loop antenna [23]. On the other hand, an oscilloscope and a function generator were established to measure lower frequencies that are < 12 Hz [23]. In [24], ANSYS HFSS 3D finite element electromagnetic solver was used for measuring current scattering within two AI blocks that resemble the inner capacitors' form (Fig. 1). It was used to confirm the suggested concept of the proximity effect, while the electrical flow in the anti-parallel arrangement will approximate each other, thus,

based on the simulation the expected phenomenon happened [24].

In terms of measuring the stray inductance and capacitance of an Integrated Power Electronics Module (IPEM), Maxwell Q3D Extractor was used in [25] and Agilent 4294A precision impedance analyzer for verification of the simulation results. Using these tools, impedance measurement confirmed the loop inductances when the switching to on and off is employed as well as the capacitances between three terminals and the ground. Hence, this was done to measure the parasitic inductance. Relatedly, another tool for impedance measurement was mentioned in [22]. The OMICRON Bode100 vector network analyzer was used to verify measurements, thus, upon comparing the simulations and measurements, it conceived a good match that operated in a frequency of 10 Hz to 40 MHz [22].

One tool for measuring insertion gains and network criteria in an EMI filter structure is the Agilent four-port balanced network analyzer E5070B which was utilized in [11]. It is known that the effects of stray capacitance can be illustrated through numerical simulations using finite element analysis (FEA) [18]. Thus, to execute the modeling, a Comsol Multiphysics® software package was utilized in a study to develop FEA models used to calculate the charge distribution of conductors with an operating frequency of 50 Hz [18]. Apparently, voltage noise measurement is important to consider in exploring the effects of parasitic elements as it also leads to the mitigation of those strays. In [26], the Hewlett-Packard 35665A dynamic signal analyzer was utilized in calculating all the voltage noise incorporated in the study that is open from the test circuits output. The modulation frequency used in the experiment is 100 kHz [26]. Overall, these tools and software mentioned in different studies were able to establish an important initiative to mitigate the known parasitic elements and analysis of various electronic systems. Hence, the summary of these tools and software with a measured frequency range between <12 Hz and 110 MHz were presented in Table 1.

TABLE I. SUMMARY OF THE TOOLS AND SOFTWARE USED IN MITIGATING PARASITIC ELEMENTS.

Tools / Software	Function	Measured Frequency Range
LCR meter [23]	measure a loop antenna electrical circuit	12 Hz -187 Hz
Oscilloscope and a function generator set-up [23]	measure a loop antenna electrical circuit	< 12 Hz
ANSYS HFSS 3D finite element electromagnetic solver [24]	compute the current distribution	50 MHz
Maxwell Q3D Extractor [25]	calculate the parasitic elements of an Integrated Power Electronics Modules (IPEM)	-
Agilent 4294A precision impedance analyzer [25]	measure IPEM for verification of Maxwell Q3D simulation results	40 Hz – 110 MHz
OMICRON Bode100 vector network analyzer [22]	verify impedance measurements of the L-shaped conductors of Partial element equivalent circuit (PEEC) model	10 Hz – 40 MHz
Agilent four-port balanced network analyzer E5070B [11]	measure all insertion gains and network criterions	-
Comsol Multiphysics® software package [18]	develop FEA models used to calculate the charge distribution of conductors	50 Hz
Hewlett-Packard 35665A dynamic signal analyzer [26]	calculate the voltage noise coming from the circuits' output	100 kHz

B. Optimization Methods

Over the years, there are various techniques and optimization methods that were developed by researchers for the attenuation of noise and performance improvement for EMI filters. One simple technique used in [24] is placing two or more capacitors in two positions which are parallel and anti-parallel arrangements (Fig. 2). Using these, it significantly reduced their equivalent parasitic inductance, which increases the level of reduction of a filter at the HF range [24]. The proposed anti-parallel arrangements of capacitors can also enhance the behavior of EMI filters in HF range [24].

Cancellation techniques were utilized in various research papers, one is the ESL cancellation technique which basically uses an X-capacitor-inductor structure [8]. In Fig. 3(a) and 3(c), the ESL cancellation concept was presented where two capacitors are connected diagonally. It can be seen that two inductors, L, are attached on the upper and lower sides. Thus, if the total value of inductance is the same as the ESL (Fig. 3(b) and 3(d)), then the reaction of ESL on the diverted path are canceled as well as the cancellation resistance and ESR [8]. However, ESR and ESL cannot be completely canceled, but they can be successfully lessened and improve the capabilities of the capacitors [8].

In mitigating parasitic capacitance, a great work of estimating it applicable to CM noise in actual time was employed using the Unscented Kalman Filter (UKF) [17]. The UKF was used as a state simulator and the complement circuit of CM noise. Together, they were able to determine the effective parasitic capacitance. Thus, the proposed method used a typical

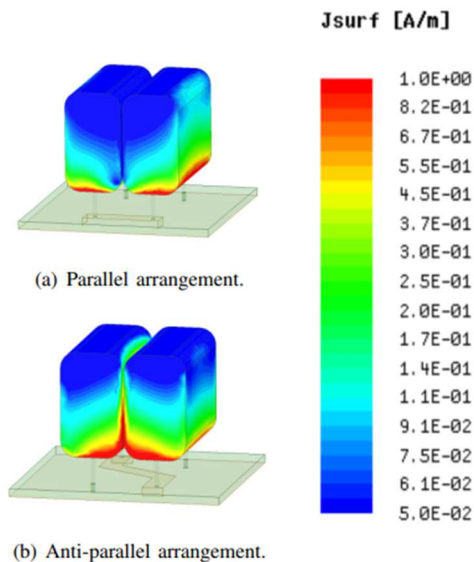


Fig. 1. The current surface density of two aluminum blocks was calculated with ANSYS HFSS at 50 MHz [24].

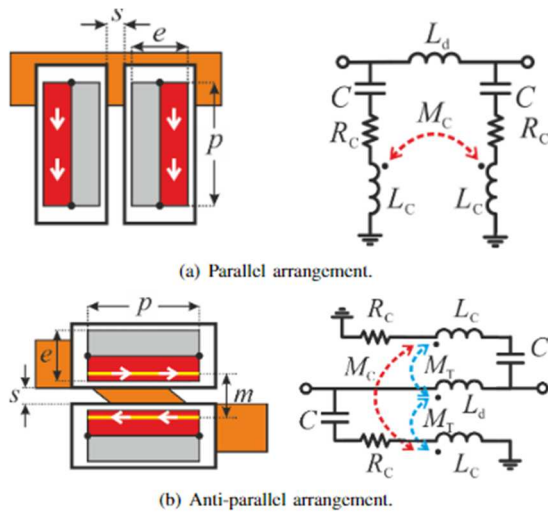


Fig. 2. Two smart arrangements applied in two capacitors that serve as diverted filters together with their incorporated circuit representation: (a) parallel and (b) anti-parallel [24].

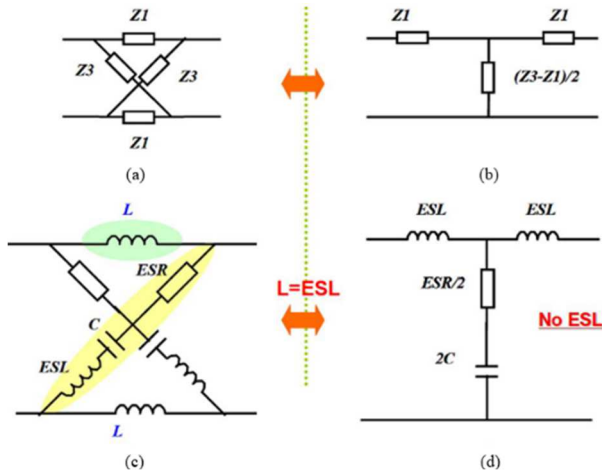


Fig. 3. ESL cancellation for capacitors: (a) two capacitors were connected diagonally, (b) resultant circuit of (a), (c) two small inductors L , are attached to the upper and lower sides, and (d) resultant circuit of (c) showing that inductance is equal to equivalent series inductance (ESL) showing no effect [10].

DC-DC buck converter in various working set-ups for testing [17].

Another study was conducted [18] where the effects of dispersed parasitic capacitance have been examined by conducting assessments through 3-D Finite Element Analysis (FEA). This allows the calculation of the parasitic capacitance and the alteration in the divider voltage output [18]. Additionally, the use of grading electrodes to alleviate the results of stray capacitance was utilized. It found that only limited correction and enhancement in the function of the resistive divider was gained when the toroidal grading rings were employed while grading hollow cylinders that function as a Faraday cage for the electrical lines, showing an excellent improvement in the response of the divider [18].

Another application that accumulates stray capacitance is in a cryogenic experimental set-up where it uses long coaxial

cables for the minimization of heat leak [26]. Unfortunately, these cables have a huge amount of capacitance per unit and a lock-in amplifier can be utilized to get rid of the outcome of the demodulator at around 100 kHz [26]. The effectivity of the coax cable and the vessel in terms of disallowing EMI should be considered in noise performance. Thus, an optimized trans-impedance capacitive sensor circuit was utilized that obtained functional noise operation and an allowable rejection of RFI [26]. This circuit introduced a capacitor that is in series with the stray capacitance and this greatly resulted in reduced noise gain in comparison to other test circuits employed [26].

A summary of the various optimization methods was presented in Table 2. However, the frequency range utilized in those methods is mostly for high-frequency operations as research papers incorporated from LF to VLF range is very limited to none. Thus, it can be concluded that parasitics in high frequency really gained a lot more interest and exploration compared to low frequencies.

TABLE II. SUMMARY OF THE OPTIMIZATION METHODS FOR MITIGATION OF PARASITIC ELEMENTS

Method / Technique	Function	Frequency Range
Parallel and antiparallel arrangement of thin-film capacitors [24]	Significantly decrease the equivalent inductance	high frequency
X-capacitor-inductor structure [8]	ESL cancellation technique for HF noise suppression	high frequency
UKF (Unscented Kalman Filter) algorithm [17]	Estimates parasitic capacitance in real-time	very high-frequency processing unit (250 MHz)
Three-dimensional FEA simulations [18]	Allows calculation of both the parasitic capacitance and output voltage	high frequency
Grading hollow cylinder [18]	Provides excellent performance in the reduction of stray capacitance in a resistive divider.	high frequency
Proposed noise gain modification method [26]	Reduces the noise gain from the shield's low impedance	100 kHz

V. PROPOSED AI-FRAMEWORK IN MINIMIZING THE BUILD-UP OF ANTENNA PARASITICS FOR SUBSURFACE IMAGING

Among the various mitigation techniques from different literature, another possible application is underground imaging antenna operations. It is known that geophysical imaging technique such as electrical resistivity tomography (ERT) has been of great use in the identification of underground utilities [27]. On that note, it can be functional in an infrastructure problem such as in the instances of 'hits' or phenomena that unintentionally damage utility lines or even worksites due to a lack of information on existing utilities and entities located underground.

In creating a subsurface imaging system, it is also important to consider that it requires high sensitivity in the recording of electrical resistivity during land surveying. Thus, the initial development of a parasitic antenna filter will be of great help in

securing the subsurface imaging antenna array against unnecessary noise or parasitic elements. Moreover, the authors have proposed a framework using the AI technique that can be utilized as an initial step in the development of a parasitic antenna filter for an underground imaging array (Fig. 4). The framework is divided into two phases, the first one is the design and modeling of the underground imaging antenna array and the parasitic antenna filter up to their data acquisition process and the second is the application of AI technique for optimization. Thus, after considering the best material selection and parameters for the parasitic antenna filter, simulations will take place and the gathering of data. Then, it will undergo the second phase which starts with data analysis. This is where AI techniques such as genetic algorithms and evolutionary computing can be applied. Genetic algorithm (GA) is incorporated into an optimization algorithm that is influenced by natural selection [28]. Various research papers used GA as an optimization technique in designing and modeling shields related to electromagnetic waves [28-35]. Hence, it is evident that this AI technique can be functional in the development and optimization of a parasitic antenna filter.

To assess the proposed AI based-framework, initial simulations of single pair (transmitter and receiver electrodes) antenna array for an underground imaging system showing with and without the application of parasitic antenna filter were conducted to prove its functionality. In Fig. 5(a), a single pair antenna including the transmitter and receiver circuits and their electrodes was simulated using Altair Feko software operating in 1 MHz frequency. The pattern of radiations in Fig. 5(a) are all coming from the transmitter and flow continues to the receiver as expected since there is no added filter yet in the system. Since the circuits and electrodes are near field locations, the generation of electromagnetic waves can be observed which

employs the presence of parasitic elements. The reading of the receiver below the ground could be affected by the generated electric field above that could result in inaccurate data in surveying.

On the other hand, another simulation was conducted, this time with a parasitic element filter as presented in Fig. 5(b). The parasitic antenna filter used in this simulation is in the form of box enclosures. The electrodes and circuits are secured with enclosures to filter unnecessary electromagnetic waves in the system. It can be seen that the radiations from the transmitter did not continuously flow in the receiver circuit as compared with the scenario in Fig. 5(a). The box enclosure in the receiver circuit was able to absorb and reflect the radiation and function well as a shield. To verify the result of the shielding mechanism, a near field graph that corresponds to the electric field around the enclosure can be observed (Fig. 6). The blue line in the graph represents the circuit without a filter which shows a continuous line with increasing value of the electric field. This indicates that the receiver circuit can also acquire direct signals which could affect the reading of data. However, in the green line that is incorporated into the circuit with filter, it displays its peak in the -0.73 m position and decreased intensely from -0.7 m to -0.61 m, then afterward started to increase again. It can be perceived that the location where the line was in the lowest position is the area where the box enclosure is located, meaning it successfully blocks and absorb the electromagnetic waves coming to the receiver circuit.

Furthermore, AI techniques such as the genetic algorithm using MATLAB can be applied for the optimization of the best combination of thickness and material of the parasitic antenna filter that is ideal for the underground imaging antenna array. This can help in identifying the best alloy mixture of materials considering their conductivity and thickness. However, in the simulations conducted in Altair Feko, a graphite material was used as a metal shield since it is the best option for a filter enclosure due to its dense quality and conductivity, but it is rarely used because it's expensive [36].

One of the limitations that can be perceived in using GA is the premature convergence due to the early completion of exploitation. This is incorporated into the loss of population diversity [37]. With that, other bio-inspired optimization algorithms should be explored using the same dataset. Conversely, the main advantage of the proposed technique is providing an accurate and optimal combination of a material that can be functional for the development of the parasitic antenna filter. This can be applied in land surveying on roads or even in uneven terrain through an underground imaging system apparently being towed with a small vehicle for economic improvement [38]. The filter will be a great help in securing the accuracy of data acquired by blocking unnecessary signals or EM waves as the imaging operation requires high sensitivity to soil resistivity readings. The success of employing this parasitic filter can guarantee a good result in generating maps of utilities underground.

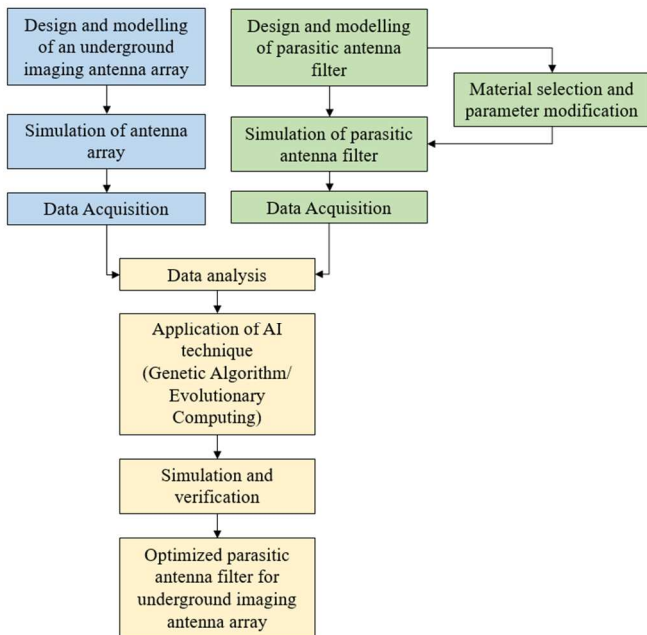


Fig. 4. Proposed framework in using AI technique in the initial development of parasitic antenna filter for the underground imaging antenna array.

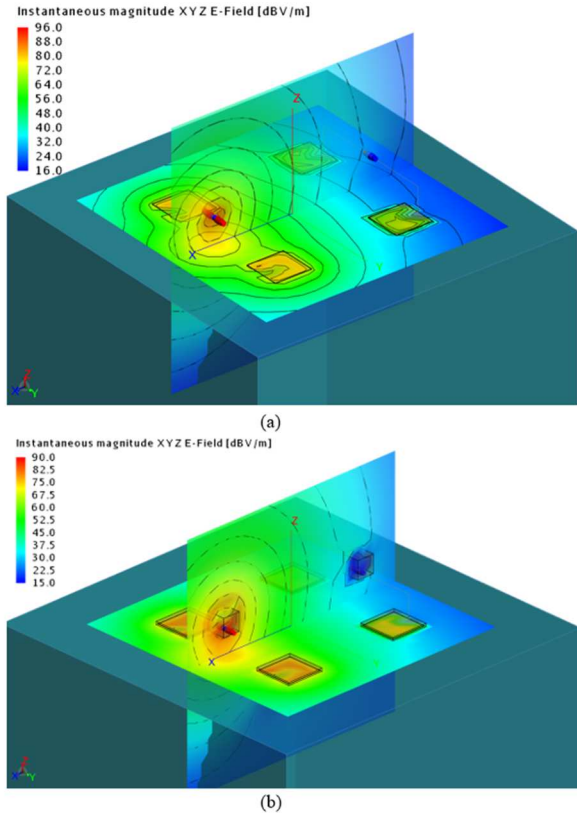


Fig. 5. Single pair antenna array simulation using Altair Feko software (a) without and (b) with parasitic antenna filter.

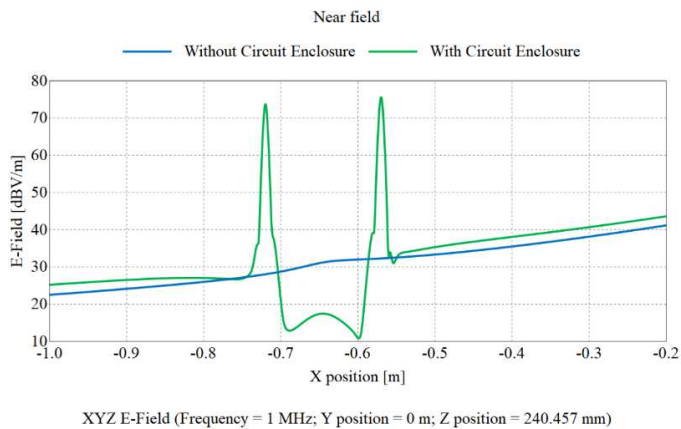


Fig. 6. Near-field comparison of with and without parasitic antenna filter.

CONCLUSION

This paper gives rise to the initial establishment of an effective filtering and shielding mechanism for an underground imaging system as it focused on the systematic analysis of parasitics in the low and very low-frequency range. We also proposed a framework for using the AI technique to reduce antenna parasitic in a subsurface imaging application. However, based on gathered literature, it found that in low-frequency circuits, the inductance is not of great concern in low frequency as strays become crucial when passing frequencies from upper

HF to the VHF region. At those frequency ranges, the stray inductance becomes a significant component of the total circuit inductance. On the other hand, the effects of such unwanted capacitance are amplified with HF compared to the LF. Moreover, as the frequency increases, those parasitic elements obtained a massive proportion of their totality. In terms of very low-frequency range, most of the studies found that are related to the VLF antennas are very outdated which can be perceived that researchers are not exploring too much in VLF especially in parasitics compared to high-frequency range that has a huge impact on this matter. Furthermore, the proposed AI-based framework was employed, and a simulation of a single pair antenna array was conducted to assess it. Genetic Algorithm as an AI technique can be utilized to provide the best combination of alloy material that could ideally perform as a parasitic antenna filter. Mitigation techniques and optimization methods from other research papers as well as the functional tools and software were also presented to give insights into the current strategies for the attenuation of EMI noise.

ACKNOWLEDGMENT

The researchers would like to express gratitude to the DOST-PCIEERD and Intelligent Systems Laboratory of De La Salle University-Manila for their support and motivation to complete this research work and pursue the technological advancement of a subsurface imaging system in the Philippines.

REFERENCES

- [1] K. Francisco et al., "Analytical Hierarchical Process-based Material Selection for Trailer Body Frame of an Underground Imaging System," in 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Nov. 2021, pp. 1–6. doi: 10.1109/HNICEM54116.2021.9732048.
- [2] K. Francisco et al., "Systematic Analysis of the Implementation of Sustainable Development Goals on Energy, Industrialization, Infrastructure, and Innovation: A Multifaceted Philippines," in 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Nov. 2021, pp. 1–6. doi: 10.1109/HNICEM54116.2021.9732043.
- [3] M. Kaur, S. Kakar and D. Mandal, "Electromagnetic interference," *2011 3rd International Conference on Electronics Computer Technology*, 2011, pp. 1-5, doi: 10.1109/ICECTECH.2011.5941844.
- [4] A. C. Baisden, D. Boroyevich, and F. Wang, "Generalized terminal modeling of electromagnetic interference," *IEEE Transactions on Industry Applications*, vol. 46, no. 5, pp. 2068–2079, Sep. 2010, doi: 10.1109/TIA.2010.2058836.
- [5] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, "Multiantenna Techniques," in *5G Physical Layer*, Elsevier, 2018, pp. 199–252. doi: 10.1016/b978-0-12-814578-4.00012-6.
- [6] G. Engelmann, S. Quabeck, J. Gottschlich and R. W. De Doncker, "Experimental and simulative investigations on stray capacitances and stray inductances of power modules," *2017 19th European Conference on Power Electronics and Applications (EPE'17 ECCE Europe)*, 2017, pp. P.1-P.10, doi: 10.23919/EPE17ECCEEurope.2017.8099158.
- [7] S. Wang, F. C. Lee, W. G. Odendaal and J. D. van Wyk, "Improvement of EMI Filter Performance with Parasitic Coupling Cancellation," *2005 IEEE 36th Power Electronics Specialists Conference*, 2005, pp. 1780-1786, doi: 10.1109/PESC.2005.1581872.
- [8] S. Wang, F. C. Lee, and W. G. Odendaal, "Cancellation of capacitor parasitic parameters for noise reduction application," *IEEE Trans. PowerElectron.*, vol. 21, no. 4, pp. 1125–1132, Jul. 2006.
- [9] P. Lunkenheimer, S. Emmert, R. Gulich, M. Köhler, M. Wolf, M. Schwab, and A. Loidl, "Electromagnetic-radiation absorption by water." *Physical Rev. E*, vol. 96, no. 6, Dec. 2017, Art. no. 062607. doi: 10.1103/PhysRevE.96.062607

- [10] H. C. P. Dymond and B. H. Stark, "Investigation of a parasitic-inductance reduction technique for through-hole packaged power devices," *IEEE Energy Conversion Congress and Exposition (ECCE)*, 2018, doi: 10.1109/ECCE.2018.8558152
- [11] "What is Stray Capacitance?" <http://www.learningaboutelectronics.com/Articles/What-is-stray-capacitance> (accessed Mar. 17, 2022).
- [12] T. Kenny and W. Kester, "Sensor Fundamentals," *Sensor Technology Handbook*, pp. 1–20, Jan. 2005, doi: 10.1016/B978-075067729-5/50041-0.
- [13] J.J. Carr, "Introduction to RF electronics," 2004. [Online]. Available: www.digitalengineeringlibrary.com
- [14] S.C. Tietsworth, "A new system for measurement of Low Frequency Radio Transmitting Antenna Parameters in Real Time," Naval Ocean Systems Center, San Diego, June 1991.
- [15] H. A. Wheeler, "Fundamental relations in the design of a VLF transmitting antenna," *IRE Transactions on Antennas and Propagation*, vol. 6, Issue 1, pp. 120-122, Jan. 1958.
- [16] D. A. Gurnett, & B.J. O'Brien, "High-latitude geophysical studies with satellite Injun 3: 5. Very-low-frequency electromagnetic radiation" *Journal of Geophysical Research*, 69(1), 65–89., Jan. 1964. doi:10.1029/jz069i001p00065
- [17] S. Karimi, E. Farjah, T. Ghanbari, F. Naseri, and J. L. Schanen, "Estimation of Parasitic Capacitance of Common Mode Noise in Vehicular Applications: An Unscented Kalman Filter-Based Approach," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 8, pp. 7526–7534, Aug. 2021, doi: 10.1109/TIE.2020.3007088.
- [18] J.R. Riba, F. Capelli, and M. Moreno-Eguilaz, "Analysis and Mitigation of Stray Capacitance Effects in Resistive High-Voltage Dividers" *Energies*, 12(12), 2278, June 2019, doi:10.3390/en12122278
- [19] M. Yang, W.Cai, M. Zhou, and Q. Guo, "Low Stray inductance packing technology," 4th International Symposium on Power Electronics and Control Engineering, ISPECE 2021, vol. 12080, no. 1208011, Sept. 2021.
- [20] R. Arora, and W. Mosch, "High Voltage and Electrical Insulation Engineering," John Wiley & Sons: Hoboken, NJ, USA, 2011.
- [21] L. Taylor, W. Tan, and X. Margueron, "Reducing of parasitic inductive couplings effects in EMI filters." 15th European Conference on Power Electronics and Applications (EPE), Sept. 2013, doi: 10.1109/EPE.2013.6634643
- [22] I. F. Kovacevic, T. Friedli, A. M. Musing, and J. W. Kolar, "3-D electromagnetic modeling of parasitics and mutual coupling in EMI filters," *IEEE Transactions on Power Electronics*, vol. 29, no. 1, pp. 135–149, 2014, doi: 10.1109/TPEL.2013.2254130.
- [23] L. M. Carducci, R. Alfonso, and W. G. Fano, "ELF magnetic field receiver: frequency performance and natural signals detection," *Elektron*, vol. 5, no. 2, pp. 105–111, Dec. 2021, doi: 10.37537/rev.elektron.5.2.135.2021.
- [24] P. Gonzalez-Vizuete, F. Fico, A. Fernandez-Prieto, M. J. Freire, and J. B. Mendez, "Calculation of Parasitic Self- and Mutual-Inductances of Thin-Film Capacitors for Power Line Filters," *IEEE Transactions on Power Electronics*, vol. 34, no. 1, pp. 236–246, 2018, doi: 10.1109/TPEL.2018.2824658.
- [25] J. Z. Chen, L. Yang, D. Boroyevich, and W. Gerhardus Odendaal, "Modeling and Measurements of Parasitic Parameters for Integrated Power Electronics Modules," 2004.
- [26] C. Gettings and C. C. Speake, "A method for reducing the adverse effects of stray-capacitance on capacitive sensor circuits," *Review of Scientific Instruments*, vol. 90, no. 2, Feb. 2019, doi: 10.1063/1.5080016.
- [27] M. L. Enriquez *et al.*, "Prediction of Weld Current Using Deep Transfer Image Networks Based on Weld Signatures for Quality Control," in *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Nov. 2021, pp. 1–6. doi: 10.1109/HNICEM54116.2021.9731979.
- [28] R. Concepcion, E. Dadios, and J. Cuello, "Non-destructive in situ measurement of aquaponic lettuce leaf photosynthetic pigments and nutrient concentration using hybrid genetic programming," *Agrivita Journal of Agricultural Science*, vol. 43, no. 3, pp. 597-617, 2021. doi: 10.17503/agrivita.v43i3.2961.
- [29] P. L. Sergeant, L. R. Dupré, M. de Wulf, and J. A. A. Melkebeek, "Optimizing Active and Passive Magnetic Shields in Induction Heating by a Genetic Algorithm," *IEEE Transactions on Magnetics*, vol. 39, no. 6, pp. 3486–3496, Nov. 2003, doi: 10.1109/TMAG.2003.819460.
- [30] M. Ziolkowski and S. R. Gratkowski, "Genetic algorithm and bezier curves-based shape optimization of conducting shields for low-frequency magnetic fields," *IEEE Transactions on Magnetics*, vol. 44, no. 6, pp. 1086–1089, Jun. 2008, doi: 10.1109/TMAG.2007.915994.
- [31] S. Cui, D. S. Weile, and J. L. Volakis, "Novel planar absorber designs using genetic algorithms," in *IEEE Antennas and Propagation Society, AP-S International Symposium (Digest)*, 2005, vol. 2 B, pp. 271–274. doi: 10.1109/APS.2005.1551993.
- [32] A. J. Lozano-Guerrero, A. Diaz-Morcillo, and J. V. Balbastre-Tejedor, "Resonance suppression in enclosures with a metallic-lossy dielectric layer by means of genetic algorithms," 2007. doi: 10.1109/ISEMC.2007.215.
- [33] R. Concepcion, L. Ilagan, and I. Valenzuela, "Optimization of nonlinear temperature gradient on eigenfrequency using metaheuristic genetic algorithm for reinforced concrete bridge structural health," World Congress on Engineering and Technology; Innovation and its Sustainability 2018, pp. 141-151, 2018. doi: 10.1007/978-3-030-20904-9_11.
- [34] P. Hong Thinh, N. Thi Lan Huong, H. Ngoc Nhan, W. J-I, and H. Quoc Viet-Hanoi-Vietnam, "Conception and Realization of Multilayered Composite Electromagnetic Shielding Material at Microwave Frequency by Using Genetic Algorithm."
- [35] R. C. Souza, G. Fontgalland, M. A. de Barbosa Melo, R. C. S. Freire, and R. B. Vasconcelos, "Proposal of a multilayer shield design using genetic algorithm," in *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, 2005, vol. 3, pp. 2300–2305. doi: 10.1109/imtc.2005.1604587.
- [36] S. Geetha, K. K. S. Kumar, C. R. K. Rao, M. Vijayan, and D. C. Trivedi, "EMI shielding: Methods and materials - A review," *Journal of Applied Polymer Science*, vol. 112, no. 4, pp. 2073–2086, May 2009, doi: 10.1002/app.29812.
- [37] A. Hussain and Y. S. Muhammad, "Trade-off between exploration and exploitation with genetic algorithm using a novel selection operator," *Complex & Intelligent Systems*, vol. 6, no. 1, pp. 1–14, Apr. 2020, doi: 10.1007/s40747-019-0102-7.
- [38] R. Concepcion *et al.*, "The technology adoption and governance of artificial intelligence in the Philippines," in *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Nov. 2021, pp. 1–6. doi: 10.1109/HNICEM48295.2019.9072725.

IoT Enabled Smart Solar Panel Monitoring System Based on Boltuino Platform

Ashim Mondal
 Dept. of Electrical Engineering
 Aliah University
 Kolkata, India
 ashim.au.een@gmail.com

Md Jishan Ali
 Dept. of Electrical Engineering
 Aliah University
 Kolkata, India
 jishan.een.au@gmail.com

Pallav Dutta
 Dept. of Electrical Engineering
 Aliah University
 Kolkata, India
 pallav.dutta@aol.com

Abstract: Renewable energy sources have been proven to be reliable, and they are also thought to be the best way to meet our growing energy needs. The rising interest in solar power, increasing costs and functional expenses are led to a few convincing explanations behind the requirement for energy monitoring. Today, Solar Photovoltaic (PV) power is the newest and brightest advancement in the renewable energy market that reduces the requirement of fossil fuel by-products. To get the most out of solar power generation, it's important to pay a lot of attention to how it's kept and how it's used. The purpose of this project work is to display real-time PV parameters on mobile or computer remotely, manage the data and load which is connected with the PV panel and examine the statistics via graph. In this work 20-watt solar panel is used as a PV module, Boltuino (Bolt IoT+ Arduino) device is used as a controller circuit and Current sensor, Voltage sensor, Temperature & Humidity sensor are used for solar parameter collection. The Arduino Uno collects the current sensor, voltage sensor, temperature & humidity sensor data and sends it to the BOLT via serial communication. Following receipt of the statistics in BOLT, it displays all of the solar parameters and generates a graph, which stores all of the information. The alert message will be sent to the registered mobile phone through SMS, and be able to monitor the data from the PV module and remotely operate the system in the event of any unwanted situation like a short circuit or voltage cut out.

Keywords – Renewable energy, solar photovoltaic, cloud computing, internet of things, remote monitoring

I. INTRODUCTION

In today's world, innovation has effectively taken the place of civilisation as the supercurrent. Numerous advancements are taking place in the electrical and hardware domains, which are resulting in more recent and significant revelations and discoveries. The primary difficulty facing all developing countries is the production of large amounts of electrical energy. Since the rise of industry and business, power request arrives at their pick. Henceforth all are aiming towards environment-friendly power sources to meet the energy needs in a sustainable manner. Solar power station

production requires astute observation and activity in moving toward breakthroughs in low-energy engineering that is oriented toward zero-energy engineering. [1].

Internet of Things (IoT) is a fast-developing invention [2], that when connected through the correspondence convention and cloud stage, makes things more intelligent and easier to utilise. In order to determine the PV module's efficiency, parameters such as current, voltage, power, and temperature must be considered [2] [3]. Consequently, a real-time photovoltaic module monitoring system is essential for increasing the reliability of solar modules by comparing the results of the test to begin a preventative activity [4].

Deficient PV modules, dust on solar panels that lowers the efficiency, connection of the system [5], and other factors affecting plant execution are monitored by this system.

Using the Internet of Things [6], a remote real-time monitoring system for solar power plants allows for the monitoring of solar parameters from anywhere and at any time. It controls the system for execution at the solar power plant level and promotes the decisional interaction for the main control station.

The Internet of Things (IoT) allows objects to be controlled and detected remotely [7] [8], allowing for easier integration of the real world into a programmable system, reducing human intervention and increasing efficiency, precision, and economy. Real-time monitoring [9], control, and maintenance of a solar power plant can be made possible via IoT technology.

For monitoring all PV module data, an internet-based programmable graphical user interface is designed [10]. In order to improve the data characteristic of the solar module, Arduino is utilised as a controller to display and analyse collected data. It will be a future accountable technique for large PV power plants using the IoT-enabled smart solar panel monitoring system using the Boltuino platform [11].

II. LITERATURE SURVEY

When it comes to increasing the quality of life for individuals, the Internet of Things (IoT) has emerged as a reliable partner in recent years.. The Internet of Things (IoT) is a network of physical devices that connect to the internet. Due to the fact that it allows sensing and specialised devices to interact with one another in order to meet the specific requirements of individuals. It bridges the gap between the advanced virtual and physical world. The Internet of Things (IoT) is replacing human interventions with machine-to-machine communication to screen and manage the borders of the PV board.

In the year of 2021 N. A. b. Anang Othman et. al [12] designed an IoT-based Solar Battery Monitoring System utilizing two microcontrollers, Node-MCU and Arduino Uno. The information acquired will be put away in the nearby data set and can be seen over an individual website that fills in as an information log and through a representation apparatus utilizing Grafana. All through the framework, the sun-powered PV framework can be effortlessly followed by the client utilizing the internet.

Kanaga Durga D. et. al [13] in 2020 revealed a paper that proposes an installed framework to support the remote observing of inverters in a photo-Voltic power generating station. The information is a complete power created and usual parts of the framework execution must be gotten physically in the Solar power generating station. This proposed framework involves Raspberry pi as the regulator circuit. The inverters are associated with a Daisy chain style and the last inverter is associated with the Raspberry Pi. The absolute power and five safeguard boundaries grid fault, high-current, high-temperature, and short-circuit are gotten from the inverters utilizing the Modbus convention. Sensors are utilized to gauge Wind speed, Temperature, and Solar Irradiance.

R. I. S. Pereira et.al [14] in 2019 published a paper detailing about plan, improvement, execution, and approval of an Internet of Things (IoT) circuit with sensors signal melding network for observing decentralized photo-voltaic plants. The framework expects to diminish expenses of business information lumberjacks and detecting devices which necessary to control information stockpiling utilizing restrictive programming.

An IoT-based far-off ongoing energy observing network is created to display the solar power generation in 2019 by Mohd Sajid Khan et. al [15]. Several voltage and current sensors are connected with a versatile micro-controller for amassing the data. An Internet of Things research stage is adjusted to envision the amassed data and assess the energy generate for a given dispersed age structure. This aids in nonstop remote

monitoring and brings about superior productivity of the power plants.

In the year 2018 Prakhar Srivastava et. al [16] published a paper where they focused on controlling crossbreed energy frameworks utilizing IoT. The primary rules are exchanging within the two sources of energy which are solar and wind-based power with no burden by a site using the ESP8266 Wi-Fi module. The data is transferred remotely by the site to the ESP8266 Wi-Fi module that controls the wellsprings of energy. The sent data is controlled remotely using the Internet of things (IoT). This allows clients to have versatile control tools via a secured internet connection. The system assists the energy client in remotely and physically monitoring the sources parameters via an advanced mobile or computer. This system is somewhat effective, cheap, and versatile in action.

III. OPERATION, BLOCK DIAGRAM & CIRCUIT CONNECTION

This system works with the following connection displayed in Fig 2. The BOLT module and Arduino Uno get power from an external source using a USB cable. Interface the pins Rx and Tx of the BOLT module to connect the 8 and 9 digital I/O pins of the Arduino Uno. The PV module is connected with the demo load through the relay and current sensor, the single-channel relay is used as a system relay, a voltage sensor is connected with the PV module in parallel, also two loads are connected with a dual-channel relay which is used as a load relay. The current sensor, Voltage sensor, OLED display, and DHT11 (Temperature & Humidity sensor) +ve and -ve terminal which is VCC and GND connect with Arduino UNO 5V and GND pin. The data pin of the current and voltage sensor is connected with Arduino UNO A0, and A1 analog pin respectively. Pin A0 and pin 3 of the BOLT module are connected with the voltage data pin and digital pin 10 of Arduino Uno. The SCL and SDA of OLED are connected with Arduino Uno A5 and A4 analog pins. DHT11 (Temperature and humidity sensor) is connected with Arduino Uno of digital pin 2.

All relay is connected with BOLT GPIO pin system relay is connected with BOLT pin 0 and load 1 and load 2 relay data pin is connected with BOLT pin 1 and 2. The VCC and GND pin is connected with Arduino Uno 5V and GND pin.

In this proposed system relay connect with the PV module positive terminal as a switch. In the BOLT cloud system, ON and OFF switches are present which is for turning on the system relay.

At the first system, the relay gets turned on through press the button in the BOLT cloud platform and the system gets started. The prototype of the system is shown in Fig. 3.

BOLT request data from Arduino Uno and it collect data from all the sensor. Collect data from the current sensor, voltage sensor, and DHT11 (temperature & Humidity sensor). Process the data and send it to the BOLT device through serial communication. BOLT shows the data in the BOLT cloud platform. If any data get high or low, if the voltage and power of the solar panel are low and the current are high, so we can manage the load to turn OFF through the load relay which is controlled by the Lad ON and Load OFF button of the BOLT cloud plate from. If the temperature rises and touches the threshold value, press the button system OFF. So, all system is turned off.

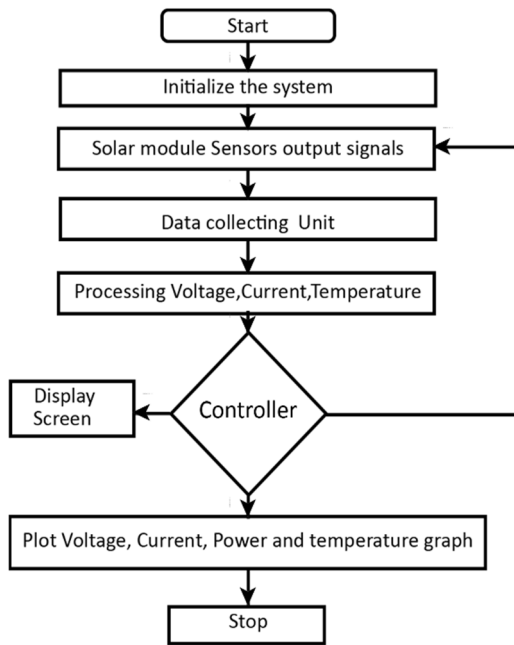


Fig.1: Flowchart of Operation

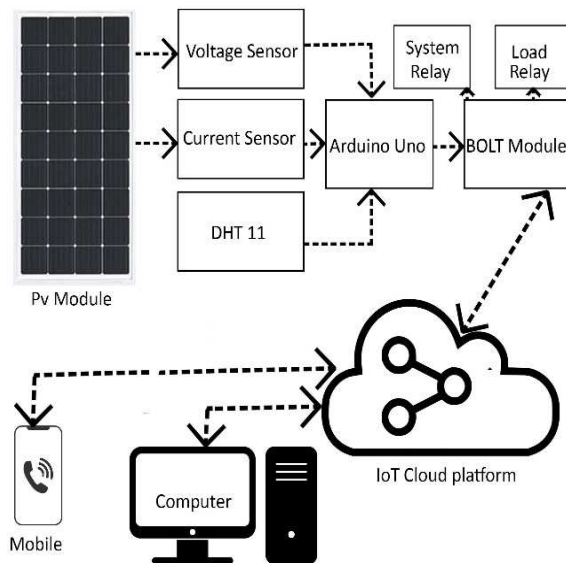


Fig 2: Block diagram of the proposed system



Fig 3: Prototype of the proposed system

All the data show in the OLED display near the system which is shown in the IoT cloud platform. If unwanted things happened like a line cut out, a short circuit, voltage drop near to zero, or current fallen down to zero the alert message is sent to the mobile.

IV. HARDWARE ARCHITECTURE

A. BOLT IoT Device

BOLT IoT module is a Wi-Fi cloud computing-based module [17]. This device has Wi-Fi connectivity based on the ESP8266 module. It collects data from any sensor or device and sends it to the cloud also shown in graphs various from [11]. With BOLT Cloud you have some control over and screen them over the web, make customized dashboards to picture the information, screen the gadget's wellbeing, run AI calculations, and part more. It has one analog pin, 5 digital input/output pins, one 3.3 Volt power pin, 5 Volt power pin, one GND pin, and also RX and TX pin for serial communication from another device. ML (Machine Learning) algorithms are simply consolidated with BOLT IoT projects to work for recognizing anomalies in the sensor information.

B. Arduino Uno

The Arduino Uno is an open-source microcontroller board in base on the microchip ATmega328P microcontroller [18] and created by Arduino. cc. The board is equipped with sets of advanced and simple input/output (I/O) pins that might be communicated to various extension sheets (safeguards) and different circuits. The board has 14 input/output (I/O) pins (six fit for PWM), 6 simple I/O pins, and Arduino IDE (Integrated Development Environment) is used to program for different purposes and projects, by a USB cable. The power source of the Arduino is through a USB cable or by an outside 9-volt DC battery, but it acknowledges voltages somewhere that vary between seven and twenty volts. Arduino Uno is used in this project to collect the data from all the sensor, display the data on the OLED screen, and when the data requested Arduino sends it to the BOLT.

C. Current Sensor (ACS712ELCTR-30A-T)

The ACS712ELCTR-30A-T current sensor is a totally integrated, linear current sensor IC in 8 pins SOIC package which has a hall effect [19]. The sensor consists of an accurate, low-offset, copper conduction path with a linear hall IC circuit which is located near the surface of the die. It generates a magnetic field for flowing through the applied current to the copper conduction path which is sensed by the integrated Hall

sensor and converted into a proportional voltage. It has a low-noise analog signal path. Its reaction time is 5 μ s to step input current and 1.5% at TA = 25 $^{\circ}$ C complete result blunder. It has 1.2 ohms inside guide opposition. 66mV/A result responsiveness. Incredibly steady result offset voltage. Almost Zero Magnetic Hysteresis. This current sensor is used in this system to measure the current flow from the solar panel to the load is loaded condition also no-load condition.

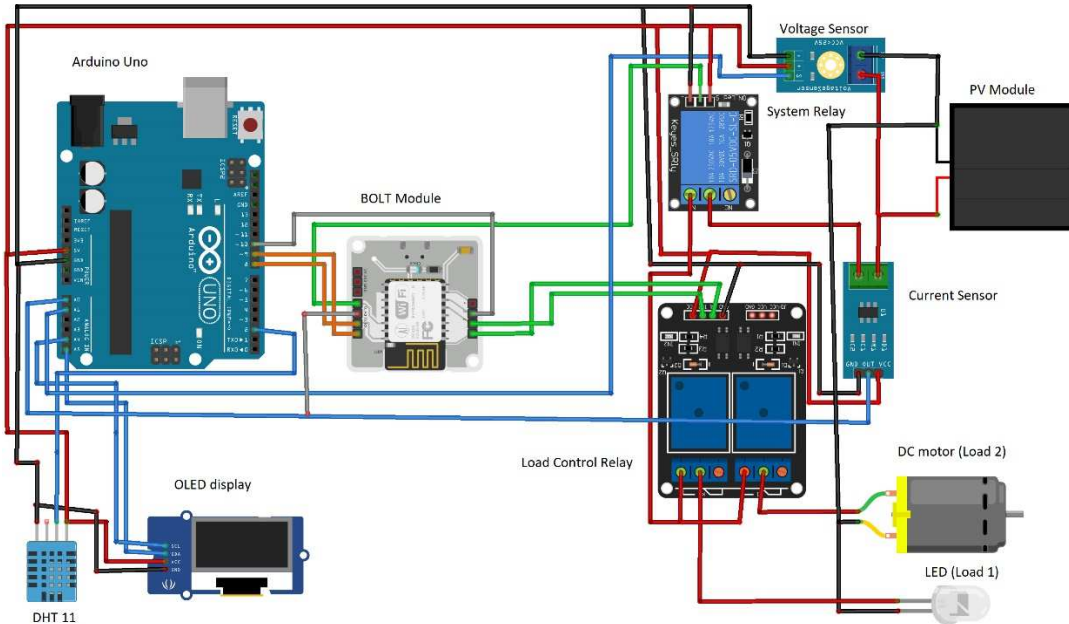


Fig 4: Circuit & connection diagram of the proposed system

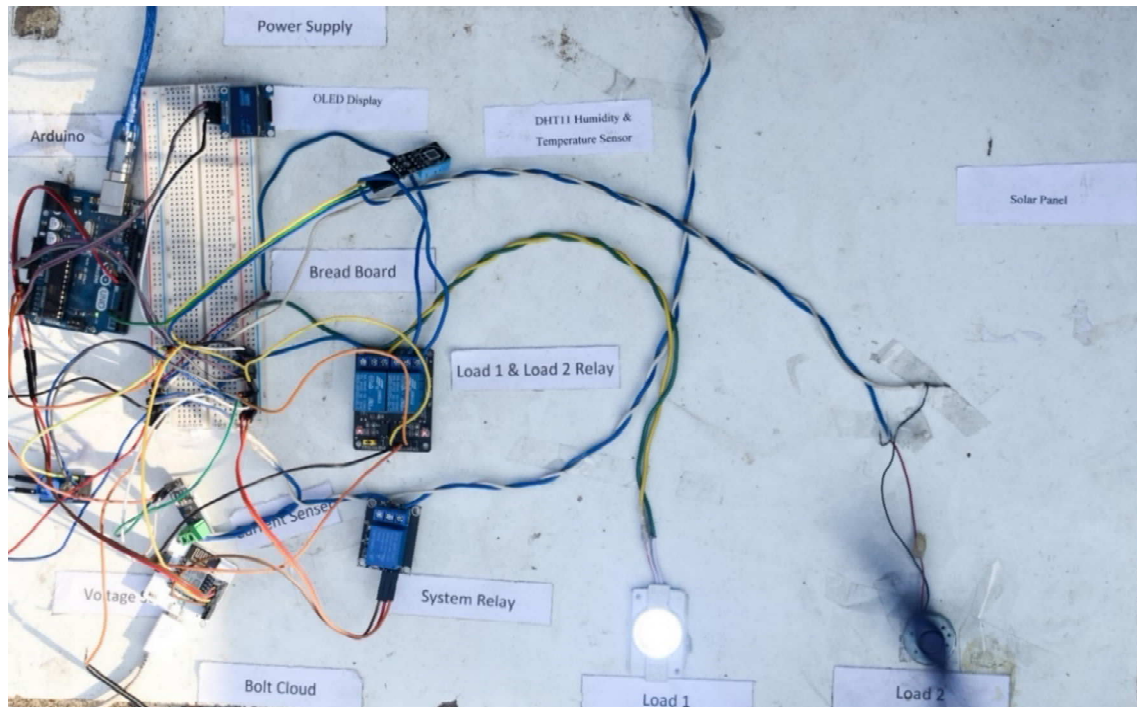


Fig 5. Actual prototype of the proposed system

D. Voltage Sensor

The voltage Detection sensor module [15] [19] could be a straightforward and extremely helpful module that uses a possible divider to reduce any input voltage by an element factor of five. This permits us to use the input pin which is an analog pin of a microcontroller to detect the potential difference higher than it is capable of sensing. For example, with a 0 Volt to 5 Volt input limit which is analog, you can sense a voltage up to 25 Volt. In this framework voltage sensor use to measure the voltage of the solar panel. It connects parallel with the terminal of the solar output.



Fig 6: Details of solar panel

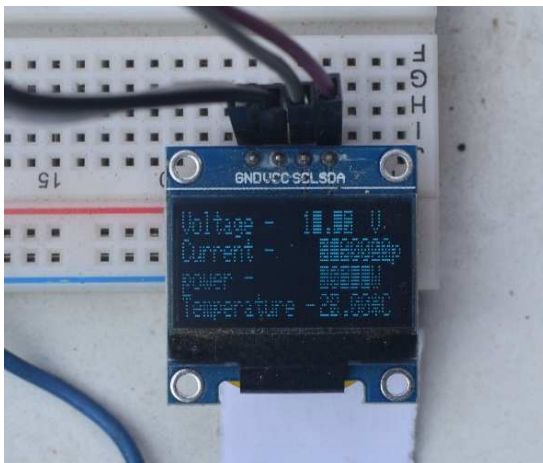


Fig 7: Output of organic light-emitting diode (OLED)



Fig 8: Output from BOLT IoT Module

E. Temperature & Humidity Sensor (DHT11)

The Temperature & Humidity Sensor (DHT11) is a necessary, chipset and advanced digital temperature and humidity sensing device [20]. It measures humidity through a capacitive sensor and temperature through a thermistor of the surrounding environment and sends a digital data signal on the input/output (I/O) pin of the microcontroller (no analog pins are used for this device). It's easy and simple to use but requires careful timing to collect the data. This device's working voltage is 3 to 5 Volt and 2.5 mA maximum current use during transformation (while mentioning information). Great for 20% to 80% dampness readings with 5% exactness. Great for 0 to 50°C temperature data readings with ±2°C accuracy.

F. Photo-Voltaic (PV) module

A solar panel, or photograph voltaic (PV) module, is a gathering of photograph voltaic cells mounted in a system for a creating functional solar panel. PV panels utilize sunlight as solar power which is a source of energy to produce DC electricity. A collection of Photo-Voltic modules is known as a Photo-Voltic panel or solar panel, and a framework of solar panels is known as an array [7] [21]. The collection of arrays of a photo-voltaic system produces solar electricity for

various electrical hardware. We use a 20-Watt max power solar panel whose max voltage is 17.0 V and current in max power is 1.18 Amp.

G. OLED Display

The OLED (Organic Light-Emitting Diode) show is an option for the LCD show. The OLED is super-light, nearly paper-slight, adaptable, and produces a more brilliant and crisper picture.

V. RESULTS

The solar parameter is seen through BOLT IoT and also see-through OLED. Voltage, Current, Power & temperature is showing BOLT IoT clouded through graph for remotely use and analysis purpose also showing in OLED. We can see the data through a laptop or Mobile. When any fault occurs or unwanted things happened, we can get an alert message [22] on mobile.

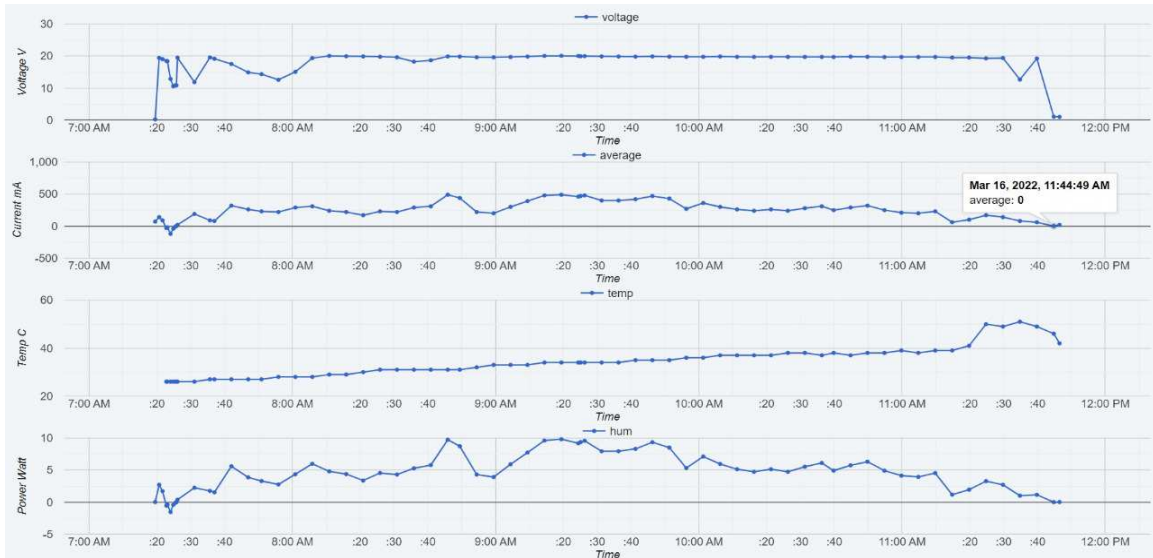


Fig 9: Line graph obtain from BOLT IoT Module

VI. CONCLUSION

The purpose of this project is to monitor real-time solar parameters from any location at any given time and also to control the system and load through a relay that is used as a switch. The proposed system has the capability to display the present and past recorded data. It has also an emergency alerting system when a fault occurs or voltage and current drop dangerously and also when the temperature crosses the threshold value. Moreover, the proposed system has the ability to remotely start and shut down the overall system and also ON/OFF the load with respect to the generation and demand. The Current Sensor, Voltage sensor, and relay are interfaced with Arduino Uno and BOLT IoT devices. The Voltage, current and DHT 11 sensor sense the data and send Arduino Uno, Arduino communicates BOLT device through serial communication and sends the data to the BOLT device. Arduino IDE is used for writing and compiling the program in C language and the BOLT device uses

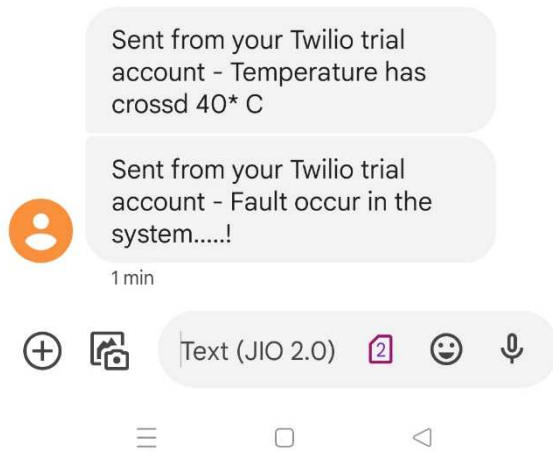


Fig 10: Fault alert SMS

JavaScript to collect data from the BOLT serial pin that is TX and TR and show the data through the graph. This proposed system will be beneficial for research objectives as well as for industrial applications. When the temperature exceeds the threshold value, it can shut down the entire system; likewise, when the voltage and current are inappropriate, the operator will be able to halt the load whenever deemed necessary. In addition, the system shows the data in a variety of ways, such as a bar graph or a line graph, in an excel datasheet that can be studied anytime for further research.

VII. REFERENCES

- [1] P. Visconti and G. Cavalera, "Intelligent system for monitoring and control of photovoltaic plants and for optimization of solar energy production," 2015 IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC), 2015, pp. 1933-1938, doi: 10.1109/EEEIC.2015.7165468.
- [2] R. K. Kodali and J. John, "Smart Monitoring of Solar Panels Using AWS," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), 2020, pp. 422-427, doi: 10.1109/PARC49193.2020.236645..
- [3] D. Kar and S. Kuntawar, "Remote Multi-parameters Monitoring and Controlling System for Solar panel Application," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 2020, pp. 1-6, doi: 10.1109/iSSSC50941.2020.9358894.
- [4] B. Shrihariprasath and V. Rathinasabapathy, "A smart IoT system for monitoring solar PV power conditioning unit," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), 2016, pp. 1-5, doi: 10.1109/STARTUP.2016.7583930.
- [5] Mayilvahanan, A.L., Stalin, N., & Sutha, S. (2018). Improving Solar Power Generation and Defects Detection Using a Smart IoT System for Sophisticated Distribution Control (SDC) and Independent Component Analysis (ICA) Techniques. *Wireless Personal Communications*, 102, 2575-2595.
- [6] A. S. Spanias, "Solar energy management as an Internet of Things (IoT) application," 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), 2017, pp. 1-4, doi: 10.1109/IISA.2017.8316460.
- [7] S. Rao et al., "A cyber-physical system approach for photovoltaic array monitoring and control," 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), 2017, pp. 1-6, doi: 10.1109/IISA.2017.8316458.
- [8] M. Alagumeenaakshi, S. Umamaheswari, A. Annisha Mevis, S. Seetha and G. Hema, "Monitoring and Controlling of Solar Photovoltaic Cells Using LoRa Technology," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675732.
- [9] S. M. Sorif, D. Saha and P. Dutta, "Smart Street Light Management System with Automatic Brightness Adjustment Using Bolt IoT Platform," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422668.
- [10] S. P, J. D, A. P, P. M, G. V and H. R, "Review on IoT Based Remote Monitoring for Solar Photovoltaic System," 2021 International Conference on Communication information and Computing Technology (ICCICT), 2021, pp. 1-5, doi: 10.1109/ICCICT50803.2021.9510163.
- [11] D. Saha, S. M. Sorif and P. Dutta, "Weather Adaptive Intelligent Street Lighting System With Automatic Fault Management Using Boltuino Platform," 2021 International Conference on ICT for Smart Society (ICISS), 2021, pp. 1-6, doi: 10.1109/ICISS53185.2021.9533234.
- [12] N. A. b. Anang Othman et al., "The Development of IoT-based Solar Battery Monitoring System," 2021 IEEE Regional Symposium on Micro and Nanoelectronics (RSM), 2021, pp. 34-37, doi: 10.1109/RSM52397.2021.9511610.
- [13] K. D. Deenadayalan, A. Arunraja, S. Jayanthi and S. Selvaraj, "IoT based Remote Monitoring of mass Solar Panels," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 1009-1014, doi: 10.1109/ICESC48915.2020.9155606.
- [14] R. I. S. Pereira, S. C. S. Jucá, P. C. M. Carvalho and C. P. Souza, "IoT Network and Sensor Signal Conditioning for Meteorological Data and Photovoltaic Module Temperature Monitoring," in *IEEE Latin America Transactions*, vol. 17, no. 06, pp. 937-944, June 2019, doi: 10.1109/TLA.2019.8896816.
- [15] M. S. Khan, H. Sharma and A. Haque, "IoT Enabled Real-Time Energy Monitoring for Photovoltaic Systems," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 323-327, doi: 10.1109/COMITCon.2019.8862246.
- [16] P. Srivastava, M. Bajaj and A. S. Rana, "IOT based controlling of hybrid energy system using ESP8266," 2018 IEEMA Engineer Infinite Conference (eTechNXT), 2018, pp. 1-5, doi: 10.1109/ETECHNXT.2018.8385294.
- [17] B. I. T. (I. P. Limited), "IoT Platform." BOLT. [Online]. Available: <https://www.boltiot.com/techspecs>. [Accessed: 31 Mar-2022]
- [18] M. M. A. Shah, M. S. Parvez, A. Ahmed and M. R. Hazari, "IoT Based Power Monitoring of Solar Panel Incorporating Tracking System," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1-4, doi: 10.1109/ACMI53878.2021.9528207..
- [19] P.G.Navamanikumar, S.Agnesha, P.Gowsalya, K.Indhu and N.Sivasakthi, " IOT Based Real Time Transformer Health Monitoring System and Phase Preventor," *International Journal of Emerging Technologies in Engineering Research*, Volume 6, Issue 4, April 2018.
- [20] S. Sadowski and P. Spachos, "Solar-Powered Smart Agricultural Monitoring System Using Internet of Things Devices," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 18-23, doi: 10.1109/IEMCON.2018.8614981.
- [21] J. Mathew and G. Vincent, "Realtime parameter monitoring and maximum power point estimation of solar photovoltaic array," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), 2016, pp. 1-4, doi: 10.1109/ICNGIS.2016.7854076.
- [22] "Communication APIs for SMS, Voice, Video and Authentication," Twilio. [Online]. Available: <https://www.twilio.com/>. [Accessed: 31- Mar-2022].

A Practical Approach To The Development Of A Decision-Supporting System Based On Fuzzy Neural Network In Information And Telecommunication Systems

Avaz Kuvnakov
Information technologies department
Tashkent University of Information
Technologies named after Muhammad
al-Khwarizmi
Tashkent, Uzbekistan
avaz_a@yahoo.com

Gulnora Mukhtarova
Information technologies department
Tashkent University of Information
Technologies named after Muhammad
al-Khwarizmi
Tashkent, Uzbekistan
gulnora6530@inbox.ru

Nurilla Mahamatov
Control and computer engineering
department
Turin Polytechnic University in
Tashkent
Tashkent, Uzbekistan
n.mahamatov@new.polito.uz

Nodira Malikova
Information technologies department
Tashkent University of Information
Technologies named after Muhammad
al-Khwarizmi
Tashkent, Uzbekistan
malikova@tuit.uz

Viktoria Kuznetsova
Information technologies department
Tashkent University of Information
Technologies named after Muhammad
al-Khwarizmi
Tashkent, Uzbekistan
kvb966@yandex.ru

Muqaddas Atadjanova
Information technologies department
Tashkent University of Information
Technologies named after Muhammad
al-Khwarizmi
Tashkent, Uzbekistan
matadjanova@tuit.uz

Abstract—This paper presents the result of the approach to the development of a decision support system (DSS) to improve the chosen location of base stations (BS) of mobile systems. The system is designed to improve the reliability of the decision-making and forecasting of the dynamics of the mobile transceiver systems and devices based on the uncertainty influences of a different nature. Analysis of these problems shows that an effective solution to this issue is to use the principles of the fuzzy set theory (FST) and modern geographic information systems (GIS) taking into consideration the geographically distributed topology of the information and telecommunications systems (ITS) elements. As a tool, a neuro-fuzzy inference system (ANFIS) in a Matlab environment is used to develop the decision support system and to select an optimal place geographical information system (GIS) is applied to find installation places of base stations of mobile companies taking into consideration geographic characteristics of the region. It also has been found that effective monitoring of the ITS in such information provision conditions primarily depends on the degree of compression of the input information.

Keywords— *Telecommunication systems, monitoring, fuzzy models, neural networks, Geographic information systems, decision making system.*

I. INTRODUCTION

This Modern rates of development of information and telecommunications systems (ITS) puts new demands on the development of the existing and new design of ITS, including the optimization of the existing ITS, and can significantly improve the quality of various types of information services. This is a priority task for many information and telecommunications sector enterprises that provides various services to the population [1,2,3,4].

The possibilities of existing research methods and decision-making for ITS solutions in case of territorial distribution of topological elements of ITS are limited.

The basis of the ITS is a spatially-distributed telecommunications infrastructure and to continuously monitor the status of ITS is necessary to conduct operational monitoring of ITS, which is the subject of this article.

Methodological bases used for the design, development, and optimization of the existing ITS have some disadvantages.

For statistical analysis of network characteristics, usually classical mathematical models are used, and largely focused on the use of numerical information, thus the information of fuzzy character is determined by the influence of various external factors used indirectly and insufficiently [5].

There are many methods to process data from the various type of information and telecommunications systems, which are using data mining technologies. The modern approach to processing data using data mining technologies allows the use of all types of data in modern information and telecommunications systems. Traditional methods of optimization and planning on the basis of their organizational arrangements for the maintenance and keeping gave the level of all the network resources required to provide quality services in the ITS due to industry dynamics and dominance of information heterogeneous nature does not allow a sufficient degree of control over all the resources.

Telecommunications infrastructure is distributed over large areas, thus there is the need to consider a variety of factors that characterize certain areas (topography, distance, construction, and so on.), and all needed information contained in the digital terrain models and digital maps. The main tool for working with these maps and models is the geographic information systems (GIS) [6].

GIS in the field of ITS may be used for:

- Strategic planning, demand analysis, and forecasting market development of telecommunication networks;

- Design and development of telecommunication networks, including spatial analysis and modeling of the network:
 - selection of locations for the antennas, repeaters with the relevant calculations of service areas, from certain points of view (antenna location), modeling propagation, etc.
 - determine the optimal cable routing based on the location of streets, highways and railways, various underground utilities, as well as data about the owners of the land, etc.
- The inventory of telecommunication networks of distributed infrastructure companies, technical documentation, including graphical, for a variety of distributed, often over a considerable area, and difficulties associated with the environment and each other object.
- Organization of a network of customer service and payments for delivered services.
- Analysis of the company and the quality of delivered services to the customer. Operative response to accidents and emergencies.
- Monitoring of networks and prevention of emergencies.
- Analysis of correspondence of boundaries and service area attributable to her workload, redefining areas.
- Optimization of travel and transportation, routing of official vehicles.
- Providing additional services using communication devices (Value-added services).

II. MATERIALS AND METHODS

To take into account the dynamic variability and Decision-making under conditions with the prevalence of heterogeneous information, and to obtain an adequate assessment of the status of ITS in time and space, appropriate to integrate geographic information systems modeling, as well as the application of the principles of fuzzy set theory (FST) in the decision-making process.

A complex assessment of the status of ITS should include an analysis of the topology of the network, channel capacity based on modern information visualization, and organizing interactive interfaces. The implementation of these approaches based on modern geo-information modeling technology is relevant and poorly investigated.

Realization of this approach is performed based on a considered hypothetical object – the town “Nurafshon” with an installed base station (BS). (Figure 2).

BS - complex radio transmitters and the main element of the mobile network. It is necessary to ensure that delivering affordable, high-quality mobile communications to all users in the territory of the district in which is installed. During making a call, the mobile phone is connected through a radio

channel to the nearest base station, and then on the base station network connection is established with the called party. The more stations, the better the communication quality [4].

Generally, a set of base station equipment is placed on the roofs of buildings (antenna device) and in technical areas (channeling equipment). Considering the complexity of the radio frequency (RF) environment, topology construction, and high population density in the “Nurafshon” area base stations (radio waves of low power) are placed not only on the objects that dominate the building areas but also on low buildings, including road transport infrastructure (installation on outdoor lighting, traffic lights, etc.).

The concrete location of base stations is determined by the communication operator independently as a result of complex technical calculations and radio measurements, based on the need to provide quality and security of services to:

- covering by qualitative radio signal of certain territory;
- provide electromagnetic compatibility with other radio electronic facilities;
- provide sanitary zones and set acceptable standards of radio emission levels.

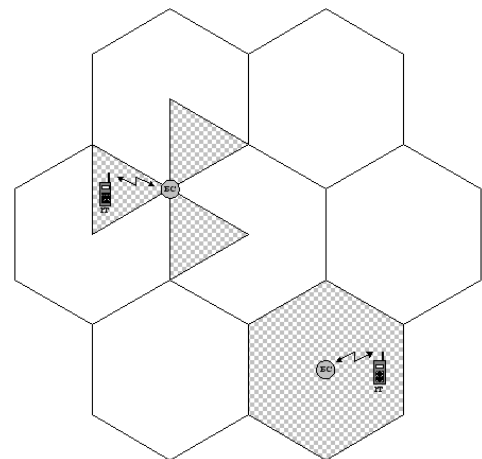


Figure.1 Mounting the BS locations. (vrednost.ru)



Figure 2. Hypothetical map of the town "Nurafshon".

The main characteristics of the BS can be listed as:

1. The height of the BS - H_{BS} (in meters)
2. The signal power - P_{BS} (dBm)
3. The number of subscribers - C_n (numbers)

Planning a cellular network starts with a plan that shows the estimated location of the installation of base stations in the territory, which are expected to deploy in the network. Such a plan is little tied to the area characteristics and existing objects in this infrastructure and is usually called a nominal plan. After forming the requirements designer must conduct a preliminary calculation of the coverage area. At this stage performed searching appropriate areas for placing the BS.

III. RESULTS

In order to analyze the place of installation of BS antennas in the district area, we need specific details about antenna suspension heights and the structure of the BS location in the areas. Further must be specified the type of user equipment and formed a detailed design of the future network.

Installation, configuration, and upgrading of the base stations are performed based on the following algorithm (Figure 3).

In addition to technical aspects of planning must be considered an important economic component of the process to minimize cellular facilities at an appropriate level of coverage and quality of services.

Further follows the construction of the BS and all the key elements of the network. Before the commercial launching, from the operator side a frequency plan, setting up hardware and software complex base stations of the system, and the test of the systems must be performed.

Many factors affect the quality of signals (for example, it may overlap the surrounding tall buildings and interfere with other radio signals), since the radio signal is distributed not in a straight line. All these factors are taken into account during planning radio networks and accordingly locations for base stations.

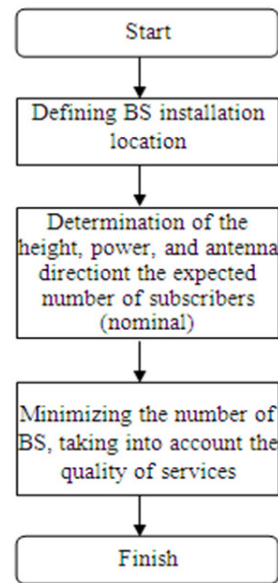


Figure 3. The algorithm of designing and construction of the BS mobile systems

The main tasks of the mobile operator is providing reliable communications to their subscribers, by placing and maintenance of the BS and prevent electromagnetic energy flux density level not more than $10 \text{ mW} / \text{sm}^2$ for most countries. In the Uzbekistan $2,5 \text{ mW} / \text{sm}^2$ is allowed. Because the last research shows that, this level of electromagnetic fields can not have a negative impact on the human body.

Antenna height varying between 20 meters to 100 meters (in case antenna installed on the roof of high-rise building). The rule is, the closer to the telephone BS is located, and the cellphone spend less radiation on setup and maintaining communication with the base station.

So the more base stations, the smaller the distance from the cellular phone to the subscriber station, the lower level of electromagnetic radiation field a subscriber cellular phone.

As analysis shows in order to find best location to install BS must be considered different internal and external factors where heterogeneous information is exists. In this case to obtain best location and adequate assessment location of BS in time and in space can be used approach by integrating geographic information systems modeling, as well as the application of the principles of fuzzy set theory (FST) [8].

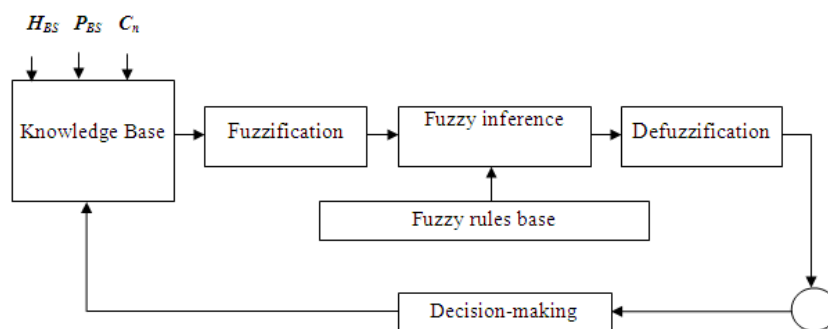


Figure.3.Fuzzy model

ITS optimization performed based on the construction of neuro-fuzzy model, on the this basis conducted a series of

numerical experiments in order to select the optimal functioning of the ITS and location of elements.

$$C_n = f(H_{BS}, P_{BS}) \quad (1)$$

Where, H_{BS} - BS elevation (in meters), P_{BS} - Signal strength (in dBm) and C_n - Number of subscribers (number).

Table 1. Rule Base

Rules №	Elevation, m		Signal power dBm		Number of Subscribers	
	H_{BS}	P_{BS}	P_{BS}	C_n	C_n	C_n
1	Very low	1	Significantly Below Normal	1	Significantly few	1
2	Low	32	Below normal	2	Few	25
3	Average	53	Normal	5	Average	74
4	High	85	Above norm	9	Big	153
5	Significantly High	100	Significantly higher than normal	10	Significantly Big	200

In the calculations, the parameters H_{BS} , P_{BS} and C_n are accepted as fuzzy values, which allows us to compress the information, which involves taking into account linguistic estimates of fuzzy variables of H_{BS} , P_{BS} and C_n .

CONCLUSION

Decision support systems for the operation of ITS depend on the use of diverse information about network topology and capacity of communication channels, subscriber access, and others. Effective monitoring of the ITS in such information provision conditions primarily depends on the degree of compression of the input information. An effective solution to this problem is to integrate the principles of the FST and modern GIS to account for the geographically distributed topology of the ITS elements. The location of the base stations of the telecommunications system is mainly determined by many characteristics, which can be taken as a time factor, changes in the terrain, the height of the transmitting antennas, and their radiation. Further research is planned as follows, including into model the results of an expert survey and adding other factors that affect the location of the Base stations, taking into account the electromagnetic radiations of the area, where the telecommunications network is deployed.

ACKNOWLEDGEMENTS

The research work has been funded by Innovation project of The Ministry for Development of Information

Technologies and Communications of the Republic of Uzbekistan №A5-022.

REFERENCES

- [1] A. Valdar, Understanding telecommunications networks: 2nd edition. 2006. doi: 10.1049/pbte071e.
- [2] R. T. Wong, "Telecommunications network design: Technology impacts and future directions," Networks, vol. 77, no. 2, 2021, doi: 10.1002/net.21997.
- [3] A. E. Kuvnakov and S. S. Kasimov, "Development internet resources in Uzbekistan: Empirical investigation," 2010. doi: 10.1109/ICAICT.2010.5612068.
- [4] G. Anandalingam and S. Raghavan, "Telecommunications network design and management," Oper. Res. Comput. Sci. Interfaces Ser., vol. 23, 2003, doi: 10.1007/978-1-4757-3762-2.
- [5] V. Raheja and S. Mahajan, "Decision Support System , Its Components , Model and Types of Managerial Decisions," Int. J. Innov. Res. Stud., vol. 2, no. 12, 2013.
- [6] G. Bonham-Carter, "Geographic information systems for geoscientists: modelling with GIS," in Computer methods in the geosciences, 2014.
- [7] L. A. Zadeh, "'Fuzzy sets', Information and Control, Vol. 8," 1965.
- [8] Z. Xakimjon and K. Muslimjon, "Modeling of Geophysical Signals Based on the Secondorder Local Interpolation Splay-Function.," 2019. doi: 10.1109/ICISCT47635.2019.9011853.
- [9] N. Mahamatov, J. Lee, and M. Lee, "Telecommunication services diffusion in CIS: Dynamics of competition and complementarity among telecommunication services," 2007. doi: 10.1109/canet.2007.4401678.

A UNIVERSAL METHOD FOR SOLVING THE PROBLEM OF BENDING OF PLATES OF ANY SHAPE

Dr. Azamatjon Yusupov, PhD
Andijan Machine Building Institute
Andijan, Uzbekistan
azamat7uzbek@gmail.com

Dr. Mirzaeva Manzura
Andijan State University
Andijan, Uzbekistan

Dr. Tajibaev Gayratjon
Andijan State University
Andijan, Uzbekistan

Uzakov Sirojiddin
Andijan State University
Andijan, Uzbekistan

Abstract— This paper presents a universal algorithm for solving boundary value problems of bending plates of arbitrary shape. To solve the problem of bending, the method of sources and sinks is used. According to this method, sources and sinks are distributed continuously along a line located outside the area occupied by the plate and similar to the contour of the original plate. By choosing their powers, the conditions at the plate boundary are satisfied.

In this paper, we use an elementary solution for the problem of bending a round supported plate and fundamental solutions from a unit transverse force and moment concentrated at a point. Mathematical expressions corresponding to various boundary conditions are given. The boundary value problem is reduced to a system of integral equations. To obtain a stable solution to the system of integral equations, the regularization method with minimization of the Tikhonov functional is applied, the numerical implementation of which will lead to systems of algebraic equations. An algorithm for solving the problem in the MATLAB system is proposed.

Keywords— boundary value problem of thin elastic plate bending, source and sink method, regularization method, fundamental solutions from concentrated forces.

I. INTRODUCTION

The study of the stress-strain state of plates is carried out on the basis of general methods for solving boundary value problems of the technical theory of elasticity and is reduced to the integration of differential equations in high-order partial derivatives. The problems of plate bending are so complex and diverse that almost all the main methods of mathematical physics have been used with varying degrees of efficiency to solve them.

Despite the presence of a large number of works devoted to the theory of calculation of plates, the problem of bending still requires further study, since interest in the diversity of the shape of buildings, structures and technical products grows over time. And most parts of the structures consist of thin plates, with a decrease in material consumption and labor intensity of manufacturing, which requires the

study of the stress-strain state. Therefore, a single algorithm is needed to create a computer tool that allows solving this problem under arbitrary loadings of a complex configuration with any fastenings, and this work is devoted to this.

Formulation of the problem. Let us consider the problem of the stress-strain state of a thin elastic plate, which occupies a region G , with a boundary G . As is known, a small transverse deflection of a thin elastic plate that satisfies the Kirchhoff-Leyva hypothesis is determined by the differential equation [1]:

$$\frac{\partial^4 w}{\partial x_1^4} + 2 \frac{\partial w}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 w}{\partial x_2^4} = \frac{\rho}{D} \quad (1)$$

$$D = Eh^3/12(1 - \nu^2), x = (x_1, x_2) \in G$$

where $w = w(x) = w(x_1, x_2)$ – plate deflection, ρ – transverse load intensity, D – bending stiffness of the plate, E – is the modulus of normal elasticity of the plate material, ν – is its Poisson's ratio, h – is the plate thickness.

The deflection $w(x_1, x_2)$ must satisfy the boundary conditions:

$$L_i w = 0, \quad i = 1, 2 \quad (2)$$

L_i – boundary condition operator.

The most common conditions on the contour are as follows [1] if the edge is pinched,

$$w(x_1, x_2) = 0, \quad \frac{\partial w}{\partial n_x} = 0$$

if the edge is freely supported, then:

$$w(x_1, x_2) = 0, \\ M_{n_x}(x_1, x_2) = 0$$

if the edge is free, then:

$$M_{n_x}(x_1, x_2) = 0, \\ Q_{n_x}^*(x_1, x_2) = 0$$

where $M_{n_x}, Q_{n_x}^*$ – respectively the bending moment and the reduced transverse force, n_x, s – the corresponding outward normal and tangent to the boundary Γ , in accordance with the assumptions made in Fig. 1 rule signed bending moment and reduced shear force

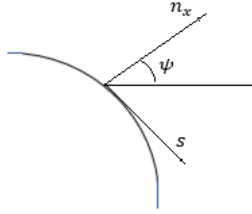


Figure. 1. Outer normal and tangent to the boundary Γ

$$\begin{aligned} \frac{\partial w}{\partial n_x} &= \frac{\partial w}{\partial x_1} \cos \psi + \frac{\partial w}{\partial x_2} \sin \psi \\ M_{n_x} &= M_{x_1} \cos^2 \psi + M_{x_2} \sin^2 \psi \\ &\quad + 2M_{x_1 x_2} \sin \psi \cos \psi \\ M_{n_{xt}} &= M_{x_1 x_2} \cos 2\psi + \frac{1}{2}(M_{x_1} - M_{x_2}) \sin 2\psi \\ Q_{n_x} &= Q_{x_1} \cos \psi + Q_{x_2} \sin \psi \\ Q_{n_x}^* &= Q_{n_x} - \frac{\partial M_{n_{xt}}}{\partial s} \\ \frac{\partial M_{n_{xt}}}{\partial s} &= \frac{\partial M_{n_{xt}}}{\partial x_1} (-\sin \psi) + \frac{\partial M_{n_{xt}}}{\partial x_2} \cos \psi \\ &\quad + \frac{1}{\rho} \frac{\partial M_{n_{xt}}}{\partial \psi} \end{aligned}$$

Where s – tangent to contour; ρ – contour curvature, $\psi = (\widehat{n_x, x_1})$, $M_{x_1}, M_{x_2}, M_{x_1 x_2}, Q_{x_1}, Q_{x_2}$ – are expressed through deflection according to known formulas [1]:

$$M_{x_i} = -D \left(\frac{\partial^2 w}{\partial x_i^2} + \nu \frac{\partial^2 w}{\partial x_j^2} \right), \quad i = 1, 2, \quad j = 1, 2, \quad i \neq j$$

$$M_{x_1 x_2} = -M_{x_2 x_1} = D(1 - \nu) \frac{\partial^2 w}{\partial x_1 \partial x_2}$$

$$Q_i = -D \frac{\partial}{\partial x_i} \left(\frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} \right), \quad i = 1, 2$$

For example, consider a mixed boundary value problem for plates having the shape shown in Fig. 2.

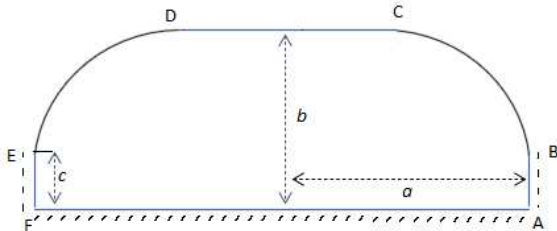


Figure. 2. Mixed boundary value problem for thin plates.

$$\begin{aligned} w(x_1, 0) &= 0, \quad \left. \frac{\partial w}{\partial n_x} \right|_{x_2=0} = 0, \quad -a \leq x_1 \leq a \\ AB \text{ и } EF &\text{ – freely supported boundaries:} \\ w(a, x_2) &= 0, \quad M_{n_x}(a, x_2) = 0 \\ w(-a, x_2) &= 0, \quad M_{n_x}(-a, x_2) = 0 \\ 0 \leq x_2 &\leq c \\ BC &\text{ – free edge, etc.} \end{aligned}$$

for CD edge: $M_{n_x}(x_1, b) = 0, Q_{n_x}^*(x_1, b) = 0, -a + r \leq x_1 \leq a - r$, here $r = b - c$

for BC edge: $M_{n_x}(x_1, x_2) = 0, Q_{n_x}^*(x_1, x_2) = 0, a - r \leq x_1 \leq a; c \leq x_2 \leq b$ for DE edge: $M_{n_x}(x_1, x_2) = 0, Q_{n_x}^*(x_1, x_2) = 0, -a \leq x_1 \leq -a + r; c \leq x_2 \leq b (x_1 - (a - r))^2 + (x_2 - c)^2 = r^2$,

II. METHODOLOGY

The solution $w(x_1, x_2)$ of equation (1), due to the linearity of the differential operator, can be represented as $w(x_1, x_2) = w_0(x_1, x_2) + \bar{w}(x_1, x_2)$ (3)

here w_0 – is a particular solution of equation (1) with the right side, and the function \bar{w} is a general solution of a homogeneous equation in the domain G .

As w_0 , you can take any function that satisfies (1). Here we use an elementary solution for the problem of bending a round supported plate [2].

$$w_0(x_1, x_2) = \frac{\rho r^4}{64D\pi}, \quad r = \sqrt{x_1^2 + x_2^2} \quad (4)$$

As \bar{w} , we take the solution from the transverse force and moment distributed continuously along the line G' (Fig. 3), located outside the region $G (G' \in \mathbb{R}^2/G)$ and similar to the contour of the original plate. In this case decision is

$$\bar{w}(x) = \int_{G'} [W_1(x, y)q(y) + W_2(x, y)m(y)] dl_y \quad (5)$$

where $y = (y_1, y_2) \in G', W_1(x, y), W_2(x, y)$ – fundamental solutions from a single transverse force and moment concentrated at the point y , respectively $q(y), m(y)$ – (source power), located along G' .

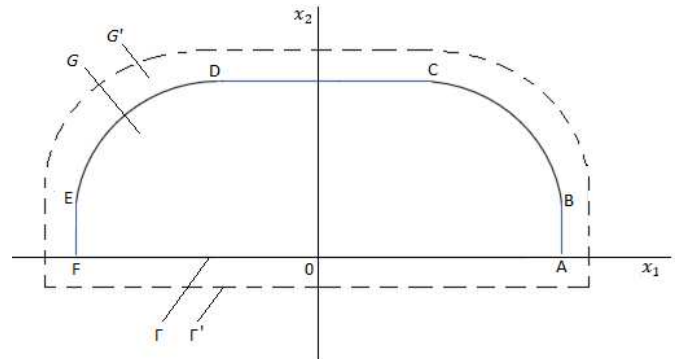


Fig 3. Auxiliary circuit for applying the method.

Fundamental solutions are known in [3].

$$W_1(x, y) = \frac{1}{16\pi D} \bar{r}^2 \ln \bar{r}^2, \quad W_2(x, y) = -\frac{\partial W_1(x, y)}{\partial n_y} \quad (6)$$

here $\bar{r} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$, n_y – outward normal to Γ' .

From (3)-(6) it follows that the solution of equation (1), expressed through the unknown functions $q(y)$ and $m(y)$, has the form

$$w(x) = w_0(x) + \int_{\Gamma'} [W_1(x, y)q(y) + W_2(x, y)m(y)] dl_y, \quad (7)$$

The function $w(x)$ exactly satisfies equation (1). The unknown functions are determined from the conditions that the given relations (2) hold on the boundary Γ . At the same time, no restrictions are imposed on the forms of the boundaries Γ and on the type of fastening, they can be liked, i.e. the algorithm of this method is universal.

Applying to $w(x)$ in (7) the corresponding differential operators L , we obtain the expressions $\frac{\partial w}{\partial n}$, M_n , Q_n , Q_n^* . As is known, these differential operators are linear: $L[w(x)] = L[w_0(x)] + L[\bar{w}(x)]$ or by (7) we have

$$L[w(x)] = L[w_0(x)] + \int_{\Gamma'} [L[W_1(x, y)]q(y) + L[W_2(x, y)]m(y)]dl_y \quad (8)$$

With this in mind, below are the expressions for each

$$L[W_i(x)], \quad i = 0, 1, 2$$

These expressions were used the following notation:

$$r^2 = x_1^2 + x_2^2, \quad \bar{r}^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2$$

$$\begin{aligned} (x_1, x_2) \in \Gamma, & \quad (y_1, y_2) \in \Gamma', \\ \theta = (\widehat{y_1, n_y}), & \quad \psi = (\widehat{x_1, n_x}), \\ S(a, b, \alpha) = a \cos \alpha + b \sin \alpha, & \quad C_1 = \\ S(x_1, x_2, \psi), & \\ C_2 = S(x_2, x_1, \psi), & \quad S_1 = \\ S(-x_2, x_1, \psi), & \\ S_2 = S(x_1, -x_2, \psi), & \quad C_\theta = S(x_1 - y_1, x_2 - \\ & \quad y_2, \theta), \end{aligned}$$

$$\begin{aligned} S_\theta = S(x_1 - y_1, -(x_2 - y_2), \theta), \\ C_\psi = S(x_1 - y_1, x_2 - y_2, \psi), \\ S_\psi = S(x_1 - y_1, -(x_2 - y_2), \psi), \end{aligned}$$

A. Expressions for $w(x)$

$$1^\circ) w_0(x) = \frac{pr^4}{64\pi D}$$

$$2^\circ) W_1(x, y) = \frac{\bar{r}^2 \ln \bar{r}^2}{16\pi D}$$

$$3^\circ) W_2(x, y) = \frac{(\ln \bar{r}^2 + 1)C_\theta}{8\pi D}$$

B. Expressions for $\frac{\partial w}{\partial n_x}$

$$1^\circ) \frac{\partial w_0}{\partial n_x} = \frac{pr^2 C_1}{16\pi D}$$

$$2^\circ) \frac{\partial W_1}{\partial n_x} = \frac{(\ln \bar{r} + 1)C_\psi}{8\pi D}$$

$$3^\circ) \frac{\partial W_2}{\partial n_x} = \frac{(-C_\psi C_\theta (\ln \bar{r}^2 + 3) + S_\psi S_\theta (\ln \bar{r}^2 + 1))}{8\pi D \bar{r}}$$

C. Expressions for M_n

$$1^\circ) M_n^0 = \frac{p}{16} (r^2(-1) + 2(S_2^2 + \gamma C_2^2))$$

$$2^\circ) M_n^1 = -\frac{1}{8\pi} \left\{ (1 + \nu)(\ln \bar{r}^2 + 1) + \frac{2}{\bar{r}^2} [C_\psi^2 + \nu S_\psi^2] \right\}$$

$$3^\circ) M_n^2 = \frac{1}{4\pi \bar{r}^2} [C_0(1 + \nu) + 2S_\psi C_\psi S_\theta(1 - \nu)]$$

D. Expressions for Q_n

$$1^\circ) Q_n^0 = -\frac{pC_1}{2}$$

$$2^\circ) Q_n^1 = \frac{-1}{2\pi \bar{r}^4} C_\psi$$

$$3^\circ) Q_n^2 = -\frac{1}{2\pi \bar{r}^4} [C_\theta C_\psi - S_\theta S_\psi]$$

E. Expressions for Q_n^*

$$1^\circ) Q_{0n}^* = \frac{p}{8\pi} (\rho(\nu - 1)[C_1^2 - S_1^2] - (5 - \nu)C_1)$$

$$2^\circ) Q_{1n}^* = \frac{1}{4\pi \bar{r}^2} \left[(1 - \nu)(C_\psi^2 - S_\psi^2) \left(\rho - \frac{C_\psi}{\bar{r}^2} \right) - 2C_\psi \right]$$

$$3^\circ) Q_{2n}^* = \frac{1}{4\pi \bar{r}^4} \left[(C_\theta C_\psi - S_\theta S_\psi) \left((1 - \nu) \frac{1}{\bar{r}^2} (S_\psi^2 - C_\psi^2) - 2 \right) + 4S_\theta C_\psi S_\psi \left(\frac{C_\psi}{\bar{r}^2} - \rho \right) \right]$$

Using 1^o), 2^o), 3^o) it is easy to obtain representations of the components of the stress-strain state $\frac{\partial w}{\partial n}$, M_n , Q_n , Q_n^* of the plate in terms of arbitrary functions $q(y)$, $m(y)$.

Substituting expression (7) into the boundary conditions, we obtain a system of two integral equations of the 1st kind with respect to q , m with continuous kernels:

$$\int K(x, y)z(y)dl_y = u(x), \quad x \in \Gamma \quad (9)$$

or in operator form

$$Kz = u,$$

here $K(x, y)$ – matrix with elements $K_{ij}(x, y) = L_i W_j$, $i = 1, 2$, $j = 1, 2$; $z(y)$ – unknown vector with components q , m ; $u_i(x) = -L_i[w_0(x)]$; L_i – differential operator of boundary conditions.

Having solved system (9) and substituting the found functions $q(y)$, $m(y)$ into formula (7), we obtain the solution of this boundary value problem.

The solution of the system of integral equations is unstable. It was shown in [4] that the stability of the approximate solution of system (9) depends on the distance

$$d = \rho(\Gamma, \Gamma')$$

In [5], a technique was developed for obtaining a stable solution using the regularization method. In this technique, the approximate solution of system (9) is obtained by minimizing the Tikhonov functional [6]:

$$M_\alpha(z) = \|K_h z - u_\delta\|_\Gamma^2 + \alpha \|z\|_{\Gamma'}^2, \quad \alpha > 0$$

with the choice of the regularization parameter α from the «principle of least residual».

$$\|K_h z_\alpha - u_\delta\|_\Gamma + h \|z_\alpha\|_{\Gamma'} \rightarrow \min$$

here K_h – numerical analogue of the integral matrix K , u_δ – numerical analogue of the vector $u(x)$

$$h > 0, \quad \delta > 0, \quad \|K_h - K\| \leq h, \quad \|u_\delta - u\| \leq \delta$$

With the help of quadrature formulas, we replace the system of integral equations (9) with a system of linear algebraic equations (SLAE) and from the condition of the minimum of the functional

$$\frac{\partial M_\alpha(z)}{\partial z} = 0$$

We obtain a system of algebraic equations

$$(A'BA + \alpha D)z = A'Bu \quad (10)$$

here A – order matrix $2n_1 \times 2n_2$; n_1, n_2 – the number of nodes on Γ, Γ' , respectively; B and D – symmetric positive definite matrices of square forms of discrete analogs of scalar products in spaces L – a vector of functions on the curves Γ, Γ' , respectively; u, z – vectors with components $u(x_i), z(y_i)$. at the split points of the curves Γ and Γ' .

Solutions to system (10) are obtained using MATLAB, with the choice of the regularization parameter from the conditions

$$(B(Az_\alpha - u_\delta), Az_\alpha - u_\delta)^{\frac{1}{2}} + h(z_\alpha, z_\alpha)^{\frac{1}{2}} \rightarrow \min_\alpha,$$

at

$$(B(Az_\alpha - u_\delta), (Az_\alpha - u_\delta)) \leq \varepsilon,$$

where ε – accuracy.

According to the obtained values of z and according to the numerical analogues of formula (7):

$$\tilde{w}(x) = w_0(x) + \sum_{j=1}^{n_2} C_j [W_1(x, y_j)q(y_j) + W_2(x, y_j)m(y_j)]$$

It is possible to approximately determine the value of the deflection at arbitrary points of the plates.

Solution algorithm. We solve this problem numerically in MATLAB.

When programming, there is no need to write equations for each edge separately. It is enough to assign a parameter for the boundary condition and compose the algorithm so that the above expressions are selected by the value of this parameter. The parameter values are the boundary condition code.

For a complete solution of any considered bending problem, it is necessary to specify the following information:

LF – number of edge rows, separated by boundary conditions and by curvature;

B1(I), B2(I) – coordinates of the starting point of the 1st side;

O1(I), O2(I) – center of curvature of the 1st curved side;

KK(I) – view of the I-th side of the edge of the plate:

KK(I)=0 – straight edge,

KK(I)=1 – curvilinear edge convex down,

KK(I)=2 – curvilinear edge convex up;

L(I) – type of boundary conditions specified on the 1st side:

1 - pinched,

2 - freely supported,

3 - free;

MN(I) – number of nodes of points on the 1st side;

AN – parameter determining from $AN = h_2/h_1$,

h_2 – step on the additional contour,

h_1 – on the main line;

D(I) – the distance between the additional and main circuits on each I - th side;

Q – external load intensity;

(M,N) – dimension of the resulting system of linear algebraic equations;

HP – plate thickness;

E – elastic modulus;

ν – Poisson's ratio.

The implementation of this method is carried out according to the following algorithm:

1. Entering and printing data;

2. Geometry subroutine;

3. Organization of calculations $\frac{\partial w}{\partial n}, M_n, Q_n, Q_n^*$ by L(I) using the formula A)-E);

4. Formation and solution of SLAE (10);

5. Calculation and verification of residual $\|Kz - u\| \leq \varepsilon$;

6. Selection of the regularization parameter;

7. Calculation of the stress-strain state;

8. Print results.

III. CONCLUSION

This paper proposes a technique for numerically solving basic and mixed linear boundary value problems in the theory of thin plate bending. The resulting expressions are presented, which make it possible to write out a complete system of relations describing the bending of thin plates. Algorithms for the implementation of the developed methodology are proposed.

Here, a small transverse deflection of a thin elastic plate is considered, the equation of state of which is described by a linear differential equation. The linearity of the equation of state made it possible to reduce boundary value problems to integral equations. This method can be applied to any linear boundary value problem, only in this case it is necessary to have a particular solution and fundamental solutions for the differential equation under consideration. When choosing these solutions, it is necessary to take into account the differentiability of the required degree, otherwise integral equations will not be obtained.

Solving a system of integral equations is not always possible. Therefore, here we use numerical methods. The numerical solution is ill-conditioned, since the smaller the step of partitioning the interval, the closer to zero the determinant of the matrix of the numerical analogue of the problem. Therefore, the regularization method is used, which minimizes the Tikhonov functional [6] with the choice of the regularization parameter. In this case, we obtain a system of algebraic equations with the regularization parameter (10). We solve this system with different values of the regulation parameter until the accuracy of the calculations is satisfied. To form and solve a system of algebraic equations, we consider it more practical to use the MATLAB system. It is also convenient for the graphical representation of the result. Therefore, an algorithm for implementing the method in the integrated MATLAB environment has been developed.

REFERENCES

- [1] Timoshenko S.P., Voynovskiy-Kruger S. *Plastini i obolochki*. – M.: Fizmatiz, 1963. – 635 s.
- [2] Konchikovskiy Z. *Pliti. Sistematicheskie rascheti*. – M.: Stroyizdat, 1984. – 481 s.
- [3] Vensel' E.S. *Primenenie metoda kompensiruyushix nagruzok k raschetu plastin slojnoj formu*. – Dokl.AN USSR, 1980, №3, ser.A. – 43 – 45 s.
- [4] Raxmatulin X.A., Tkacheva G.D. *Reshenie zadach ob udare uprugogo kol'sa o jestkuyu pregradu s uchetom volni razrusheniya*. – V sb. statey. *Mexanika deformiruyushix tel i konstruksiy*. – M. «Mashinostroeniye», 1975. – s. 409-414.
- [5] Vensel' E.S., Kobilinskiy V.G., Levin A.N. *Primenenie metoda regulyarizatsii dlya chislennogo resheniya zadachi izgiba tonkix uprugix platinok*. – J.vichisl.matem. i matem.fiz., №2,1984. – S 323-328.

- [6] Tixonov A.N., Goncharskiy A.V., Stepanov V.V., Yagola A.G.
Regulyariziruyushie algoritmi i apriornaya informatsiya. – M.:
«Nauka», 1983. – 198 s.

DEVELOPMENT OF STOCHASTIC DISTRIBUTION MODEL OF CONTAMINATED WATER TREATMENT COMPLEX

1st Bakhadir Begilov

Faculty of Computer Engineering

Nukus branch of Tashkent University of

Information Technologies named after Muhammad al-Khwarizmi

Nukus, Uzbekistan

bakhadirbegilov@gmail.com

Abstract—The object of the research is the treatment facilities of the city of Nukus.

Purpose of the study. The aim of the study is to develop and analyze mathematical models and algorithms for solving deterministic problems of polluted water purification, establishing the ratio of duality and optimality conditions, as well as establishing a marginal ratio for these tasks and determining the optimal size of costs for the purification of polluted water with various technological treatment schemes, development of a software package for solving deterministic problems of the complex for the purification of polluted waters.

In general, on the research topic under study, theoretical studies were carried out to determine the optimality conditions for solving optimization problems and the stability of the deterministic model of the complex for the purification of polluted waters.

The results of the study allow, determine the optimal size of the cost of treating polluted water with various technological treatment schemes.

Index Terms—Stochastic, water, treatment, model

I. INTRODUCTION

Water is the most valuable natural resource. It plays an exceptional role in the metabolic processes that form the basis of life. Water is of great importance in industrial and agricultural production. It is well known that it is necessary for the everyday needs of man, all plants and animals. For many living creatures, it serves as a habitat [1], [2].

The shortage of fresh water is already becoming a global problem [3]–[5]. The ever-increasing needs of industry and agriculture for water are forcing all countries and scientists of the world to look for various means to solve this problem.

At the present stage, the following directions of rational use of water resources are determined: more complete use and expanded reproduction of fresh water resources; development of new technological processes to prevent pollution of water bodies and minimize the consumption of fresh water.

The main schemes of treatment facilities were described back in 1949 by Professor, Doctor of Technical Sciences B.O.

Botuk [6]. In his book Domestic Wastewater Treatment, he describes the first laboratory studies on the use of an aeration tank for biological wastewater treatment, carried out in 1912 [7]–[11].

The authors of the article Formation, treatment and use of wastewater gave the following definition to the biological treatment process: Biological treatment involves the purification of the dissolved part of wastewater pollution (organic pollutants - COD, BOD; biogenic substances - nitrogen and phosphorus) with special microorganisms (bacteria and protozoa) or earthworms, which are called activated sludge or biofilm [12].

Ts.I. Rogovskaya in her 1967 work "Biochemical method of industrial wastewater treatment" writes about the development of industry, in particular chemical, and changes in the composition of wastewater, which led to changes in the composition and characteristics of activated sludge biocenoses [13]–[22]. It is unacceptable to discharge un-treated industrial wastewater into water bodies; it is necessary to intensify the purification process. With this information, it is possible to speed up and reduce the cost of the biological treatment process. The author also gives the maximum allowable concentrations of some compounds for adapted microflora when they enter aerobic treatment facilities operating for complete purification [23].

In 1977, in the reference manual Methods of industrial wastewater treatment, the authors (A.I. Zhukov, I.L. Mongait, I.D. Rodziller) noted that various types of aerotanks are most often used for biological treatment of large volumes of wastewater, one of the important advantages of which is the ability to effectively influence the speed and completeness of the process of biochemical oxidation occurring in them. That the processes occurring in the aerotank are controllable [20].

Doctor of Technical Sciences E.D. Gelfand in the textbook "Fundamentals of Biological Wastewater Treatment" highlights some parameters of biological treatment [21]–[26].

Tukalevsky Sergey Leonidovich in 1992 in his dissertation work "Economic and mathematical models and methods for solving special classes of nonlinear distributive problems" proposed for a discrete-continuous goal with a concave function

of the distribution problem of optimizing the water treatment process, a new two-stage approach to finding a solution was proposed, which consists in determining the lower estimate of the minimum and the optimal values of the dual estimates at the first stage and the algorithm for obtaining a feasible solution close to the optimal one at the second.

In Uzbekistan, scientists academician M. Mirsaidov, Sh.Kh. Rakhimov, I.K.Khuzhaev, E.I. Makhmudov, S.I. Khudaykulov, O. Golovatsky, A.V. Kabulov, N. U. Uteuliev, A. Zh. Seitov, D. Sh. Bazarov, E. Shermatov and others.

Purpose of the study. The aim of the study is to develop and analyze mathematical models and algorithms for solving deterministic problems of polluted water treatment, establishing a duality ratio and optimality conditions, as well as establishing a marginal ratio for these tasks and determining the optimal cost of polluted water treatment under various treatment flowsheets, development a software package for solving deterministic problems of a polluted water treatment complex [27]–[29]. Research tasks:

- Development of mathematical models that optimize costs in the treatment of polluted wastewater;
- Development of an algorithm for solving the problems of contaminated water treatment complex;
- Establishment of secondary relations and optimal conditions for the optimization of wastewater treatment;
- Development of software to solve the problem of optimization of wastewater treatment;

Object of research. The object of research is the treatment of polluted wastewater in municipal wastewater treatment plants.

Subject of research Mathematical model and numerical methods of optimizing the cost of wastewater treatment.

Research methods. The study used mathematical programming theories and methods and discrete optimization methods.

II. SETTING OF DETERMINISTIC DISTRIBUTION PROBLEM OF COMPLEX OF CONTAMINATED WATER TREATMENT

This part discusses the deterministic distribution problem of the contaminated water treatment complex.

A. Statement of the problem

Thus, the objective is to select such an acceptable combination of technologies and such a distribution of the volume of contaminated water between them so that the cost of purification is minimal and at the same time the concentration of impurities after purification at the control solution does not exceed the specified maximum permissible values. In reality, the problem of water treatment is solved for a certain water treatment complex in the presence of several spillways and the concentration of controlled impurities at the control range depends on the quality of water protection events at all spillway facilities.

Enter the following symbols:

– q_k volume of contaminated water supplied to the treatment complex on the k - volume of the spillway;

– x_{ki} amount of contaminated water to be treated on i cleaning process diagram on k - volume of the spillway. Thus, $X_k = \{x_{ki}\}, i = \overline{1, N}$, vector of the distribution of the total amount of contaminated water between treatment technologies on the k - th of the spillway;

x_{ki}^- and x_{ki}^+ - values limiting production capacity i process diagram during volume discharge cleaning q_k ;

C_{0kj}, C_{kj} - concentrations of j -type impurities before purification and maximum permissible concentrations (MPC) after purification, respectively;

P_{kij} - degree of cleaning of j -th impurity according to the i -th cleaning technology on k -volume spillway;

A_{ki}, λ_{ki} - coefficients of approximation of the function of costs for treatment of contaminated water according to the i -th process diagram on k - volume of the spillway ($A_{ki} \geq 0; 0 \leq \lambda_{ki} \leq 1, k = \overline{1, L}, i = \overline{1, N}$).

Then the mathematical model of the contaminated water treatment complex, reflecting the process of finding the best treatment method with minimal costs, has the following form:

$$F(x) = \sum_{k=1}^L \sum_{i=1}^N A_{ki} x_{ki}^{\lambda_{ki}} \rightarrow \min \quad (1)$$

$$\sum_{i=1}^N B_{kij} x_{ki} \leq C_{kj}, \quad k = \overline{1, L}; j = \overline{1, M}, \quad (2)$$

$$\sum_{i=1}^N x_{ki} = q_k, \quad k = \overline{1, L}; \quad (3)$$

$$x_{ki}^- \leq x_{ki} \leq x_{ki}^+, \quad k = \overline{1, L}; j = \overline{1, M}, \quad (4)$$

$$A_{ki} \geq 0, \quad 0 \leq \lambda_{ki} \leq 1, \quad k = \overline{1, L}, i = \overline{1, N}$$

The problem in which it is necessary to distribute wastewater to process schemes in such a way that the limitations of the balance and process type are fulfilled, and at the same time obtain the optimal value of the objective function, is called a non-linear problem of the distribution type.

It should be noted that the set mathematical problem arises not only in connection with the use of equipment, but also in many other issues. Such a model and problems on the rational use of water resources are considered in special studies by L.V. Kantorovich and V.A. Zalgaller [4], reflecting both the mathematical and purely technological aspect of the question. In addition to the distribution problem of optimal load of equipment, following the transport problem and assignment problem, one of the first detailed problems of the distribution type studied was the linear problem of optimal distribution of interchangeable resources.

B. Now let us dwell on the economic content of the model (1) – (4).

The objective function (1) reflects the total cost of the water user to carry out water treatment events. It will be appreciated that the dependencies reflecting the relationship

of cleaning implementation costs to the loading volume of cleaning technologies are concave functions. Minimizing the objective function even on the convex region presents increased complexity.

Constraints (2) to (4) specify the best solution search area. Limitation (2) means that in the control solution the concentration of j- of this type of impurity should not exceed the permissible one.

“Cleanability” factor denoted by B_{kij} , calculated as follows:

$$B_{kij} = C_{0kj}(1 - P_{kij}) \quad (5)$$

Limitation (3) requires that the entire volume of k- of this discharge on that spillway be distributed according to process diagrams.

Limitation (4) means that the amount of contaminated water to be treated shall not exceed the production capacity of the process diagrams.

III. ALGORITHM FOR SOLVING DETERMINISTIC DISTRIBUTION PROBLEM OF CONTAMINATED WATER TREATMENT

In this part we will give the algorithm of solution of the deterministic and limiting problem of the complex of contaminated water purification, reflecting the process of search for the best method of purification with minimum costs (1) - (4), proposed in work [2].

To justify the procedure proposed in work [2], write down the Lagrange function for problem (1) - (4), which has the form:

$$L(x, u) = \sum_{k=1}^L \sum_{i=1}^N A_{ki} x_{ki}^{\lambda_{ki}} + \sum_{k=1}^L \sum_{j=1}^M u_{kj} C_{kj} + \sum_{k=1}^L u_k q_k \quad (6)$$

If we know u^* - the optimal solution of the dual problem to the problem (1) - (4), then one can consider its equivalent setting:

$$L(x, u^*) \rightarrow \min_{x \in X} \quad (7)$$

where X is a set of vectors satisfying the conditions (2), (4).

From problem (7) we obtain, taking into account (7), k-independent subproblems for each catchment

$$\sum_{k=1}^L \sum_{i=1}^N A_{ki} x_{ki}^{\lambda_{ki}} - \sum_{k=1}^L \sum_{i=1}^N \left(\sum_{j=1}^M u_{kj}^* B_{kij} - u_k^* \right) x_{ki} \rightarrow \min_{x \in X} \quad (8)$$

Problem (8) is a problem of geometric programming with a concave nonlinear objective function and it can be solved by the conditional gradient method.

Thus, with the known solution u^* The problem (1) to (4) is decomposed. The same will be the case with arbitrary u , only then the solution of problem (8) obtained in this case will not be the optimal solution of problem (1) - (4).

Let us now give the algorithm for solving the deterministic distributive problem (1) - (4):

In fact, the exact optimal solution to the dual problem is unknown. However, you can build a definition procedure u^* , which has the form:

1) As an initial approximation $u^{(0)}$ we take an arbitrary number. Let us know $u^{(s)}$, $s = 0, 1, \dots$ Then we determine $x^{(s)}$, solving a problem

$$\begin{cases} \sum_{i=1}^N (A_{ki} x_{ki}^{\lambda_{ki}} - u_k^{(s)} x_{ki}) \rightarrow \min \\ x \in X \end{cases} \quad (9)$$

2) New value $u^{(s+1)}$ is defined using the formula

$$u_k^{(s+1)} = \max\{0, \rho_s (q_k - \sum_{i=1}^N x_{ki}^{(s)})\} \quad (10)$$

where ρ_s - some step factor.

3) The step factor is defined as follows:

$$\rho_s \geq 0, \quad \rho_s \rightarrow 0 \quad \text{at} \quad s \rightarrow \infty, \quad \sum_{s=0}^{\infty} \rho_s = \infty, \quad \sum_{s=0}^{\infty} \rho_s^2 < \infty \quad (11)$$

These conditions are satisfied, for example, by the following step factor value

$$\rho_s = \frac{1}{(1+s)^\alpha}; \quad \frac{1}{2} < \alpha \leq 1 \quad (12)$$

It should be noted that $q_k - \sum_{i=1}^N x_{ki}^{(s)}$ is the (10) generalized gradient of the objective function $\varphi(u) = \min_{x \in X} L(x, u)$ a dual problem. Under the assumptions made, this function will not be differentiable under those $u^{(s)}$, when at least one of the problems (9) has more than one optimal solution. The procedure (10) itself will be a general gradient descent method (8) for solving the dual problem.

IV. STOCHASTIC DISTRIBUTION PROBLEM OF COMPLEX FOR TREATMENT OF CONTAMINATED WATER

In this part, let's talk about the problem of distributing a certain amount of contaminated discharge water to the process treatment schemes. By cleaning process circuit is meant a set of some cleaning devices arranged in series. The part of the discharge allocated for cleaning according to some scheme should not go to the cleaning devices of another scheme in the future.

It should be noted that in the problem according to the available literary sources [5], dependencies reflecting the relationship of the water user's costs for the implementation of water protection events with the load volume of purification process circuits are concave functions and are well approximated by power functions with a fractional indicator. In addition, the coefficients of the objective function may depend on random factors, then we will obtain a stochastic model of the complex for treating contaminated waters.

Here is a stochastic version of the distribution problem of the contaminated water treatment complex:

The objective is to select such an acceptable combination of technologies and such a distribution of the volume of contaminated water between them so that the media total purification costs are minimal and the concentration of impurities after purification at the control solution does not exceed the specified maximum permissible values. In reality, the problem of water treatment is solved for a certain water treatment complex in the presence of several spillways and the concentration of controlled impurities at the control range depends on the quality of water protection events at all spillway facilities.

Enter the following symbols:

$A_{ki}(\theta)$, λ_{ki} - coefficients of approximation of the function of costs for treatment of contaminated water according to the i -th process diagram on k -th of the spillway, θ - random number ($A_{ki}(\theta) \geq 0$; $0 \leq \lambda_{ki} \leq 1$, $k = \overline{1, L}$, $i = \overline{1, N}$).

Then the stochastic distribution model, reflecting the process of finding the best cleaning method with average total minimum costs, has the following form:

$$F(x, \theta) = \sum_{k=1}^L \sum_{i=1}^N MA_{ki}(\theta)x_{ki}^{\lambda_{ki}} \rightarrow \min \quad (13)$$

$$\sum_{i=1}^N B_{kij}x_{ki} \leq C_{kj}, \quad k = \overline{1, L}; \quad j = \overline{1, M}, \quad (14)$$

$$\sum_{i=1}^N x_{ki} = q_k, \quad k = \overline{1, L}; \quad (15)$$

$$\begin{aligned} x_{ki}^- \leq x_{ki} \leq x_{ki}^+, \quad k = \overline{1, L}; \quad j = \overline{1, M}, \\ A_{ki} \geq 0, \quad 0 \leq \lambda_{ki} \leq 1, \quad k = \overline{1, L}, i = \overline{1, N} \end{aligned} \quad (16)$$

Problems (13) - (16) are problems of stochastic convex programming.

To solve the problem (13) - (16), the algorithm proposed in operation [2] is used.

This algorithm is applicable to solve the problem (13) - (16). To do this, we construct the Lagrange function for problem (11) - (14)

If we know u^* - the optimal solution of the dual problem to the problem (13) - (16), then one can consider its equivalent setting:

$$\begin{cases} L(x, u^*, \theta) \rightarrow \min \\ x \in X \end{cases} \quad (17)$$

where X is a set of vectors satisfying the conditions (14), (16).

From problem (17) we obtain, taking into account (??), k -independent subproblems for each catchment

$$\begin{cases} \sum_{i=1}^N (MA_{ki}(\theta)x_{ki}^{\lambda_{ki}} - u_k^*x_{ki}) \rightarrow \min \\ x \in X \end{cases} \quad (18)$$

We do not know the exact solution to the dual problem. But instead, you can use the approximation obtained, say, by the method of stochastic generalized gradient descent.

Based on this, we give the following algorithm for solving the stochastic problem (13) - (16) and the dual to it:

1) As an initial approximation $u^{(0)}$ we take an arbitrary number. Let us know $u^{(s)}$, $s = 0, 1, \dots$. To calculate $u^{(s+1)}$ perform the following actions:

2) We solve the following problem using one of the methods of stochastic programming [?].

$$\sum_{i=1}^N (MA_{ki}(\theta)x_{ki}^{\lambda_{ki}} - u_k^*x_{ki}) \rightarrow \min_{x \in X} \quad (19)$$

solution of the problem we denote $x^{(s)}$;

3) Find the gradient of the Lagrange function (16) by the dual variables u ;

4) we calculate $u^{(s+1)}$ by formula:

$$u_k^{(s+1)} = \max\{0, \rho_s(q_k - \sum_{i=1}^N x_{ki}^{(s)})\} \quad (20)$$

where ρ_s - some step factor. Selection of step multiplier is determined by conditions (13) - (14).

It should be noted that $q_k - \sum_{i=1}^N x_{ki}^{(s)}$ is the (20) generalized gradient of the objective function $\varphi(u) = \min_{x \in X} L(x, u, \theta)$ a dual problem. Under the assumptions made, this function will be undifferentiated under those $u^{(s)}$, when at least one of the problems (19) has more than one optimal solution. The procedure (20) itself is a method of stochastic quasi-gradients [?] for solving a dual problem.

V. NUMERICAL EXPERIMENT OF ENVIRONMENTAL-ECONOMIC MODELS OF CONTAMINATED WATER TREATMENT COMPLEX

Numerical experiment of the calculation of MPD (maximum permissible discards) substances for a water stream and a water body. We consider the river basin section (Figure 1), which includes 3 water quality control sections, 2 wastewater discharges, 2 water intakes and 1 reservoir. The volume of filling the reservoir is $1km^3$. Water quality is estimated by 6 indicators biochemical oxygen consumption (BOD), BOD_{full} (complete oxidation) or BOD_{20} (almost complete oxidation is achieved within 20 days), ammonium nitrogen, nitrite nitrogen, nitrate nitrogen, dissolved oxygen and petroleum products. Initial data for MPD calculation are given in table 1-2.

Let us introduce the following notation. Complete biological treatment - 101, complete biological purification with simultaneous precipitation to improve the purification of phosphorus - 102, complete biological purification with nitrification-denitrification to improve purification by impurities of the nitrogen group - 103.

Realization. The operation diagram of the treatment facilities is as follows: from the receiving chamber, waste water enters the pond as a settling tank for deposition of suspended

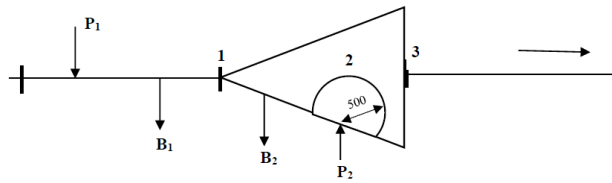


Fig. 1. Linear diagram of the river basin section: 1, 2, 3 check doors; p1, p2 wastewater discharges; 1, 2 water intakes.

substances and processing of sediment. To increase stability and reliability of the effect of clarification, the sump is divided into two consecutive stages, the interruption of the effluents of each of them is 3 or 2 days, respectively.

TABLE I
CHARACTERISTICS OF THE WATER BODY

	Activities	Reset conditions	
		mg/liter	MPD g/hour
1	suspended substances	140.0	53050.4
2	BOD ₅	5.0	1894.3
3	mineral composition	2000.0	757720.0
4	Chlorides	350.0	132601.0
5	Sulfates	500.0	189430.0
6	Fats	5.5	2083.73
7	Nitrate nitrogen	—	—
8	Nitride nitrogen	—	—
9	Ammonium nitrogen	—	—

To improve reliability of compliance with KMK requirement 2.04.02-97 (Water supply External networks and structures), it is also envisaged to divide all earthen capacitance structures into two parallel operating sections, each of which is designed to pass 50% of the design flow in normal operation mode and 100% in forced operation.

Lean liquid clarified during 5 days is divided into two streams in accordance with 30% or 70%. A smaller flow is passed through the cultivator pond for 6 days for the development of algal biomass and the accumulation of the sum of oxidants, after which it is mixed with the main flow. To accelerate cultivation, 10% of the flow is supplied to the cultivator - starter pond, from which it enters the cultivator after 4.5 days.

The resulting mixture continuously enters the biocogulator pond, where during 2 days a complex biocogulation process occurs, as a result of which the content of suspended and dissolved contaminants is reduced.

Next to the pond biocogulator treatment stage is aerobic treatment stage under flow conditions in the pond insulator, the main process of which is bacterial oxidation of dissolved organic and mineral components of waste water.

In the flow pond waste liquid is 2 days, BOC (biochemical oxygen consumption) brought to 15mg/liter. After biological treatment, water enters post-treatment ponds with the highest water vegetation, where the quality of the treated effluents

is brought to the requirements that meet the rules of the Protection of Surface Waters from Pollution by Waste Water.

Approved MPD and waste water composition (discharge of substances not specified below is prohibited) according to the table.

Water protection measures according to formulae (1) to (4) are described as follows:

$$f_1 = 9,67 \cdot 365 \cdot (x_{10} + 602,9 \cdot x_{11} + 741,5 \cdot x_{12} + 1083,6 \cdot x_{13}) = 3530 \cdot x_{10} + 2128237 \cdot x_{11} + 2617495 \cdot x_{12} + 3825108 \cdot x_{13};$$

$$C_{11} = 193,7x_{10} + 103,1 \cdot x_{11} + 34,4 \cdot x_{12} + 20,6 \cdot x_{13};$$

$$C_{12} = 7,6 \cdot x_{10} + 4 \cdot x_{11} + 1,4 \cdot x_{12} + 0,8 \cdot x_{13};$$

.....

$$x_{11} + x_{12} + x_{13} + x_{10} = 1;$$

Results of solution of formed problem of calculation of MPD of substances and optimal water protection measures for their achievement obtained with the help of OCHISTKA software complex are given in Table 2.

TABLE II
OPTIMAL WATER CONSERVATION MEASURES TO ACHIEVE MPD

Dumping	Event cipher without cleaning	Waste water flow rate			Reported costs, thousand sum/year
		thousand m ³ /day	in %	x _{ii} - decisions	
	202	—	—	—	—
	206	8.9177	92.22	0.9222	2413792

TABLE III
OPTIMAL WATER CONSERVATION MEASURES TO ACHIEVE MPD

No.	Components	Without cleaning	Complex cleaning by permissible concentration
1	suspended substances	193.7	140.0
2	BOD ₅	7.6	5.0
3	mineral composition	2769.0	2000.0
4	chlorides	712.5	350.0
5	sulfates	404.4	500.0
6	fats	1	5.5
7	nitrate nitrogen	47.1	20
8	nitride nitroge	0.08	0.04
9	ammonium nitrogen	0.91	0.4
10	costs, thousand of units per day	0	6613.7

It should be understood that in order to achieve acceptable pollutant concentrations, it is necessary to clean mainly the process some allowable values are not achieved, for example, mineral composition, chlorides and nitrogen. Comparison of cleaning costs by different process diagrams and complex cleaning is given in Table 3.

VI. CONCLUSION

In the proposed work, we considered the deterministic and stochastic distribution problems of the contaminated water treatment complex, consisting in the distribution of a certain volume of contaminated discharge water through treatment schemes, and we also proposed algorithms for solving these problems.

We also presented the results of a numerical experiment with the task of choosing such an acceptable combination of technologies and such a distribution of the volume of polluted water between them that the average total cost of cleaning is minimal and at the same time the concentrations of impurities after cleaning at the control site do not exceed the specified maximum permissible values. All possible types of visualization of the results of a computer experiment, as a rule, are reduced to the representation of tabular dependencies of a function of one variable, a function of two variables, as well as lines of the level of fields of vector functions.

The results of the study in the form of software, algorithms, and methods for the purification of polluted waters have been implemented: a software tool for calculations on deterministic and stochastic ecological and eco-nomic models of the complex for the purification of polluted waters has been introduced in the purification plant of the city of Nukus. The implementation made it possible to obtain and determine the optimal costs for the treatment of polluted waters with various technological treatment schemes; An improved mathematical model of the process of the polluted water treatment complex was used in the purification plant of the city of Nukus. The results of scientific research made it possible to save costs several times less than the previous one (by 15%).

REFERENCES

[1] A. P. Karmanov, I. N. Polina. Technology of wastewater treatment. - Syktyvkar: SLI, 2015. - 207 p.
 [2] S. V. Yakovlev. Rational water use for cities and industrial enterprises // Water supply and sanitary engineering. - 1994. - No. 7. - p. 3-6.
 [3] Yu. M. Yermol'ev. Methods of stochastic programming. - M.: Nauka, 1976. 176 p.
 [4] Yu. V. Voronov, S. V. Yakovlev. Water disposal and wastewater treatment. - Moscow: MGSU Publishing House, 2006. -702 p.
 [5] S. V. Yakovlev, Yu. V. Voronov. Water disposal and wastewater treatment. 4th ed., reprint. - M. Ed. of the Association of Construction Universities, 2006. -704 p.
 [6] A. T. Adugna, H.A. Andrianisa, Y.Konate, A.Ndiaye and A.H. Maiga. Performance comparison of sand and fine sawdust vermifilters in treating concentrated grey water for urban poor // Environmental Technology, 2015.
 [7] A. Kabulov, I. Saymanov. Application of IoT technology in ecology (on the example of the Aral Sea region). International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
 [8] A. Kabulov, I. Saymanov, I. Yarashov, F. Muxammadiev. Algorithmic method of security of the Internet of Things based on steganographic coding. 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
 [9] A. Kabulov, M. Berdimurodov. Parametric Algorithm for Searching the Minimum Lower Unity of Monotone Boolean Functions in the Process Synthesis of Control Automates. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-5.

[10] A. Kabulov, M. Berdimurodov. Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal). International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
 [11] A. Kabulov, I. Saymanov, M. Berdimurodov. Minimum logical representation of microcommands of cryptographic algorithms (AES). International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
 [12] A. Kabulov, E. Urunboev, I. Saymanov. Object recognition method based on logical correcting functions. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2020, pp. 1-4.
 [13] A. Kabulov, A. Babadzhanov, I. Saymanov. Completeness of the linear closure of the voting model. AIP Conference Proceedings, 2022 (accepted).
 [14] A. Kabulov, A. Babadzhanov, I. Saymanov. Correct models of families of algorithms for calculating estimates. AIP Conference Proceedings, 2022 (accepted).
 [15] A. Kabulov, I. Normatov, S. Boltaev, I. Saymanov. Logic method of classification of objects with non-joining classes. Advances in Mathematics: Scientific Journal, 2020, 9(10), p. 8635-8646.
 [16] A. Kabulov, I. Kalandarov, I. Saymanov. Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules. Smart transport conference 2022 Conference, pp. 1-11.
 [17] A. Kabulov. Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions. Cybernetics, 1979, 15(3), p. 305-311.
 [18] A. Kabulov, G. Losef. Local algorithms simplifying the disjunctive normal forms of Boolean functions. USSR Computational Mathematics and Mathematical Physics, 1978, 18(3), p. 201-207.
 [19] A. Kabulov. Local algorithms on yablonskii schemes. USSR Computational Mathematics and Mathematical Physics, 1977, 17(1), p. 210-220.
 [20] Anil K Dwivedi. Researches in water pollution: a review // International Research Journal of Natural and Applied Sciences, 2017, pp. 118-142.
 [21] Sunita Narain. Ganga-the River, its pollution and what can we do to clean it // Centre for Science and Environment, 2014, pp. 1-32.
 [22] Rajat Kaushik. Mathematical modelling on water pollution and self-purification of river Ganges // Pelagia Research Library.
 [23] S. Siddiqua, A. Chaturvedi, R. Gupta. A Review-Mathematical Modelling on Water Pollution and Its Effects on Aquatic Species // Advances in Applied Mathematics Conference, 2020, pp. 835-842.
 [24] H. Khujamatov, I. Siddikov, E. Reypnazarov, D. Khasanov. Research of Probability-Time Characteristics of the Wireless Sensor Networks for Remote Monitoring Systems // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
 [25] I. Siddikov, D. Khasanov, H. Khujamatov, E. Reypnazarov. Communication Architecture of Solar Energy Monitoring Systems for Telecommunication Objects // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
 [26] I. Siddikov, K. Khujamatov, E. Reypnazarov, D. Khasanov. CRN and 5G based IoT: Applications, Challenges and Opportunities // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
 [27] K. Khujamatov, A. Lazarev, N. Akhmedov, E. Reypnazarov, A. Bekturdiyev. Methods for Automatic Identification of Vehicles in the its System // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
 [28] S. Tanwar, H. Khujamatov, B. Turumbetov, E. Reypnazarov, Z. Allamuratova. Designing and Calculating Bandwidth of the LTE Network for Rural Areas // International Journal on Advanced Science, Engineering and Information Technology, 2022, 12(2), pp. 437-445.
 [29] H. Zaynidinov, D. Singh, S. Makhmudjanov, I. Yusupov. Methods for Determining the Optimal Sampling Step of Signals in the Process of Device and Computer Integration // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 471-482.

Predictive Maintenance and Condition Monitoring in Machine Tools: An IoT Approach

Brett Sicard

Department of Mechanical Engineering
 McMaster University
 Hamilton, Ontario, Canada
 sicardb@mcmaster.ca

Naseem Alsadi

Department of Mechanical Engineering
 McMaster University
 Hamilton, Ontario, Canada
 alsadin@mcmaster.ca

Petros Spachos

Department of Mechanical Engineering
 University of Guelph
 Guelph, Ontario, Canada
 petros@uoguelph.ca

Youssef Ziada

Global Manufacturing Engineering
 Ford Motor Company
 Livonia, Michigan, USA
 yziada@ford.com

S. Andrew Gadsden

Department of Mechanical Engineering
 McMaster University
 Hamilton, Ontario, Canada
 gadsden@mcmaster.ca

Abstract—To maximize efficiency, quality of products, and profits, it is important to maintain machine tools to reduce downtime and maximize output. Predictive maintenance is the most efficient method of condition monitoring and maintenance. An Internet of Things approach can help implement an autonomous predictive CM system in manufacturing facilities. The critical parameters of sensor selection, communication, and data analysis have been examined. The components that make up an effective IoT CM system have been discussed and analyzed. An IoT approach has been shown to eliminate the disadvantages of traditional manual CM approaches.

Index Terms—Feed drives, Machine tools, Condition monitoring, IoT, Internet of Things, Industry 4.0.

I. INTRODUCTION

The manufacturing world has entered the fourth industrial revolution, often referred to as *Industry 4.0*. The fourth industrial revolution builds upon the automation and digitization of the third industrial revolution. It includes the implementation of the Internet of things (IoT), machine to machine communication, and improved communication and condition monitoring (CM).

The manufacturing sector is one of the most competitive sectors in developed nations. They must compete not only with other firms in developed nations but also with developing nations, which have the advantage of lower wages and less strict labour and environmental laws. To remain competitive, firms must embrace industry 4.0. Embracing Industry 4.0 can help to reduce costs, reduce errors, increase efficiency, and improve throughput. These advantages can help firms propel themselves above the local competition and compete with the prices of firms in developing nations. Advanced manufacturing is very capital dependent. Expensive machine tools such as CNC lathes, mills, laser cutters, and grinders are staples of modern manufacturing facilities. It is of utmost importance to maximize the output of these machines. Substantial costs are accrued if machines are not working at high levels of output.

It is in the best interest of these manufacturing facilities to maintain these machines adequately. Machines are maintained through maintenance and CM.

Maintenance and machine condition has several associated costs: Expected downtime, unexpected downtime, quality, and replacement parts. The cost of expected downtime is the lost production due to the machine being down for scheduled maintenance. Unexpected downtime is the cost of lost production and damage due to sudden failures and crashes. The cost of quality is the cost of defects that arise due to poor machine conditions, such as not meeting dimensional tolerance. Replacement part cost is the cost associated with replacing a part before it has reached the end of its useful life.

Predictive and prognostic maintenance is the latest evolution of maintenance paradigms. The first level of maintenance paradigms is reactive maintenance. Reactive maintenance is simply fixing machines as they break. This is the least effective form of maintenance. Reactive maintenance increases unexpected downtime as unexpected failures arising from neglect force the machines down. The next level of maintenance is planned maintenance, where maintenance occurs at set intervals. This is a better method compared to reactive maintenance but is still not ideal. Doing maintenance at set intervals increases the cost of expected downtimes and can mean replacing parts that are still far away from the end of their life. Predictive maintenance involves analyzing data from the machine and predicting when it requires maintenance. This paradigm can reduce expected, and unexpected downtime costs as machines are serviced only when necessary and before critical failures occur. Cost of quality and replacement part costs can be reduced as well as parts are used for the majority of their useful life and not past the point where the cost of quality begins to rise. A summary of the relative costs of each paradigm is seen in Table I below.

Predictive maintenance requires a great deal of data to

TABLE I
COSTS OF MAINTENANCE

Paradigm	Associated cost			
	Expected downtime	Unexpected downtime	Quality	Replacement parts.
Reactive	Low	High	High	Moderate
Scheduled	High	Low	Low	High
Predictive	Moderate	Low	Low	Low

analyze. Traditional CM is labour intensive and prone to human error. It often requires a worker to inspect and test machines over time individually. This approach is too expensive to inspect at a high enough frequency, and inspection is done too infrequently, therein faults and wear are detected too late. IoT can facilitate autonomous CM. Sensors can be installed on machines to collect a large stream of data in real-time. Data collected from the machines can be analyzed to determine the machine’s condition. This information can be used to take corrective action if required. Autonomous CM will maximize machine performance and up-time, which will, in turn, maximize manufacturing efficiency.

Machine tools have several components whose condition is worth monitoring. The primary investigation of this paper will be on machine tools with ball screw feed drives. Machine tools feed drives and spindles are typically run by AC servo motors. Ball screw feed drives are the most common feed drive for machine tools. The main components of a ball screw system can be seen below in Fig.1. Ball screws are normally supported on either end by fixed and free ball bearings. A table is connected to the ball nut, and this table is a set of linear guides that support it to slide back and forth with minimal friction.

This paper will cover the following: Section II will cover the current literature on the subject of CM. Section III will cover the sensors used in the machine tool CM system. Section IV will cover the communication within the system. Section V will cover the data analysis used for CM. Finally Section VI will cover the conclusion and future work.

II. STATE OF CONDITION MONITORING IN THE LITERATURE

Condition monitoring has been a field that has been studied extensively in the literature. A paper by Martin [2] covered machine tool CM and fault detection technology. Numerous

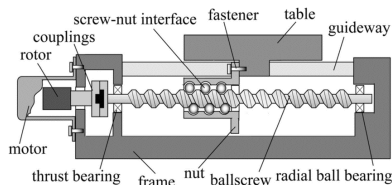


Fig. 1. Ball Screw System [1]

CM analysis methods for bearings and other moving parts exist in the literature. A survey by Zhou, Habetler, and Harley [3] outlines many common methods for CM. Many of these methods will be outlined in section V. A journal by Tandon, Yadava, and Ramakrishna [4] compared four methods for CM to determine which was most effective at detecting a fault in a bearing. They examined velocity vibration, motor stator current, acoustic emission and shock pulse method. They found that acoustic emission and shock pulse methods were the most effective method in detecting a bearing fault. In another study by Tandon [5] several vibration analysis techniques were compared to determine which was the most effective at detecting a bearing failure. It was found that peak vibration and vibration power were good predictors of wear. Temperature is another method for CM. A study by Touret et al. [6] reviews sensors and methods involving temperature for detecting faults. An IoT system for CM of CNC machine was proposed in a paper by Al-Naggar et al. [7].

There are several studies to detect wear in machine tools. In a study by Pichler, Klinglmayr, and Pichler-Scheder [8] they examined methods such as vibration, temperature and current to detect wear in a ball screw system. In one study by Verl et al. [9] they used information from the CNC controller to measure positioning error, repeatability and reversal error to determine if the system had encountered a fault. In a study by Huang, Tan, and Lee [10] they used Kalman filtering to detect mechanical and sensor failure in a ball screw system. Methods for detecting preload loss and determining levels of preload in a ball screw have been covered by several studies. In a study by Frey, Walther, and Verl [11] they inserted a force sensor between nuts in a system preloaded with a double nut setup. By doing this, they were able to determine the force of preload and measure changes over time and over the stroke of the screw. Studies by Feng and Pan [12] [13], Chang et al. [14], Tsai, Cheng, and Hwang [15], and Nguyen, Ro, and Park [16] used ball pass frequency, temperature, motor current and vibration analysis to detect preload loss.

IoT approaches to CM exist in the literature. Machine learning has become a popular data analysis method over the past decade. In a study by Ayvaz and Alpay [17] they implemented an IoT system and used machine learning to predict the time until failure for machine equipment. In another study by Kanawaday and Sane [18] they could predict a bad production cycle with a high degree of accuracy using machine learning. Both edge computing [19] and cloud computing [20] approaches to IoT CM have been proposed in two studies.

III. SENSORS AND DATA COLLECTION

Sensors on a machine tool can broadly be categorized into *integrated sensors* and *external sensors*. Integrated sensors are sensors that are built into the CNC controller or are required for normal operation. External sensors are sensors not normally included with the CNC controller or required for normal operation. External sensors can collect information from many different components such as the bearings, spindle, motor, and linear guides.

A. Integrated Sensors

Integrated sensors are the sensors and information that are required for the normal operation of the machine tool. These typically include both linear and rotary encoders for measuring the position of the worktable in each axis, as well as the torque signal which is sent to the motor from the CNC controller.

1) *Encoders*: Position measurements are performed using linear and rotary encoders. Encoders can either be incremental or absolute encoders. Absolute encoders do not require a reference position, while incremental encoders do. As a result, incremental encoders need to have zero points established after being powered down. Glass scale encoders are the most common type of linear encoder used, as seen in Fig. 2.

Most servo motors are equipped with optical rotary encoders to measure the rotational position of the output shaft. The derivation of readings from encoders can be used to determine velocity and acceleration. Disagreement between the two encoders can be used to detect faults. If there is disagreement during changing of directions, it could mean there is substantial backlash in the system. Constant disagreement between the two encoders could represent miscalibration or misalignment of the system.

2) *Measuring Torque and Current*: The torque input signal of the system and motor current information can be retrieved from the CNC controller. These can be used to determine the torque applied to the system. If torque is greater or less than expected, it could represent a fault [16]. Analyzing the torque signal behaviour [22] or motor current behaviour [23] can also provide useful analysis.

B. Temperature Sensors

Temperature can be a useful characteristic to measure. Heat is normally generated as a result of friction. Excess heat can result in increased wear in both mechanical and electronic components. It can indicate faults in components such as bearings and motors. When selecting a sensor, it is important to consider what temperature range will be examined, the sensitivity, response time, and linearity of measurement. Temperature sensors are either contact or non-contact. A summary of popular temperature sensors can be seen below in Table II.

A comparison of the performance of popular contact temperature sensors can be seen below in Table III. Contact temperature measurement sensors are ideal if it is possible to make a solid physical connection without interfering with the operation of the machine.

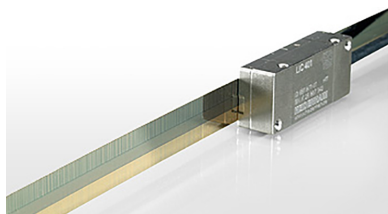


Fig. 2. Glass scale linear encoder [21]

TABLE II
COMMON TEMPERATURE SENSORS

Contact	Non-Contact
Thermistor	Radiation thermometer
Resistance temperature detectors	Optical pyrometer
Thermocouple	Fiber optic sensors
Semiconductor thermometer	

TABLE III
COMMON CONTACT TEMPERATURE SENSORS

Sensor type	Thermistor	RTD	Thermocouples
Range (C°)	-55 to 125	-200 to 850	-600 to 2000
Sensitivity	High	Low	Medium
Response time	Fast	Slow	Fast to Slow
Linearity	Exponential	Fairly linear	Fairly linear

Thermocouples are one of the most common contact temperature sensors. They can sense a large range of temperatures, have linear sensing of temperatures, and can have fast response times. They do not require an external power supply. Contact sensors may be useful to measure non-moving objects such as the bearings and motors. The expected temperatures for these in normal operating conditions are around 50° to 80° Celsius. IoT specific thermocouple exist and have built-in communication capabilities.

Non-contact sensors may be more useful in situations where contact with the object is not feasible or easy, the object is moving, or there are multiple objects to be measured at once. The most common of these sensors are radiation thermometers. These sensors measure the infrared radiation emitted from objects. They are often used for measuring the temperature of electronic components, and bearings [24].

C. Vibration Sensors

Vibration sensors are sensors that measure the mechanical vibrations of components. Papers by Goyal and Pabla [25], and Xu et al. [26] provide an overview of sensors used to detect vibrations. Vibration sensors can mostly be categorized into displacement, velocity, and acceleration sensors. An overview of these sensors can be seen below in Table IV. These sensors can also be categorized as contact and non-contact.

Certain characteristics must be considered when selecting a vibration sensor [27]. Some of these characteristics include Level of vibration, the frequency range of interest, environmental conditions, size constraints, contact type, and linearity of measurement.

Anticipating the expected sensitivity and frequency range of the vibrations is important. With high vibration amplitudes, low sensitivity sensors are preferable. The expected frequency of vibrations must reside within the range of the sensor. Different frequencies can represent different faults, as discussed in section V-A. Environmental conditions such as humidity, extreme

TABLE IV
COMMON VIBRATION SENSORS

Type of sensor	Common sensors	Frequency range
Displacement	Eddy current	Low frequency
	Inductive proximity	
Velocity	Coil and magnet	Low to mid
Acceleration	Piezoelectric accelerometer	All frequencies
	Capacitive MEMS	

temperatures, hazardous chemicals, and corrosion must be considered. Machine tools can expect a great deal of exposure to metal chips and lubrication fluid. Shielding the sensor from these is important to ensure accurate measurement. Sensors must fit into the machine tool without interfering with moving parts. For non-moving components such as bearings or motors, wired contact sensors are adequate. For moving components such as the cutting tool spindle or ball nut, it may be necessary to use a wireless and/or non-contact sensor. Otherwise, special considerations may be needed for cable management so that interference does not occur. Linearity of sensor measurements makes vibration analysis easier.

The piezoelectric sensor and the capacitive micro-electromechanical system (MEMS) are two very common sensors used for vibration sensing. They are useful for a wide range of frequencies and have low noise characteristics. They also boast high linearity over a large range. MEMS sensors are usually incredibly small. A study compared the performance of each of these two sensors and found that they had similar performance in metrics such as maximum load before failure, peak amplitude and linearity. [28]. A 3 axis IoT vibration sensor can be seen in Fig. 3.

D. Other Useful Sensors

Other than vibration and temperature sensors, there are a few other external sensors which can provide useful data. Oil quality sensors are useful devices for analyzing the quality of lubricants. Sensors are available which can provide information about the oil such as pH, temperature, capacitance [30].

Sound pressure sensors can be used to measure the sound energy emitted from moving parts. Components such as bearings will emit more noise as they wear or in the cases of



Fig. 3. Wireless 3 axis vibration sensor [29]

misalignment. One study used sound to detect faults in machine tool carbide inserts during facing operations [31]. They are not often used for CM because there is a great deal of background noise during the use of the machine.

Force sensors are useful for measuring process forces such as cutting or drilling as well as component forces such as the preload of the ball screw [11]. Common force sensors include load cells, strain gauges, and force-sensing resistors.

Cameras are useful for collecting visual data such as wear on the components like the ball screw. Many low cost cameras for IoT exist, such as the camera module available for the Raspberry Pi as seen in Fig.4.

IV. COMMUNICATION AND NETWORKING

Two key levels of communication exist in a CM system: the machine sensing level and the decision making factory level. A system overview which shows the interaction between the nodes can be seen in Fig. 5 below. The machine-level consists of individual machine tools with CNC controllers and external sensors. The factory level consists of all the machines, a central data analysis node, data storage, and nodes for each department. This is a similar framework as proposed in papers by Yaseen, Swathi, and Kumar [33], and Takemura et al. [34].

Four communication technologies are considered for IoT CM: Wired, Bluetooth, Wi-Fi, and Zigbee. An overview is seen below in Table. V. Zigbee is a popular communication technology used in literature. In one study by Lee et al. [35] they prototyped a smart factory to simulate measuring plastic extrusions. They used Zigbee with star topology for their communication. In another study, Zigbee was used for CM of motors [36]. Bluetooth technology was used in a study for the CM of welding stations [37] and a study on an IoT monitored factory [38]. In one study, the performance of Wi-Fi 2.4 GHz was examined for use in an IoT factory [39].

A. The Machine Level Communication

The machine level is the interaction between the sensors, CNC controller and machine computing and communication unit. This level represents a single machine. Table VI outlines



Fig. 4. Raspberry Pi camera [32]

the key considerations when selecting communication technology. The subsections below cover communication technology recommendations based on each design factor. Some computation occurs on this level for applications requiring low latency or immediate corrective action.

1) *Range*: Machine tools are typically never larger than a few meters. Sensors in the system can either be wired or wireless, depending on the sensor and application. If a sensor can be cannot be wired, Bluetooth or Zigbee wireless connection makes the most sense.

2) *Latency*: Certain analysis and control methods such as Kalman filtering may require low latency. Sensors for methods requiring very low latency should use a wired connection. For analysis methods with more lenient latency requirements such as temperature analysis, Bluetooth or Zigbee would be adequate.

3) *Throughput*: Certain types of sensors, such as cameras, will require a large throughput; these sensors should be wired if possible. Sensors with lower throughput, such as temperature sensors which may only pass on a single value every few seconds, can use either Zigbee or Bluetooth.

4) *Scalability*: Sensors may need to be added or removed. If a large number of sensors are required, issues may arise with the number of physical connections for wired sensors. Bluetooth has a maximum of 7 devices connected to a local network at once. There is not expected to be an excessive amount of wireless sensors, so Bluetooth is as adequate as Zigbee.

5) *Topology*: The topology of this level is a star. Communication occurs between the sensors and the central computing and communication node and between the CNC controller and the central computing and communication node. Communication between sensors or communication of external sensors

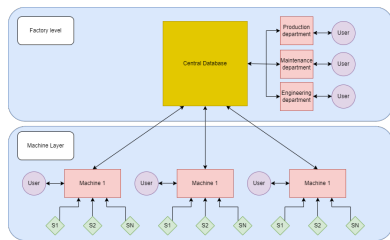


Fig. 5. Overview of the levels of communication

TABLE V
COMMUNICATION TECHNOLOGY COMPARISON

	Wired	Bluetooth	Wi-Fi	Zigbee
Typical range	5 m	10 m	100 m	50 m
Latency	Lowest	Moderate	Low	Moderate
Throughput	High	Low	Moderate	Low
Scalability	Difficult	Easy	Easy	Easy
Power	-	Moderate	High	low
Topology	Star	Star	Star	Mesh or Star

TABLE VI
COMMUNICATION CONSIDERATIONS AT MACHINE LEVEL

Range	A few meters maximum
Latency	Low latency required for some application
Throughput	Low throughput for most sensors
Scalability	Ability to add additional sensors
Topology	Star topology

with the CNC controller is unnecessary. Bluetooth or Zigbee could be used here.

6) *Optimal Communication Technology*: When possible, a wired connection is preferred. Wired connections have the lowest latency, do not require a battery, have the highest throughput, and have adequate scalability for up to a dozen or so sensors. For sensors that must be wireless, Bluetooth technology is the ideal choice. Bluetooth’s short range is fine for the expected range of communication, has low enough latency, moderate throughput and good scalability up to about seven devices.

B. The Factory Layer

The factory level is the interaction between the individual machine tools, a central computation and data storage server, and several relevant departments. Table VII outlines the key considerations when selecting communication technology. The subsections below cover communication technology recommendations based on each design factor. Most serious computation, data analysis, and data storage occur at this level.

1) *Range*: Communication on this level would be over the entire factory. Typically factories would be about 100 m to 1000 m across. Nodes at this level are all stationary, so wired communication is an adequate solution. If wireless is preferred, 2.4 GHz Wi-Fi would be adequate for these ranges if there are receivers throughout the factory, or Zigbee could be used, and communication could be routed between machines.

2) *Latency*: Most communication on this level does not require low latency. If wireless is preferred, Wi-Fi 2.4 GHz or Zigbee would be more than adequate.

3) *Throughput*: Most communication at this level has high throughput, typically on the scale of KB or MB. For this consideration, a wired or Wi-Fi wireless approach is favourable

4) *Scalability*: It may be necessary to add additional machines to the network. These machines will need power infrastructure so connecting them via a wired connection is not very

TABLE VII
COMMUNICATION CONSIDERATIONS AT FACTORY LEVEL

Range	Up to a few hundred meters
Latency	Low latency not required for most applications
Throughput	High throughput for some applications
Scalability	Ability to add additional machines
Topology	Star topology

cumbersome. Adding machines via Wi-Fi or Zigbee is very easy, and a large number of devices can be easily connected this way.

5) *Topology*: The topology of this level is a star. Communication occurs between the machines and a central data processing server and the server and relevant departments. This communication resembles a star topology. The larger range of communication could mean that a cluster topology (Fig. 6.) may be preferable in some situations.

6) *Optimal Communication Technology*: In many situations, wired connections would be adequate given the static position of each node in this level of communication. If a wireless approach is preferred, Wi-Fi is the preferred technology. The high levels of throughput necessitate communication technology that can transfer large quantities of data; thus, Zigbee is not as viable.

C. Human Interaction in Each Level

There must be a method of interaction between humans and the IoT CM system. At both levels, there will be human interaction, at the individual machine level and at the factory level.

1) *Interaction with the Machine Level*: Operators and maintenance staff need access to information about each machine. The machine interface could provide information to operators, such as if all the sensors are functional, if wireless sensors need to be recharged, and if the machine is still in operating condition. For individuals who are performing maintenance, the machine could tell them the specific issues that are occurring as well as components to inspect or replace. A dashboard similar to the one seen below in Fig. 7 would be available to workers to gather information about the condition of the machine quickly.

2) *Interaction with the Factory level*: Many different users from different departments can interact with the IoT system at the factory level. The maintenance department can access

information about the health of each machine and get alerts when there are issues, faults, or crashes. Production can use information about what machines are available as well as information about the expected availability of each machine to make production decisions. Engineering can examine data from the machines to make decisions about future machine purchasing and program creation. A dashboard should be available for each department which provides them with relevant information for decision making, such as in Fig. 8.

V. ANALYSIS METHODS

Data collected from sensors must be analyzed. This data analysis can be used to determine the condition of the machine tool. Many different types of analysis can be used to determine the condition of the machine. The following methods will be discussed: Vibration analysis, temperature analysis, Kalman filtering, lubrication analysis, machine learning, and statistical process control.

One thing that must be considered when doing analysis is where the analysis will be computed. Some methods require low latency and immediate action. These methods are best done at the machine level. Methods requiring a great deal of data or high levels of complexity should be done on the server level of the factory.

A. Vibration Analysis

Vibration analysis is one of the most common methods of CM available. Vibration measurements are used to detect many faults such as misalignment, imbalance, wear, and looseness [43]. Several factors of vibration can be considered, such as Vibration peak amplitude, velocity, and acceleration. Analyzing the different frequencies of vibration can be useful. Using the Fast-Fourier transform, you can transform a vibration from the time domain to the frequency domain. In the frequency domain, different frequency components of vibration can be analyzed. Different faults manifest at different frequencies. Lower frequency vibration analysis includes imbalance, misalignment and looseness. Higher frequency vibration analysis includes bearing faults and gear faults, among others.

Vibration analysis would primarily be done at the factory level of computing, where trends of vibration could be analyzed. Other vibration analysis methods could be done at the machine level, such as alerting the user if vibrations exceed a certain amplitude.

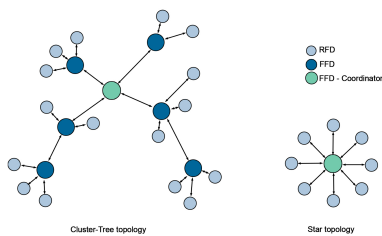


Fig. 6. Cluster topology [40]



Fig. 7. Machine monitoring dashboard [41]



Fig. 8. Factory level monitoring dashboard [42]

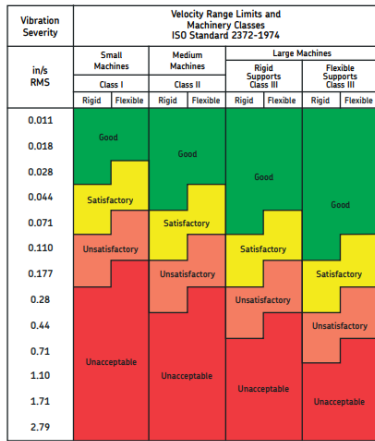


Fig. 9. ISO 2372-1974 Standard for machine vibration [44]

B. Temperature Analysis

The temperature in moving components such as bearings is directly correlated to certain variables such as speed, applied load, and lubrication [45]. Increases in rotational speed applied load, and lubrication velocity will increase temperature. Faults and wear will also increase temperature. The temperature will increase over the course of the run time of machinery. Fig. 10 shows a typical temperature increase of a bearing over its run time. Knowing a system’s speed, applied load, and lubrication, you can build a model for the expected temperatures of a component. Substantial variation from this model can indicate substantial wear or faults. Readings will need to be taken over a long period and compared to models and previous data. Most temperature analysis will occur at the factory level of computation as it involves looking at trends of data.

C. Kalman Filtering

Kalman filtering is an estimation method which uses the systems dynamic model, system inputs, and measurements to estimate the variables of the system. It is often used to improve target tracking and the control of moving objects. Kalman filtering has many derived methods, such as the extended Kalman filter and unscented Kalman filter, which are better estimators for non-linear systems. [46]. Most systems have multiple models to describe their behaviour. Interacting

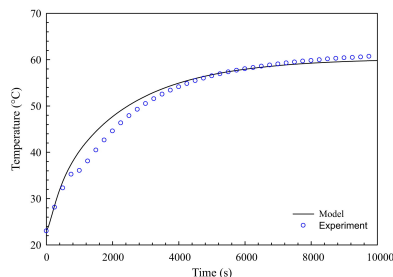


Fig. 10. Temperature change under running conditions [45]

multiple models (IMM) is a method to use multiple filters to improve estimation. IMM has often been used for fault detection [47]–[50] in actuators. Kalman filtering methods require a steady stream of low latency data. Kalman filtering is best to be performed at the machine level.

D. lubrication Analysis

Lubrication analysis involves analyzing the oil used to lubricate moving parts. This analysis is often only performed on large machines (> 50 horsepower) that use circulating oil systems. Qualities such as oil pH, capacitance, and temperature can be used to determine the condition of lubrication. The presence of chips and debris could also be detected to predict pitting or spalling on a bearing. An overview of lubrication oil CM is covered in a paper by Zhu et al . [30]. Data about the oil condition over time can be analyzed at the factory level. Simple analysis such as alerting the user if oil temperature has gone above a certain level can be computed at the machine level.

E. Machine Learning and Machine Vision

Machine learning is algorithms which can improve automatically through experience and collection of large quantities of Data. Machine learning algorithms are *black box* systems, that is that we only understand the inputs and outputs but not the calculations and variables used to determine outputs. Machine learning is an excellent method for estimating parameters given a large number of inputs making it an ideal choice for CM.

Machine vision is extracting data from images. Visual data collected from cameras could be processed for useful information. Machine vision has been used to detect faults in machine equipment [51] and detect wear in machine cutting tools [52]. Machine vision can be combined with temperature analysis in the form of thermography. Temperatures of each component can be taken without contact using a single infrared camera. Temperatures of the entire system can be analyzed this way.

These methods of analysis require high levels of data storage and computation, so they are best done at the factory level of computation.

F. Statistical process control and trend analysis

Data collected from the IoT CM system can be used for statistical process control (SPC) and trend analysis. Data collected can be used for SPC methods such as control charts Fig. 11 and histograms which can be useful for CM and fault detection applications. These methods would be computed at the factory level as they require data sets over long periods of time. [53]–[55].

VI. CONCLUSION AND FUTURE WORK

After examining available IoT technology and current techniques for CM, it is apparent that IoT technology is advantageous for the application of CM. The primary disadvantage of CM is the requirement to manually collect a great deal of data at a high frequency and the possibility of human error. These

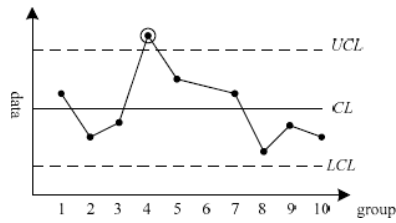


Fig. 11. Control chart [56]

disadvantages are eliminated if data can be collected and analyzed autonomously using an IoT system. An IoT autonomous CM system can help increase production throughput, reduce costs associated with maintenance, and help with decision making. When designing an IoT system for CM, the following design considerations have been examined: Types of available analysis, communication technology, and what sensors are available to collect the needed data. After considering these design criteria, and IoT CM system can be created.

To future augment the literature in the field, additional real-life applications of IoT CM systems should be studied. With more real-life examples, their benefits can be further examined and studied.

REFERENCES

[1] Y. Altintas, A. Verl, C. Brecher, L. Uriarte, and G. Pritschow, "Machine tool feed drives," *CIRP annals*, vol. 60, no. 2, pp. 779–796, 2011.

[2] K. Martin, "A review by discussion of condition monitoring and fault diagnosis in machine tools," *International Journal of Machine Tools and Manufacture*, vol. 34, no. 4, pp. 527–551, 1994.

[3] W. Zhou, T. G. Habetler, and R. G. Harley, "Bearing condition monitoring methods for electric machines: A general review," in *2007 IEEE international symposium on diagnostics for electric machines, power electronics and drives*. IEEE, 2007, pp. 3–6.

[4] N. Tandon, G. Yadava, and a. K. Ramakrishna, "A comparison of some condition monitoring techniques for the detection of defect in induction motor ball bearings," *Mechanical systems and signal processing*, vol. 21, no. 1, pp. 244–256, 2007.

[5] N. Tandon, "A comparison of some vibration parameters for the condition monitoring of rolling element bearings," *Measurement*, vol. 12, no. 3, pp. 285–289, 1994.

[6] T. Touret, C. Changenet, F. Ville, M. Lalmi, and S. Becquerelle, "On the use of temperature for online condition monitoring of geared systems—a review," *Mechanical Systems and Signal Processing*, vol. 101, pp. 197–210, 2018.

[7] Y. M. Al-Naggar, N. Jamil, M. F. Hassan, and A. R. Yusoff, "Condition monitoring based on iot for predictive maintenance of cnc machines," *Procedia CIRP*, vol. 102, pp. 314–318, 2021.

[8] K. Pichler, J. Klinglmayr, and M. Pichler-Scheder, "Detecting wear in a ball screw using a data-driven approach," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 3123–3128.

[9] A. Verl, U. Heisel, M. Walther, and D. Maier, "Sensorless automated condition monitoring for the control of the predictive maintenance of machine tools," *CIRP annals*, vol. 58, no. 1, pp. 375–378, 2009.

[10] S. Huang, K. K. Tan, and T. H. Lee, "Fault diagnosis and fault-tolerant control in linear drives using the kalman filter," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 11, pp. 4285–4292, 2012.

[11] S. Frey, M. Walther, and A. Verl, "Periodic variation of preloading in ball screws," *Production Engineering*, vol. 4, no. 2-3, pp. 261–267, 2010.

[12] G.-H. Feng and Y.-L. Pan, "Investigation of ball screw preload variation based on dynamic modeling of a preload adjustable feed-drive system and spectrum analysis of ball-nuts sensed vibration signals," *International Journal of Machine Tools and Manufacture*, vol. 52, no. 1, pp. 85–96, 2012.

[13] G.-H. Feng and Y. Pan, "Establishing a cost-effective sensing system and signal processing method to diagnose preload levels of ball screws," *Mechanical Systems and Signal Processing*, vol. 28, pp. 78–88, 2012.

[14] J.-L. Chang, J.-A. Chao, Y.-C. Huang, and J.-S. Chen, "Prognostic experiment for ball screw preload loss of machine tool through the hilbert-huang transform and multiscale entropy method," in *The 2010 IEEE International Conference on Information and Automation*. IEEE, 2010, pp. 376–380.

[15] P. Tsai, C. Cheng, and Y. Hwang, "Ball screw preload loss detection using ball pass frequency," *Mechanical Systems and Signal Processing*, vol. 48, no. 1-2, pp. 77–91, 2014.

[16] T. L. Nguyen, S.-K. Ro, and J.-K. Park, "Study of ball screw system preload monitoring during operation based on the motor current and screw-nut vibration," *Mechanical Systems and Signal Processing*, vol. 131, pp. 18–32, 2019.

[17] S. Ayvaz and K. Alpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using iot data in real-time," *Expert Systems with Applications*, vol. 173, p. 114598, 2021.

[18] A. Kanawady and A. Sane, "Machine learning for predictive maintenance of industrial machines using iot sensor data," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2017, pp. 87–90.

[19] B. Chen, J. Wan, A. Celesti, D. Li, H. Abbas, and Q. Zhang, "Edge computing in iot-based manufacturing," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 103–109, 2018.

[20] J. Wang, L. Zhang, L. Duan, and R. X. Gao, "A new paradigm of cloud-based predictive maintenance for intelligent manufacturing," *Journal of Intelligent Manufacturing*, vol. 28, no. 5, pp. 1125–1137, 2017.

[21] D. Collins, M. says, and Momentsleutel, "Where are glass scale linear encoders used?" Dec 2018. [Online]. Available: <https://www.linearmotiontips.com/where-are-glass-scale-linear-encoders-used/>

[22] F. Koumboulis, G. Petropoulos, and C. Mavridis, "Fault detection in machining via torque estimation," *IFAC Proceedings Volumes*, vol. 33, no. 20, pp. 217–221, 2000.

[23] J. Kim, I. Yang, D. Kim, M. Hamadache, and D. Lee, "Bearing fault effect on induction motor stator current modeling based on torque variations," in *2012 12th International Conference on Control, Automation and Systems*. IEEE, 2012, pp. 814–818.

[24] R. Bogue, "Sensors for condition monitoring: a review of technologies and applications," *Sensor Review*, 2013.

[25] D. Goyal and B. Pabla, "The vibration monitoring methods and signal processing techniques for structural health monitoring: a review," *Archives of Computational Methods in Engineering*, vol. 23, no. 4, pp. 585–594, 2016.

[26] S. Xu, F. Xing, R. Wang, W. Li, Y. Wang, and X. Wang, "Vibration sensor for the health monitoring of the large rotating machinery: review and outlook," *Sensor Review*, 2018.

[27] M. R. Akhondi, A. Talevski, and T.-H. Chou, "Criteria for hardware selection of wireless vibration sensor," in *2010 International Conference on Broadband, Wireless Computing, Communication and Applications*. IEEE, 2010, pp. 836–841.

[28] A. Béliveau, G. T. Spencer, K. A. Thomas, and S. L. Roberson, "Evaluation of mems capacitive accelerometers," *IEEE Design & Test of Computers*, vol. 16, no. 4, pp. 48–56, 1999.

[29] "Industrial iot wireless predictive maintenance sensor," Nov 2021. [Online]. Available: <https://store.ncd.io/product/industrial-iot-wireless-predictive-maintenance-sensor/>

[30] J. Zhu, D. He, and E. Bechhoefer, "Survey of lubrication oil condition monitoring, diagnostics, and prognostics techniques and systems," *Journal of chemical science and technology*, vol. 2, no. 3, pp. 100–115, 2013.

[31] C. Madhusudana, H. Kumar, and S. Narendranath, "Face milling tool condition monitoring using sound signal," *International Journal of System Assurance Engineering and Management*, vol. 8, no. 2, pp. 1643–1653, 2017.

[32] "Raspberry pi camera pinout," Dec 2020. [Online]. Available: <https://www.arduino.com/raspberry-pi-camera-pinout/>

[33] M. Yaseen, D. Swathi, and T. A. Kumar, "Iot based condition monitoring of generators and predictive maintenance," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2017, pp. 725–729.

[34] D. Takemura, A. Murata, K. Miyakoshi, T. Yokotani, Y. Kobayashi, and Y. Kobayashi, "System designing and prototyping on iot for a factory,"

- in *2019 International Conference on Information Networking (ICOIN)*. IEEE, 2019, pp. 327–329.
- [35] H. Lee and T. Kim, “Prototype of iot enabled smart factory,” *ICIC Express Lett. Part B Appl*, vol. 7, no. 4, pp. 955–960, 2016.
- [36] L. Hou and N. W. Bergmann, “Novel industrial wireless sensor networks for machine condition monitoring and fault diagnosis,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 10, pp. 2787–2798, 2012.
- [37] D. Eyers, R. I. Grosvenor, and P. W. Prickett, “Welding station condition monitoring using bluetooth enabled sensors and intelligent data management,” in *Journal of Physics: Conference Series*, vol. 15, no. 1. IOP Publishing, 2005, p. 024.
- [38] D. Tandur, M. Gandhi, H. Kour, and R. Gore, “An iot infrastructure solution for factories,” in *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2017, pp. 1–4.
- [39] F. Rincon, Y. Tanaka, and T. Watteyne, “On the impact of wifi on 2.4 ghz industrial iot networks,” in *2018 IEEE International Conference on Industrial Internet (ICII)*, 2018, pp. 33–39.
- [40] “Wireless topologies.” [Online]. Available: <https://www.emerson.com/documents/automation/training-wireless-topologies-en-41144.pdf>
- [41] “Machine condition monitoring: Increased uptime and revenue.” [Online]. Available: <https://www.ixon.cloud/knowledge-hub/machine-condition-monitoring-benefits-parameters-and-solutions>
- [42] M. Albert, “Making sure mtconnect is a good fit.” [Online]. Available: <https://www.mmsonline.com/articles/making-sure-mtconnect-is-a-good-fit>
- [43] S. Ebersbach and Z. Peng, “Expert system development for vibration analysis in machine condition monitoring,” *Expert systems with applications*, vol. 34, no. 1, pp. 291–299, 2008.
- [44] “Spectrum analysis,the key features of analyzing spectra.” [Online]. Available: <https://www.skf.com/binaries/pub12/Images/0901d1968024acef-CM5118-EN-Spectrum-Analysis-tcm-12-113997.pdf>
- [45] J. Takabi and M. Khonsari, “Experimental testing and thermal analysis of ball bearings,” *Tribology international*, vol. 60, pp. 93–103, 2013.
- [46] S. J. Julier and J. K. Uhlmann, “New extension of the kalman filter to nonlinear systems,” in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. International Society for Optics and Photonics, 1997, pp. 182–193.
- [47] S. Kim, J. Choi, and Y. Kim, “Fault detection and diagnosis of aircraft actuators using fuzzy-tuning imm filter,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 3, pp. 940–952, 2008.
- [48] Y. Zhang and J. Jiang, “Integrated active fault-tolerant control using imm approach,” *IEEE Transactions on Aerospace and Electronic systems*, vol. 37, no. 4, pp. 1221–1235, 2001.
- [49] L. Cork and R. Walker, “Sensor fault detection for uavs using a nonlinear dynamic model and the imm-ukf algorithm,” in *2007 Information, Decision and Control*. IEEE, 2007, pp. 230–235.
- [50] C. Rago, R. Prasanth, R. K. Mehra, and R. Fortenbaugh, “Failure detection and identification and fault tolerant control using the imm-kf with applications to the eagle-eye uav,” in *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, vol. 4. IEEE, 1998, pp. 4208–4213.
- [51] V. Chauhan and B. Surgenor, “A comparative study of machine vision based methods for fault detection in an automated assembly machine,” *procedia manufacturing*, vol. 1, pp. 416–428, 2015.
- [52] J. Zhang, C. Zhang, S. Guo, and L. Zhou, “Research on tool wear detection based on machine vision in end milling process,” *Production Engineering*, vol. 6, no. 4, pp. 431–437, 2012.
- [53] Z. Lu, M. Wang, and W. Dai, “A condition monitoring approach for machining process based on control chart pattern recognition with dynamically-sized observation windows,” *Computers & Industrial Engineering*, vol. 142, p. 106360, 2020.
- [54] S. J. Bae, B. M. Mun, W. Chang, and B. Vidakovic, “Condition monitoring of a steam turbine generator using wavelet spectrum based control chart,” *Reliability Engineering & System Safety*, vol. 184, pp. 13–20, 2019.
- [55] Z. Jiao, W. Fan, and Z. Xu, “An improved dual-kurtogram-based control chart for condition monitoring and compound fault diagnosis of rolling bearings,” *Shock and Vibration*, vol. 2021, 2021.
- [56] H. Yan, G. Gao, J. Huang, and Y. Cao, “Study on the condition monitoring of equipment power system based on improved control chart,” in *2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*. IEEE, 2012, pp. 735–739.

Research on Ambient Backscatter Communication Signal Detection Algorithm Based on Digital Terrestrial Multimedia Broadcast

Chen Ruan

College of Electronics Science
National University of Defense Technology
Changsha, China
ryuanlin@163.com

Hongyi Wang

College of Electronics Science
National University of Defense Technology
Changsha, China
wanghongyi2011@163.com

Liming Zheng

College of Electronics Science
National University of Defense Technology
Changsha, China
lmzheng@nudt.edu.cn

Jianfei Wu

College of Electronics Science
National University of Defense Technology
Changsha, China
wujianfei990243@126.com

Fengxian Ma

Tianjin Institute of Advanced Technology
National University of Defense Technology
Tianjin, China
1443665385@qq.com

Abstract—Digital Terrestrial Multimedia Broadcast (DTMB) has high signal coverage and strong stability, so it can be used in Ambient Backscatter Communication (AmBC). To solve the problem of signal detection performance degradation of traditional algorithm in high backscattering rate scenes, this paper presents a signal detection algorithm based on DTMB frame structure. By introducing preamble, the receiver can not only detect the reflected link signal from the received signal, but also perceive the corresponding relationships between the information symbol and the DTMB signal frame; by introducing training symbol, the receiver can get the decision threshold of the information symbol, and then complete the demodulation of the information symbol. Simulation results show that the proposed algorithm significantly improves the signal detection performance.

Keywords—DTMB, AmBC, signal detection algorithm, signal frame

I. INTRODUCTION

The Internet of Things (IOT) is a network for the interconnection of all things. It has been nearly 30 years since it was proposed. As a product of the rapid development of science and technology, the IOT has formed a very close relationship with various fields of society, and constantly brings convenience to daily life.

Ambient Backscatter Communication (AmBC) uses Radio Frequency (RF) signals present in the environment to communicate, and shares spectrum with communication systems that provide RF signals. Because of less dependence on the dedicated RF sources, AmBC can reduce communication energy consumption and maintenance costs, which means AmBC has broad research space and application prospects.

Digital Terrestrial Multimedia Broadcast (DTMB) has three key advantages as ambient RF signals. First, DTMB has a high signal coverage. As of 2011, the coverage of DTMB in China

has reached 94.52%. Second, the television tower works day and night, so the DTMB signal is stable. Finally, the DTMB system adopts DMB-TH technology so that the DTMB signal has strong anti-interference ability, and the terminal has excellent receiving performance under mobile conditions.

In this paper, DTMB signals are combined with AmBC to study signal detection algorithm under the scene of high backscattering rate.

II. AMBIENT BACKSCATTER COMMUNICATION SYSTEM

The AmBC system is composed of an ambient RF source, a backscattering device and a receiver. The RF signal can be Wi-Fi, FM, LoRa, BLE, base station and other signals present in the environment [2-6]. The information to be transmitted by the backscattering device is superimposed on the radio frequency signal, then the receiver captures the backscattered signal and demodulates it.

A. Backscattering device

Fig. 1 shows the structure of the backscattering device, which is mainly made up of antenna, energy harvesting module, micro-control unit, backscatter modulation module, etc.

- The energy harvesting module provides power to the backscattering device by collecting RF energy in the environment.
- The micro-control unit encodes the information to obtain the binary sequence. According to the binary sequence, the modulation signal of the ambient RF signal is generated, which controls backscattering modulation module to complete the Binary Amplitude Shift Keying.
- The backscattering modulation module realizes the absorption and reflection of ambient RF signal by RF switch. When the impedance of RF switch and antenna match, the RF signal energy is absorbed by the load,

indicating the binary data "0"; when the impedance mismatch, the RF signal is reflected by the antenna, representing the binary data "1" [11-12].

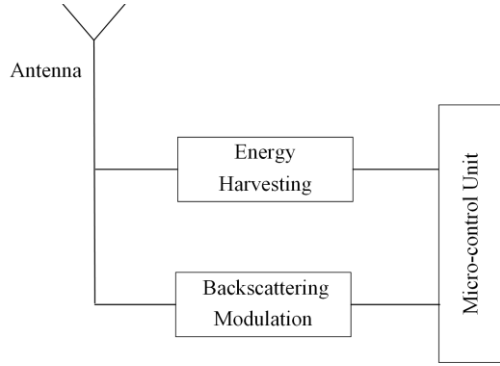


Fig. 1. Structure of backscattering device

B. Receiver

The path of the RF signal from the RF source to the receiver is the direct link, while the path from the RF source to the receiver through the backscattering device is the reflected link.

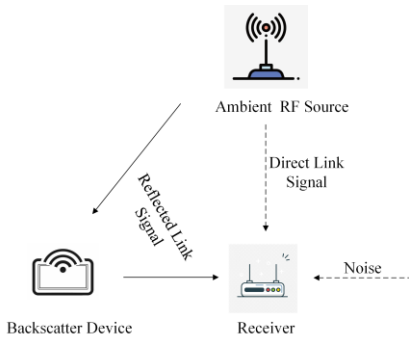


Fig. 2. Signal captured by the receiver

As shown in Fig. 2, the signal $y(t)$ is the superposition of the direct link signal, the reflected link signal, and the noise.

$$y(t) = s(t) + \alpha B(t)s(t) + u(t) \quad (1)$$

In the above equation, $s(t)$ is the signal sent by RF source, $B(t)$ is the modulation signal, and α is the complex attenuation coefficient of the reflected link signal relative to the direct link signal. The role of backscatter receiver is to demodulate the reflected link signal.

C. Digital Terrestrial Multimedia Broadcast Signal

The AmBC based on DTMB is studied in this paper. DTMB is the Chinese digital video broadcasting standard. The signal frame, as the basic transmission unit, consists of a frame header and a frame body, which is shown in Fig. 3.

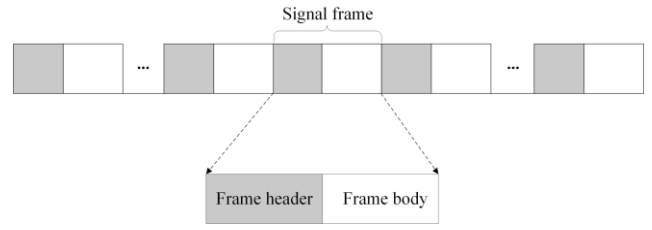


Fig. 3. DTMB signal structure

In order to overcome the multipath interference caused by different delays, the standard provides frame headers of three lengths. The parameters are shown in Table 1.

TABLE I. DTMB FRAME HEADER PARAMETERS

Signal Frame	Frame header structure	Frame header length (μs)	Maximum distance (km)
Frame structure 1	PN420	55.5	16.7
Frame structure 2	PN595	78.7	23.6
Frame structure 3	PN945	125	37.5

The baseband symbol rate is 7.56 Mbps. Take the frame structure 1 as an example, the frame header symbol contains 420 symbols, then the length of the frame header is 55.5 μs , and the transmission distance of the RF signal within the frame header is 16.7km, i.e., the maximum distance of the single frequency network.

The average power of frame header 2 is the same as that of frame body, while the average power of frame header 1 and frame header 3 is twice that of frame body [9-10].

III. AMBC SIGNAL DETECTION ALGORITHM BASED ON DTMB

A. Traditional Signal Detection Algorithm

An ambient backscatter communication signal detection algorithm based on the TV Tower signal was proposed in [1] as follows:

Do Nyquist sampling to (1) to get the signal $y[n]$:

$$y[n] = s[n] + \alpha B[n]s[n] + w[n] \quad (2)$$

where $B[n]$ is either "1" or "0", $s[n]$ is uncorrelated with noise, $w[n]$. If the receiver averages the instantaneous power in the N samples corresponding to a single backscattering bit, the following equation is obtained:

$$\frac{1}{N} \sum_{n=1}^N (y[n])^2 = \frac{|1+\alpha B|^2}{N} \sum_{n=1}^N (s[n])^2 + \frac{1}{N} \sum_{n=1}^N (w[n])^2 \quad (3)$$

Say P_{TV} is the average power in the received TV signal:

$$P_{TV} = \frac{1}{N} \sum_{n=1}^N (s[n])^2 \quad (4)$$

Ignoring the noise, when binary "1" is sent, the average power of the received signal is $|1+\alpha|^2 P_{TV}$; when binary "0" is sent, the average power of the received signal is P_{TV} . The

receiver can decode the information from the backscattering device by distinguishing $|1+\alpha|^2 P_{TV}$ and P_{TV} .

In this algorithm, it is assumed that the average power of the TV signal corresponding to different backscattering bits is same. If the average power of the TV signal is different in different backscattering bit duration, the performance of the above algorithm will become worse.

Therefore, if the DTMB signal of frame structure 2 is used in AmBC, the receiver can decode information from the backscattering device through the above algorithm. For the DTMB signal with frame structure 1 or frame structure 3, the receiver can accurately extract information from the received signal only in the scene of low backscattering rate.

B. An Signal Detection Algorithm Based on Frame Structure

In this section, the DTMB signal detection algorithm corresponding to frame structure 1 or frame structure 3 is studied in high backscattering rate scenes.

The DTMB signal average power of a signal frame length can be expressed as:

$$P_f = \frac{T_b}{T_h+T_b} \times P_b + \frac{T_h}{T_h+T_b} \times 2P_b \quad (5)$$

where T_h represents the duration of the frame header in a signal frame, T_b represents the duration of the frame body in a signal frame, and P_b is the average power of the frame body.

The data sent by the backscattering device is composed of preamble, training symbol and information symbol. The preamble is "100", and the training symbol is composed of alternating "10". Each preamble and training symbol lasts for the same time as the length of a signal frame.

The receiver calculates the average power as follows:

$$P_{av} = \frac{1}{b-a} \sum_{n=a}^{n=b} (y_I[n])^2 + (y_Q[n])^2 \quad (6)$$

where $y_I[n]$ and $y_Q[n]$ are the in-phase and quadrature samples of the received signal.

1) The receiver slides to calculate the average power of the output samples of a signal frame length. At the time t_1 , the receiver calculates the average power of the output signal within $(t_1 - T_f, t_1)$; at next moment t_2 , the receiver calculates the average power within $(t_2 - T_f, t_2)$, where T_f is the length of a DTMB signal frame. If data are transmitted, the average power calculated by sliding will exceed the threshold set with reference to $|1+\alpha|^2 P_f$, and the receiver records the time T_1 corresponding to the maximum average power.

2) Starting from time $t_1 + T_f$, the receiver slides to calculate the average power of the output samples of a frame header length. Since the power of the frame header is twice that of the frame body, the average power calculated by sliding will exceed the threshold set with reference to $2P_b$, and the receiver records the time T_2 corresponding to the maximum average power.

3) According to T_1 and T_2 , the receiver can perceive the corresponding relationship between training symbol, information symbol and DTMB signal frame. In DTMB signal, the frame body of the previous signal frame is connected with the frame header of the next signal frame. According to the start of transmission, there are two corresponding relationships between preamble and DTMB signal frame, as shown in Fig. 4.

a) The selection of the second preamble: The second preamble is "0" so that the average power value obtained at the end of the transmission of the first preamble "1" is maximum. Then, the end time of the first preamble is determined, i.e., T_1 .

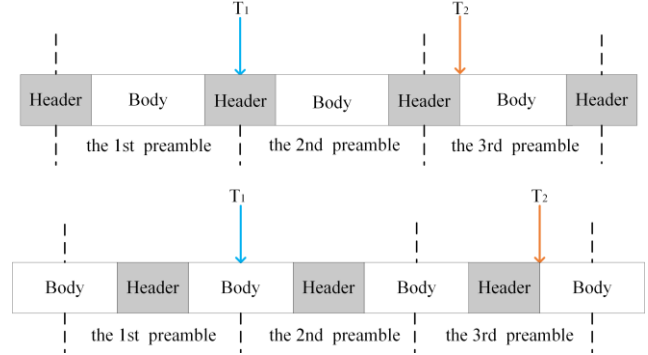


Fig. 4. Two corresponding relationships between preamble and DTMB signal frame

In order to avoid the influence of the frame header corresponding to the second preamble, the receiver start to calculate the average power from time $T_1 + T_f$.

b) The selection of the third preamble: It is the third preamble is "0" that the receiver can determine the moment when the frame header ends, i.e., T_2 . If the third preamble is "1", the receiver may not be able to accurately find when the frame header ends.

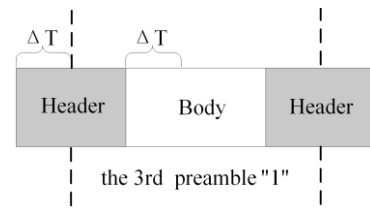


Fig. 5. The third preamble "1" may cause the receiver to go wrong.

In Fig. 5, ΔT is the time interval between the frame header and the third preamble. When the following equation is established, the time T_2 found by the receiver will be after the end of the frame header, and the receiver will wrongly perceive the corresponding relationships between the training symbol, the information symbol and the DTMB signal frame.

$$2P_b \times \Delta T < |1+\alpha|^2 P_b \times \Delta T \quad (7)$$

4) According to the corresponding relationship between the training symbol and the DTMB signal frame, the receiver obtains the decision threshold. It is complicated to derive the decision threshold from the channel estimation of the reflected link. For simplicity, the threshold is calculated by using the training symbol[7-8]. The training symbol consists of several groups of "10". The receiver can obtain the power levels of the frame header in the reflection state, the frame body in the reflection state, the frame header in the non-reflection state and the frame body in the non-reflection state, which are recorded as \bar{P}_{hr} , \bar{P}_{br} , \bar{P}_{hn} and \bar{P}_{bn} respectively.

5) The receiver calculates the average power of each information symbol according to the corresponding relationship between the information symbol and the DTMB signal frame. In the high backscattering rate scene, there are three kinds of corresponding relationship between the information symbol and the DTMB signal frame: only corresponding to the frame header; only corresponding to the frame body; and corresponding to the frame header and the frame body. If the information symbol corresponds to the frame header and the frame body, the samples of the corresponding frame header or frame body is selected to calculate the average power. Say B_N is the Nth information symbol and P_N is the average power of the Nth information symbol. The receiver decodes the backscattered information through (8).

$$B_N = \begin{cases} 1, & P_N > \frac{\bar{P}_{hr} + \bar{P}_{br}}{2} \\ 0, & \frac{\bar{P}_{hr} + \bar{P}_{br}}{2} \geq P_N > \frac{\bar{P}_{br} + \bar{P}_{hn}}{2} \\ 1, & \frac{\bar{P}_{br} + \bar{P}_{hn}}{2} \geq P_N > \frac{\bar{P}_{hn} + \bar{P}_{bn}}{2} \\ 0, & P_N \leq \frac{\bar{P}_{hn} + \bar{P}_{bn}}{2} \end{cases} \quad N=0,1,\dots, \quad (8)$$

C. Simulation Results and Analysis

The DTMB signal is generated by the module built in Advanced Design System, the preamble is "100", the training symbol is set to 5 groups of "10", and the backscattering rate of information symbol is 8kbps. At the same time, the noise is added. The performance of the AmBC signal detection algorithm based on DTMB frame structure is shown in Fig. 6 and Fig. 7.

1) Fig. 6 shows the relationship between bit error rate and signal-to-noise ratio when α equals to 0.15. It can be seen that with the increase of the signal-to-noise ratio, the bit error rate of the algorithm proposed in this paper tends to zero. In the high backscattering rate scene, this algorithm takes into account the difference of frame header and frame body power of DTMB signal, and increases the power level to distinguish the reflected and non-reflected states of information symbols from two to four, so that the demodulation performance is obviously better than the traditional algorithm. In addition, there is a certain degree of deviation between the time T_1 and T_2 determined by

receiver and the exact value. The greater the signal-to-noise ratio, the less influence of the offset on demodulation.

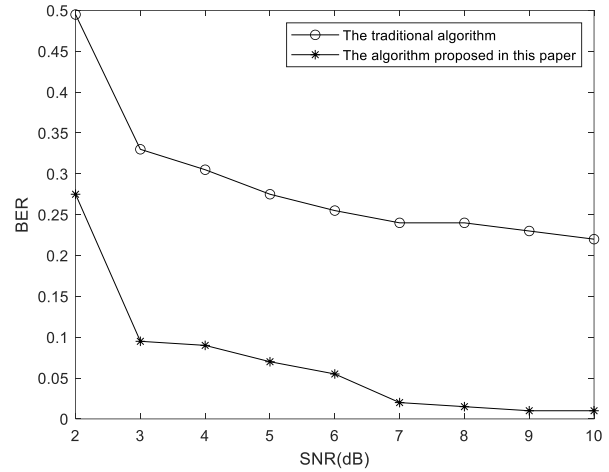


Fig. 6. Effect of signal-to-noise ratio on bit error rate

2) Fig. 7 shows the relationship between the bit error rate and the attenuation coefficient when SNR is 4dB. The attenuation coefficient is affected by antenna efficiency and path loss. When the attenuation coefficient is close to 0, the bit error rate approaches 50%, this is because the noise level is greater than the distinction of $|1+\alpha|^2 P$ and P . With the increase of attenuation coefficient, the interference of noise decreases, and the demodulation performance becomes better.

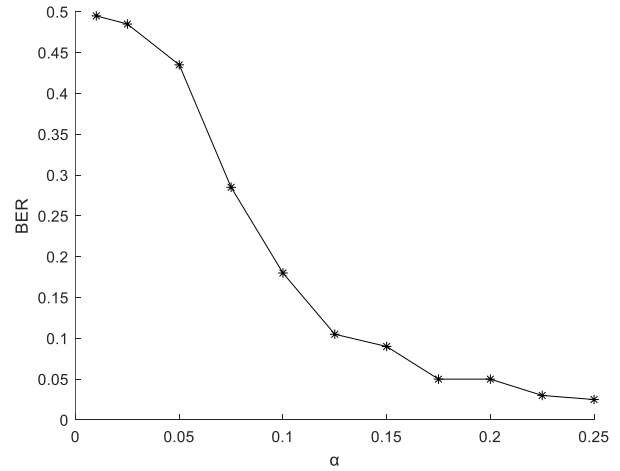


Fig. 7. Effect of attenuation coefficient on bit error rate

IV. SUMMARY

A signal detection algorithm based on DTMB frame structure is proposed in this paper, which solves the problem that the communication rate can't be higher than the frame rate of DTMB signal. Through the introduction of preamble and training symbol, the receiver can perceive the corresponding relationships between the information symbol and the DTMB signal frame, and obtain the decision threshold of demodulation information symbol. The simulation results show that compared with the traditional algorithm, the

proposed algorithm significantly improves the signal detection performance in the scene of high backscattering rate.

REFERENCES

- [1] Liu V , Parks A , et al. Ambient Backscatter: Wireless Communication Out of Thin Air[J].ACM SIGCOMM Communication Review,2013.
- [2] Gollakota S, Reynolds M S, Smith J R, et al. The Emergence of RF-Powered Computing[J]. Computer, 2014, 47(1):32-39.
- [3] Wetherall, David, Gollakota, et al. Wi-Fi Backscatter: Internet Connectivity for RF-Powered Devices[J]. Computer Communication Review: A Quarterly Publication of the Special Interest Group on Data Communication, 2014, 44(4):607-618.
- [4] Wang A. , Vikram I. , Vamsi T. , et al. FM Backscatter: Enabling Connected Cities and Smart Fabrics[C]. Symposium on Networked Systems Design and Implementation, 2018.
- [5] Yuan L, Zhang R, Yang K, et al. Protocol-Aware Backscatter Communication Using Commodity Radios[C]// 2020 IEEE/CIC International Conference on Communications in China. IEEE, 2020.
- [6] M. Zhang, J. Zhao, S. Chen, et al. Reliable Backscatter with Commodity BLE[C] // IEEE INFOCOM 2020 - IEEE Conference on Computer Communications. IEEE, 2020.
- [7] Kimionis J . Design and Implementation of Backscatter Links with Software Defined Radio for Wireless Sensor Network Applications[D]. Chania, Technical University of Crete,2011, 39-66.
- [8] Biao Zhou. Design and Implementation of Receiver Algorithm for Ambient Backscatter Communication System [D]. University of Electronic Science and Technology,2020.
- [9] GB 20600-2006, Frame Structure, Channel Coding and Modulation of Digital Terrestrial Multimedia Broadcast Transmission System [S]
- [10] Jingfeng Feng, Jun Liu, Xingwei Zhou. Interpretation of National Standard GB20600-2006 *Frame Structure, Channel Coding and Modulation of Digital Terrestrial Multimedia Broadcast Transmission System*[J].Radio and television technology,2007(05):21-22+24-28+15.
- [11] Ge Shi. Research on Passive Ambient Electromagnetism Backscatter Communication[D].Beijing University of Posts and Telecommunications, 2020.
- [12] Wenjie Wang. Design and Simulation of Ambient Radio Backscatter Communication System[D]. Southwest University of Science and Technology, 2017.

Dynamic Analysis of Demographic Sentiment

1st Joshua Weston

*Department of Computer Science
Thompson Rivers University
British Columbia, Canada
westonj18@mytru.ca*

2nd Brenden Bickert

*Department of Computer Science
Thompson Rivers University
British Columbia, Canada
bickertb16@mytru.ca*

3rd Caleb Stasiuk

*Department of Computer Science
Thompson Rivers University
British Columbia, Canada
stasiukc161@mytru.ca*

4th Fadi Alzhouri

*Department of Computer Science
Trent University
Peterborough, Canada
fadi.alzhouri@trentu.ca*

5th Dariush Ebrahim

*Department of Computer Science
Thompson Rivers University
British Columbia, Canada
debrahimi@tru.ca*

Abstract—There is no doubt that big data analysis has a very positive impact on economics, security, and other aspects for countries and enterprises alike. Where we have recently noticed the frantic competition between companies to increase their profits by analyzing the largest amount of data as quickly as possible. Especially analyzing data related to Covid-19 to make the most of information in all areas. Covid-19 has drastically affected many lives in recent years but, even in these hard times, businesses can leverage the current pandemic to make a profit. In this paper, we investigate a variety of tweets using MapReduce, Spark, and Machine Learning methods to determine the sentiment of a given tweet based on the information provided by the dataset. With this information, businesses could learn how to present Covid-19 and pandemic related goods and information in a way that will be well received by its audience. To take this a step further, we will investigate trends in sentiment across demographics tweeting about the virus. This information in sentiment is dynamically useful to understand how specific audiences feel about the pandemic. We explore which Machine Learning methods produce the best results such as Multi-Layer Perceptron neural networks and Logistic Regression.

Index Terms—Big Data, Machine Learning, Dynamic Sentiment Analysis, Spark, MapReduce, COVID-19, Tweets.

I. INTRODUCTION

As time goes on, the amount of data available for analysis increases by quite a large margin. Data mining allows us to find what is important in data, the value. Big-Data programming models such as MapReduce aim to alleviate some of the challenges of working with such large volumes of data. We aim to make use of some of these Big-Data programming models in order to retrieve some value out of a considerably large dataset. The dataset being used contains tweets that are related to the COVID-19 pandemic. The data is extracted from over 2.5 billion tweets and organized into 7 different tables. With this large amount of labelled data, we will explore the relationship between different demographics and how much each has been tweeting about COVID-19.

Our objective is to find the volume of tweets from differing demographics to discover what type of people are primarily

tweeting about COVID-19. To solve this problem, MapReduce methodologies will be used to mine and categorize the data. Although the circumstances are unpleasant, the current pandemic has been leveraged and utilized by businesses to turn a profit, like determining whether to increase or decrease productions for certain products. The findings from this paper can be used, for example, to aid businesses such as e-commerce companies in targeting ads towards these demographics for pandemic-related goods.

A number of methods are proposed, such as the following:

- A MapReduce join to create a full table along with a synthetic text.
- A Spark-based counter to gather instances.
- Machine Learning classification using various methods to determine the most useful approach for this data.

This paper is organized into a number of sections. Section 2 reviews this topic and state-of-the-art literature. Section 3, Materials and Methods, provides a description of the dataset that is being used to conduct the research of this paper, the system architecture with the definition of the problem, and the proposed solution methods. Section 4 is an overview of the simulation results. Section 5 discusses the possible aspects of the improvement, followed by section 6, which concludes this study.

II. RELATED WORK

Es-Sabery et al. [1] provide a paper, “A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier,” in which they describe the problem of classifying opinions based on sentiment in the context of big data. The analysis of opinions with sentiment is important due to the fact that its applications can be applied in beneficial ways. Many applications are mentioned such as consumer trends, marketing strategies, and developments in the stock market among other things. The challenges that come with analyzing such data include long execution time due to volume and the variety of content

that comes in Twitter posts. They propose an improved ID3 decision tree classifier to create a model for their COVID-19 related tweets dataset. As for the long runtime execution issue, they propose handling the data in a number of mapping steps that lead to a reducing step that results in the ID3 decision tree classifier. While they propose a unique enhanced ID3 Decision Tree Classifier, our work will instead focus on the comparison of unaltered existing methods like Neural Networks, Linear Models, and Decision Trees. Their COVID-19 dataset is much smaller and focuses on an 11 day time period of tweets exclusively from India (150MB), while our dataset will be magnitudes of size larger. The dataset they use has text included, whereas ours does not and we will need to construct a synthesized text out of select tables.

[2] takes a deep dive into the human element that is in opinion mining and sentiment analysis research. They cover three main ways to implement the research: keyword-based classification, lexicon-based classification, and a wide variety of machine learning-based approaches. The machine learning-based methods are constructed around two core components, feature extraction and a machine learning algorithm. Additionally, they list application areas that can be beneficial to employee use of opinion mining and sentiment analysis, including healthcare, finance, sports, politics, tourism, and consumer behaviour, as well as some other less conventional applications like box office performance predictions based off a movie trailer. In the final sections of the paper they cover some of the open challenges of big data sentiment analysis, both technical and non-technical. The aim of their paper is to provide an overview of the process and procedure behind conducting sentiment analysis, our aim is to implement and take advantage of some of the methods explored in this paper to use in our research.

[3] discusses the power of using sentiment analysis by companies to benefit their market share. The typical use case for sentiment analysis is to analyze a targeted group of people's sentiment towards different entities. These entities could be organizations, products, services, individuals, events, topics, and issues which are gathered online through text, video, and other means of communication. Since they explore the vast number of ways sentiment can be gathered, our scope is focused on a single area, Twitter. There are three categories that the collected sentiment can fall into: positive, neutral, and negative. These can then be used in various applications, which can include stock price prediction, public voice analysis, and crowd surveillance. They also talk about some of the issues that occur with analysis in big data. One of the most prevalent of these issues that occurs is flexible capacity, and having to allocate more time and resources to solve problems. Something that adds to this problem is the fact that the solution to many problems in big data is not black-and-white. The paper talks a lot about frameworks used for sentiment analysis like Hadoop Distributed File System (HDFS), SQL, and Map Reduce for example, as well as outlining general approaches and features of sentiment analysis. We determined that it

would be best to focus on the particular frameworks which fit our use case best, instead of experimenting with a number of different methods. This paper focuses on how sentiment analysis can be used and implemented, whereas our work will focus on the actual implementation of some of these methods to produce results.

In search of the best solutions for sentiment analysis of tweets, a paper which uses Apache spark for sentiment analysis was found. This paper related perfectly to our data set, as the sentiment analysis was also being done on COVID-19 tweets. Spark stores in between data in the memory, making it useful for low-latency jobs [4]. The performance of spark stands out. By using memory in this way, it increases the efficiency of the system. This should not come as a surprise since using memory is a faster option when compared to loading intermediate values on the disk. Where our research differs is in what we are doing with our analysis. While this paper offers solutions for efficiently dealing with the data, we plan to actually create some models to predict sentiment. This paper only provides a solution of how to deal with our large data set, since processes can be slow with so much data.

III. MATERIALS AND METHODS

A. Description of the Dataset

The dataset that is being used in this paper is titled "COVID-19 Tweets Dataset (over 1 billion)"[5] and it is a vast repository of COVID-19 related tweets. The total number of tweets located in the dataset is 2,648,139,275 across the years 2020, 2021, and 2021, which is quite sizable and reaches a total space in memory of over 100GB estimated compressed. When the dataset is uncompressed, the memory it takes up is around 7 times the amount, totalling a probable 700 plus gigabytes. Therefore, within reason, we are limiting our use of the data to a subset of the available 2022 data that takes up around 35 gigabytes uncompressed.

The dataset is split over 7 tables that share a tweet ID as the primary key. The first table is the details of the tweet and this includes, language, if geolocation is in the tweet, if it is a retweet, likes, retweets, country, date, and time created. The next table is hashtags which is a one-to-many relationship where one id maps to many hashtags. Mentions are what the next table is composed of and it functions much like the hashtags table. The fourth table is the sentiment table, here the tweets are labeled and some specific predictions for neutral, positive, and negative are included. The fifth is the NER (Named-Entity-Recognition) table and this includes information relating to entities found within the tweets. The last two tables are the same as the sentiment and NER tables, but with the difference that they are for Spanish (ES) content. The tables are not all stored in a single file, instead, they are spread across files and directories. First off, the years are separated into different GitHub repositories as found in the ReadMe.md of the base repo which contains the latest year data. Each table has its own directory (Example, Summary_Details) and within that, they are further split by

month. Within the month directory, one CSV file is uploaded for every hour of every day of the month.

Nowhere within the data is the raw text of what was actually the tweet, but we still get enough information like hashtags, mentions, and NER values that we should not have to worry about not having the raw tweet text. If the raw text information of a tweet was really required then we can query the Twitter API with the tweet ID provided in one of the tables, this is more limited as the base developer account for Twitter has a limit of 500k calls a month. Therefore, the Twitter API will not be relied on for the majority of this project.

B. System Architecture and Problem Definition

A MapReduce style programming model will be implemented for various parts of the data analysis and preparation, but not through a framework like Hadoop. The data may have to be joined together due to the issue with it being stored across different tables and this is something MapReduce could achieve. Some of the analysis that requires MapReduce can be done through Apache Spark.

For the Machine Learning part of the project we will be using the library Sci-kit Learn. Sci-Kit Learn has a vast library of methods including a number of ML methods. Methods include different kinds of ML algorithms such as classification and regression algorithms, along with some useful scaling methods. Classification methods will be more important for this project, and perhaps we will use many depending on training time, but first we may look into Naive Bayes. Other potentially useful ML methods include Logistic Regression, Support Vector, and Multi-Layer Perceptron.

1) *System Architecture*: Our architecture, as seen in Fig.1, for getting to Machine Learning will work kind of like a pipeline architecture. First the data goes through joins, then feature extraction, and finally training.

2) *Problem Definition*: We hope to gain information from the COVID-19 dataset and see how well a machine learning model can perform, but this is a difficult task with how the data is organized. One such issue is that of the data being segregated into a number of tables. It would be easier if all the feature data we hope to use for Machine Learning is contained within a single table. Another issue is how to deal with the text data, using the text data as it is raw is not doable for the Machine Learning process. Some solution must be proposed for this problem.

C. Solution Method

To attain some information from the dataset that could provide insight, one of the methods we implement will be a MapReduce instance count. The class instance we are going to collect is the sentiment label, which is either positive, negative, or neutral. With the result of this count we should be able to plot the amount of tweets of each instance over time as well as the total class instance imbalance. It will take two MapReduce jobs to accomplish the counts and aggregations for hourly, daily, and total. The first MapReduce job will map out the contents of each file into intermediate values which

correspond to hourly instance counts, and we can keep these. After the map we will aggregate the keys into daily counts. The second MapReduce job will simply have the mapper map out the daily counts as intermediate pairs and then the reducer aggregates them into a total count for the data we are using.

Three of the tables in the dataset include information that could be of value if counted. The NER, mentions, and hashtags tables all include a single column within them dedicated to some kind of text data. If we count all of the values of those columns we can find some of the top values included in a tweet in a specific set of time. This will be accomplished using Apache Spark to count the words and then output them into a list we can sort for top words, as seen in Algorithm 1. There will be a lower limit of how many words must appear in a single hour of data. The lower limit will allow us to filter out low usage terms, and save space on memory and disk. For example, if we set the lower limit to 10, then we are not missing out on much data. Say there are 100,000 tweets per hour, we would only miss out on a maximum of 240 instances of a term over 24 hours of 2,400,000 tweets. Popular terms will have much higher instance counts.

Algorithm 1 Spark MapReduce Count

```

f ← sc.textFile(path)
s ← f.map(Line.split(',').foreachLine)
c ← s.map((lowercase(w), 1)foreachw)
c ← c.reduceByKey(w1 + w2foreachw1, w2)
sort(c)
i ← 0
while i < length(c) do
  if c[i] ≥ somethreshold then
    break
  else
    i ← i + 1
  end if
end while
return c up to i, exclusive

```

One of the sub-problems of this problem is getting the data ready for Machine Learning algorithms. Some of the data has been prepared via MapReduce style programming. Since the data resides across a number of tables, it would be much better for the features that we are going to use to be shared in a singular table. So the proposed solution for this issue is that of a Map-side join. We use this to join the tables into one. The map-only MapReduce job works here because the data we have is split into many small files already which will be able to fit into main memory and produce the many final small tables.

We will also use this Map-side join style for a similar issue before joining every table we are using. Rather than a join, it will be a Map-side concatenate. The issue is that some tables include one-to-many relationships, and because of this we are unable to simply join unless we perhaps want duplicate records. These tables are the NER, hashtags, and mentions

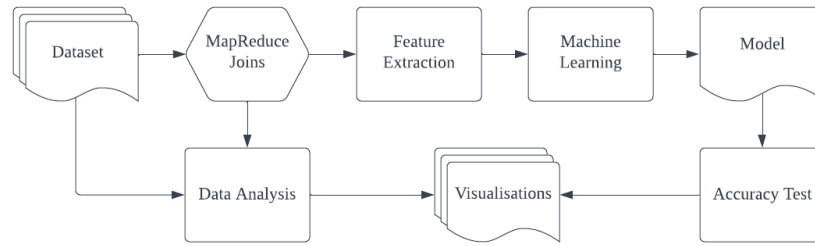


Fig. 1. Pipeline System Architecture.

tables. The proposed solution is that if we combine the many values that are mapped to a single tweet ID into a single string, it shall work as outlined in Algorithm 2. For example, say we have the hashtag value for a tweet as "#hashtag" and there also happens to be a mention for the tweet, "@mention", after the Map-side concatenate algorithm the result will be "#hashtag @mention" as the stored text value. A possible benefit of this is that we will never have to obtain the plain text of a tweet from Twitter, since we will have most of the important values that would be in plain text such as named entities, hashtags, and mentions. Furthermore, possibly less valuable plain text values such as determiners like "the", "a" and other linguistic values that provide less information will not have to be dealt with.

Once we have joined all the text values into a one-to-one relationship table, we can join this with the last remaining table, the details table, as seen in Algorithm 3. Because of the one-to-one relationship of each table, we can join the two tables using the tweet ID as the key. The result shall give us a single table to work upon for our machine learning step. Although, from this point the text and other features may need to be encoded somehow.

Algorithm 2 MapReduce Join Mapper Text

```

f ← fileName
paths ← [ner, hash, mention]
outmap ← emptyhashmap
for P in path do
  lines ← openFile(P + f)
  for l in lines do
    if outmap.in(lid) then
      outmap[lid] ++
    else
      outmap[lid] ← 1
    end if
  end for
end for
emit(f, outmap)
  
```

For the machine learning part, we will test some features and see how they compare to each other. Once the best feature has been found we will use that for further experimentation. A number of different classifiers will be tested to see which

Algorithm 3 MapReduce Join Mapper Full

```

f ← fileName
paths ← [summary, text]
txtmap ← load(paths[text] + f)
for line in open(paths[summary]+f) do
  lines ← openFile(P + fileName)
  for l in lines do
    l.append(',')
    if txtmap.in(lid) then
      l.append(txtmap[lid])
    end if
  end for
end for
  Save Updated paths[summary]+f file
  emit(f, paths[summary] + f)
  
```

provides the best possible score using the accuracy metric. These classifying methods include Ridge, PassiveAggressive, Support Vector, Multinomial Naive Bayes, Stochastic Gradient Descent, Decision Tree, Multi-Layer Perceptron, K-Nearest-Neighbors, and Logistic Regression. All of these methods will be trained from the library sci-kit learn.

Feature extraction of the joined text column of the data set is something that needs to be dealt with. The proposed solution for this is to convert the words into a vector representation that is stored inside of a sparse matrix. The sparse matrix will allow us to save a large amount of memory compared to the usage of a large dense matrix. More useless features will be dropped from the data set as we focus on the features with more potential. Date, country, tweet ID, and language will be dropped. Date and tweet ID do not seem to be much more than arbitrary values that would not provide value to models. Country is a very sparsely populated column within the data set, so the appearance of a country value is an anomaly compared to any other given example. Since we are focusing on English sentiment data, the language feature does not provide value for our usage as all the data will be English language.

IV. SIMULATION RESULTS

We were able to plot the results we received from the MapReduce instance count into a number of figures. The data used is from the start of January 2022 to February 19,

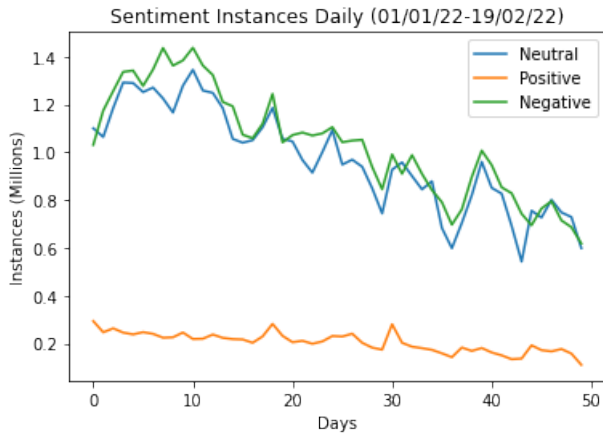


Fig. 2. Instances over time, daily.

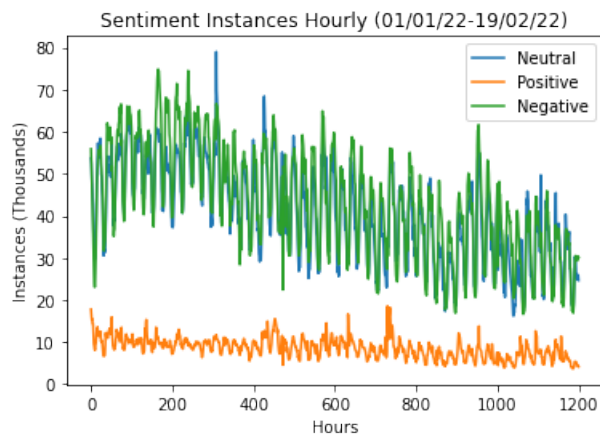


Fig. 3. Instances over time, hourly.

Total Sentiment Distribution, 01/01/22-19/02/22

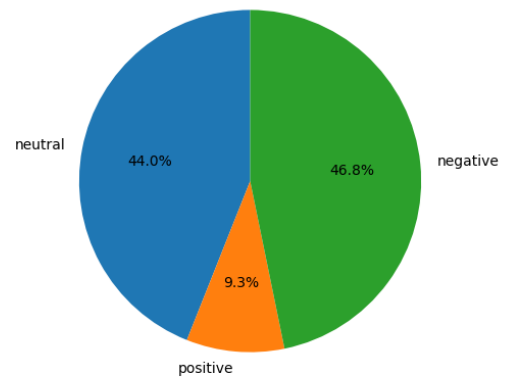


Fig. 4. Class instance distribution.

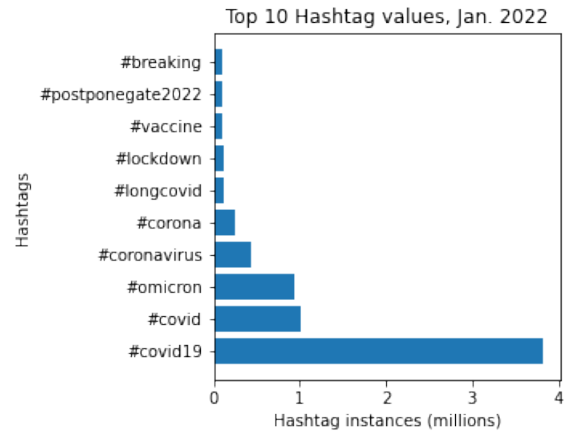


Fig. 5. Top 10 Hashtags.

2022. When viewing the daily instance counts in Fig. 2, we can see that there is a downward trend in total twitter activity relating to COVID-19. Furthermore, the negative and neutral class seem fairly correlated all throughout this time frame. The positive instances are much lower than the others throughout the entire time. The cause of the lower instance of positive tweets could be from people not being happy and hyped up about COVID-19 on twitter as it is an inherently negative thing. High amounts of neutral came from news sources possibly as news is generally written as a neutral piece of work, or anything that reports general information. Negative instances can come from the many different groups that complain about different COVID-19 related topics.

Within the hourly figure, Fig. 3, many peaks and valleys can be seen. This may just be the difference between Twitter prime and off hours, especially since we just used English data for this part. While many English speakers exist across all time zones, the areas with the highest amount of active English twitter users will influence these peaks and valleys such as North America, a highly populated typically English speaking region.

The total class distribution can be seen in Fig. 4. There

is an unequal distribution of the classes, especially between the positive class and the other classes. The class imbalance could pose a problem to the machine learning methods, since the number of examples for positive is only at 9.3% of the total data set for the time examined. Neutral and negative are pretty close to being balanced between each other with 44.0% and 46.8% respectively.

We were able to gather the results of the top seen values for hashtags, mentions, and NER values. We can derive possibly trending topics for the month of January where these values were from.

Hashtags can be good values to gather data from. Showcased in Fig. 5, the top hashtags are all strictly COVID adjacent values. Perhaps indicating how much of the data from this data set was collected. For example, the majority of the values are #covid19, #covid, and #omicron. Perhaps the usage of less trending hashtags could gain some more value on the sentiment of subtopics within COVID-19 related tweets.

In Fig. 6, the top mentioned accounts can represent highly engaged figures within the community that is active in

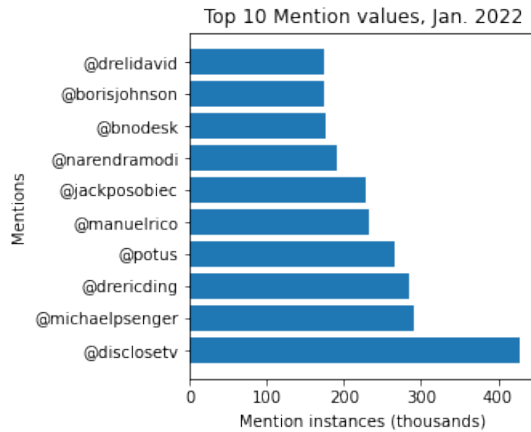


Fig. 6. Top 10 Mentions.

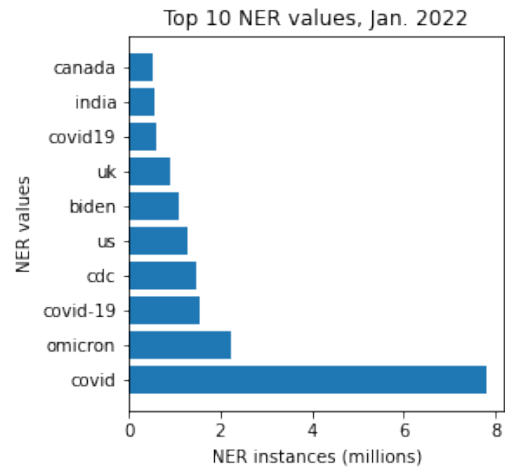


Fig. 7. Top 10 NER values.

COVID-19 twitter. As to whether these accounts foster negative or positive communities that engage with them could be of interest for further analysis. All tweets that mention a certain account could be counted for their sentiment value and the accounts can be deemed as positive or negative figures. Government officials can be found high up in the mentioned accounts due to the possibility that many people are likely to try and voice their opinions to them as they hold power in some COVID-19 related policies generally. The second highest mentioned account is suspended, suggesting that they were spreading information of questionable quality that would end up against Twitter’s terms of service. Other mentions include some health officials or experts that may have high interaction due to their knowledge.

The top NER values are presented in Fig.7, and show obvious values that would be trending among COVID-19 related tweets, but beyond that, we can find sub-topics that were of great importance for the month. For example, the values US, India, UK and Canada can indicate significant events within those nations that were related to COVID-19 in January. The CDC and Biden values could indicate important events relating to certain government entities. Now as to whether these values had negative or positive relations with the tweet, that would have to be counted to see the ratios of positive, negative, and neutral with the named entities.

Space was a constraint that made using just one month of the data set very difficult to run machine learning on. With limited resources, only the first seven days of 2022 were used when machine learning was done.

First, correlations between some of the features were investigated. Since lots of the data was categorical, it needed to be converted to numerical values to find correlation. Some of the features were not converted to numerical values, like hashtags, NER values, and mentions, since there were so many unique values in the data. Having so many unique categories would make seeing any relationship unlikely since the categories, at least in this case, are not ranked. For most features, not much correlation was found. There was, however,

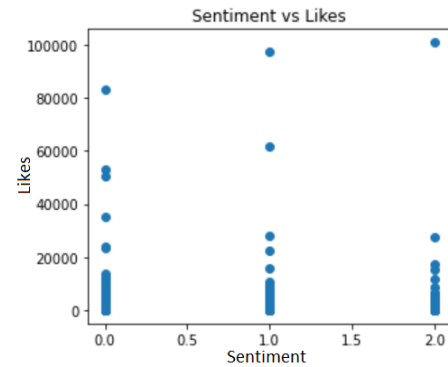


Fig. 8. Sentiment against likes.

a considerable amount of correlation between "Sentiment", "Logits Neutral", "Logits Positive", and "Logits Negative", which is to be expected since the feature "Sentiment" is found from the three "Logits" features. These relationships are not useful in developing any sort of model.

Some of the features were plotted against "Sentiment" to visualize their relationship. Sentiment was Label Encoded to numeric values (0 = Negative, 1 = Neutral, 2 = Positive). Each category had many tweets with few retweets or likes and some outliers that greatly exceeded the range of the majority of the tweets. It seemed most tweets with Sentiments of "Positive" had fewer likes and retweets, not exceeding the main clump of points. There was, however, one extreme positive sentiment outlier,

Fig. 8 and Fig. 9, that could have been the result of a COVID restriction being lifted, for instance. Tweets with the most retweets seemed to belong to the "Neutral" category, Fig. 9, while tweets with the most likes seemed to be "Negative", Fig. 8.

Using only likes and retweets, a K-Nearest Neighbors algorithm was used to predict the sentiment of a tweet. Using different values for K, it was found that a minimum of

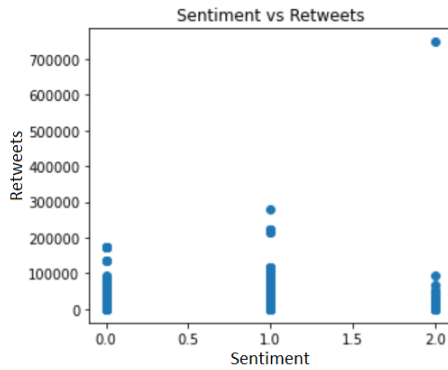


Fig. 9. Sentiment against retweets.

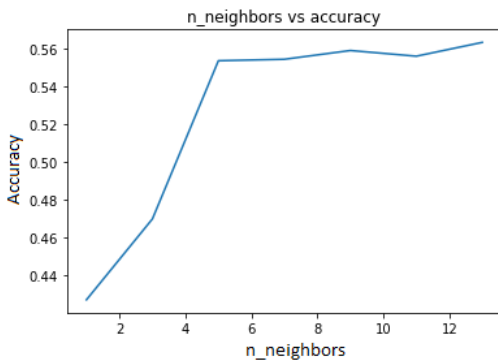


Fig. 10. K-Nearest Neighbours with different values.

K=5 provided the best accuracy scores, Fig. 10. This model, however, would not be very useful since its accuracy score is not very high.

We can derive from the results of the K-Nearest Neighbors (KNN) classification results, Fig. 10, with just likes and retweets that the features are not that useful on their own. Rather than continuing testing with these, we switched to using the vectored text from the generated text column that we made. Seeing Fig. 11, the results are much better, but still not possibly the best that could possibly have been accomplished.

Out of all the nine classifiers tested, two of them performed the worst with low 50-55% accuracy on the test set. Those classifiers were the Support Vector (SVC) and the Decision Tree. Along with being perhaps a bad model for the task, the SVC took a long time to classify examples and train compared to some of the better performing models.

The rest of the classifiers scored around 70-75% on the test set that was used. The Multi-Layer Perceptron (MLP) performed the best on the test set for accuracy above all else with 76.6% accuracy. The best accuracy did come at a cost, and that was training time. The MLP took around 3 hours to train, compared to other top models which all finished in about a minute at most. Due to the time the MLP took to train, if we were to be doing this training on live data hourly we would probably want to choose one of the faster methods such as Logistic Regression. The time to train the MLP could

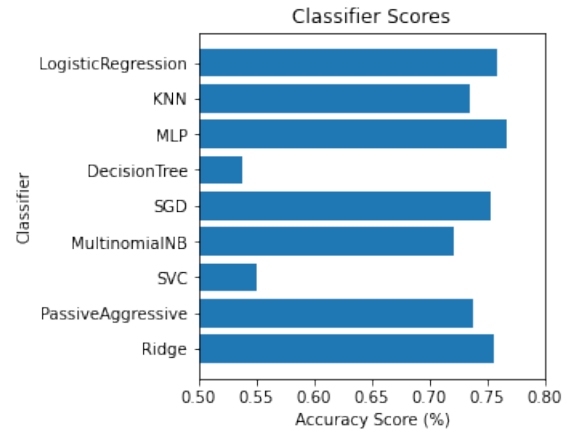


Fig. 11. Results of various classifiers.

be alleviated by using a more complex neural network library that allows for GPU acceleration as the library we were using, sci-kit learn, does not support it.

The Ridge, PassiveAggressive, SGD, and Logistic Regression classifiers all performed similarly and this may be due to the fact that they are all linear models.

As to why KNN performed much better on the text data compared to the likes and retweets, the reason can be that there is much more data to compare. Tens of thousands of words compared to only a couple of columns. The difference between a negative tweet and a positive tweet can be more easily distinguished as similarity based on the euclidean distance of any two examples.

V. DISCUSSION

As previously discussed, our best result for predicting the sentiment of a given tweet came from classifying tweets using vectored text. Even with the tested accuracy being acceptable, there are improvements that could be made. Considering the size of the data set being used, surely some trends would present themselves and *some* have. However, limiting the data that was used in machine learning, to a much smaller chunk than what was available, will have some effects on the model produced. Depending on what was happening with respect to the pandemic during the time frame that the data was taken from, it could only be effective at predicting the sentiment of a tweet during that time. The better approach(es) could be to use a larger sample of data, take small samples from different times during the pandemic, or both. Either solution could provide a more diverse set of tweets to design a model based off of, which could possibly mean a more accurate model to predict a tweets sentiment.

VI. CONCLUSION

Our paper is able to figure out many characteristics in the data using big data methods such as MapReduce and a number of machine learning methods for sentiment analysis. The counting algorithms are able to determine instance counts of various features within the data, providing insight

into possible popular trends. The machine learning methods proposed have been fairly successful at determining sentiment when trained on the synthesized text and could be improved by perhaps using more complex neural networks beyond a MLP. Businesses could potentially use the results found to direct their advertisements towards specific audiences based on trending topics and sentiment towards the pandemic found in tweets.

REFERENCES

- [1] F. Es-Sabery *et al.*, "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," in *IEEE Access*, vol. 9, pp. 58706-58739, 2021, doi: 10.1109/ACCESS.2021.3073215.
- [2] S. Shayaa *et al.*, "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges," in *IEEE Access*, vol. 6, pp. 37807-37827, 2018, doi: 10.1109/ACCESS.2018.2851311.
- [3] Sharef, Nurfadhina Mohd, Harnani Mat Zin, and Samaneh Nadali. "Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data." *J. Comput. Sci.* 12.3 (2016): 153-168.
- [4] Albaldawi, W. S. and Almuttairi, R. M., "Comparative Study of Classification Algorithms to Analyze and Predict a Twitter Sentiment in Apache Spark", in *Materials Science and Engineering Conference Series*, 2020, vol. 928, no. 3, p. 032045. doi:10.1088/1757-899X/928/3/032045.
- [5] Lopez, Christian E and Gallemore, Caleb, "An augmented multilingual Twitter dataset for studying the COVID-19 infodemic," in *Social Network Analysis and Mining*, vol. 11, no. 1, p. 1-14, 2021, doi:10.1007/s13278-021-00825-0.

Use of Drones (UAVs) for Pollutant Identification in the Industrial Sector: A Technology Review

Deyby Huamanchahua
*Department of Electrical and Mechatronics
 Engineering*
 Universidad Ingeniería y Tecnología - UTEC
 Lima, Perú
 dhuamanchahua@utec.edu.pe

Julio C. Huamanchahua
Facultad de Ingeniería
 Universidad San Ignacio de Loyola
 Lima - Perú
 julio.huamanchahua@usil.pe

Fabiola Fanny-Flores
Facultad de Ingeniería
 Universidad San Ignacio de Loyola
 Lima - Perú
 fabiola.floresi@usil.pe

Abstract— Unmanned aerial vehicles (UAVs) began to be used at the beginning of the 20th century, but only for military purposes; however, nowadays their use has become popular in other areas such as mining, construction, security, agriculture, environment, and for research purposes. Therefore, the purpose of this article is to document a systematic review of the use of UAVs related to environmental science issues. The objective of this article is to identify the different uses of unmanned aerial vehicles (UAVs) in an environmental system that will allow the identification, quantification, and data analysis of different variables related to the area of interest such as the concentration of pollutants found within the study area. For the procedure of this article, we used different databases and scientific search engines that allowed us to obtain research conducted from 2019 to 2021. The selection of the reviewed articles was carried out, leaving 25 selected. Finally, it is concluded that more research can still be developed on environmental science topics but focused on flora for its detection, identification, quantification, or monitoring.

Keywords—environmental monitoring, drone, detection, monitoring

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) have different components such as hardware and software necessary for real-time control, controllers such as altitude and heading controllers used to determine altitude and flight respectively, and sensors that are specifically based on what you want to measure, among others [1]. It also allows the collection of data or information related to the detection, supervision, environmental monitoring, and capture of 3D satellite images that favor the quantification and identification of some species [2]. Therefore, in the last decade, there has been a growing popularity for unmanned aerial vehicles (UAVs), increasing interest in the development of new types of drones of different sizes so that they can perform more tasks, which has led to their application in different areas [3].

The main uses that are currently being given to drones, are used for patrols and even surveillance systems in cities, they are also taken as tools in civil engineering to evaluate topographic measurements, in the environmental use are made as surveillance and control systems to identify and counteract the negative impacts caused by pollution, in addition, these UAVs

help to record aerial shots to observe the evolution of vegetation after a fire, in agriculture UAVs control and monitor through images captured by cameras to observe the state of crops, therefore, the use of drones allows or helps systematically, accurately, quickly and economically for monitoring works and research studies [4].

On the other hand, environmental monitoring with drones is an important tool to determine pollutant concentrations in the environment allowing constant data collection using sensors in real-time helping to determine and predict the quality of air, water, and noise. Therefore, drones provide a new opportunity for monitoring potential negative impacts on the environment [5]. In other words, by incorporating an unmanned aerial vehicle (UAV) into a monitoring system, it will be possible to collect data practically at higher altitudes and in areas that are difficult for humans to access.

An example of this is the use of UAVs for monitoring gases such as CO₂ in mining deposits, which are obtained through regression methods and neural networks to generate a mapping of the concentration of this chemical compound in the environment [6]. Such use allows for reducing variables in terms of cost and time of analysis to carry out a study. This would not have been possible at the beginning of the 20th century since its use was exclusively for military matters [7]. Likewise, another study mentions the use of drones in the mining sector for the monitoring of particulate matter (PM), which is generated by the flying that is carried out constantly in the deposits. This measurement was carried out thanks to the incorporation of a dust sensor in the UAV, also allowing the collection of data during the flight, facilitating the characterization of the plumes of each exploitation in real-time [8].

The objective of this article is to review the use of drones or unmanned aerial vehicles (UAVs) integrated into the area of environmental sciences where they perform the function of identification, quantification, analysis, monitoring, supervision, and detection of environmental variables of interest. All the information collected will be captured in a table employing certain characteristics. Finally, conclusions will be presented based on recent trends in the application of drones around interest.

II. METHODOLOGY

For the development of this research a search was conducted in the following databases: Google Scholar, Refseek, and ProQuest where keywords such as drone, environmental monitoring with drones, supervision, and detection were used, obtaining more than eight thousand studies until the year 2021, but we considered only those focused on the branch of environmental sciences and on the other hand that also provide information on the name of the drone, type of drone, sensors, type of pollution, type of action of the drone and type of industry, which considerably reduced the number of articles to be treated. Another selection criterion was access to scientific

articles, of which just over 30 research papers could not be accessed. Finally, the criteria reduced the number of studies to only 25 studies to be published between 2019 and 2021.

III. REVIEW OF PUBLICATIONS

To simplify and structure the research conducted. The information collected related to drones with a focus on environmental sciences was classified according to drone name, type of drone, sensors, type of contamination, type of drone action, and type of industry. Table I shows the collection of information

TABLE I. DRONES FOR POLLUTANT IDENTIFICATION IN THE INDUSTRIAL SECTOR

Reference	Name of drone	Type of drone	Sensors	Type of Contamination	Type of drone action	Type of Industry
Kuantama, E (2019) [9]	NE	QDC	Carbon	Air	Monitoring	ENV
Naughton, J (2019) [10]	Matrice 100	QDC	DZX3V, ZX3, DZXTR	Air	Monitoring	ENV
Resop, J (2019) [11]	Vapor 35	OnP	YSS, Core: C3D	Soil	Monitoring	ENV
Stavarakoudis, D (2019) [12]	Phantom 4	QDC	PSIM, S, SIMLI	NE	Detection	Agrarian
Buters, T (2019) [13]	P4P	QDC	C20	Soil	Monitoring	Ecology
Duan, Z (2019) [14]	M600P	HXC	SF, LDOCP	Water	Monitoring	ENV
Oviedo, B (2019) [15]	MAVIC PRO	QDC	MQ7	Air	Monitoring	ENV
Simo, A (2019) [16]	NE	NE	SGMP	Air	Monitoring	ENV
Godall, R (2020) [17]	E-DRONE	QDC	BMP085, GPS, XP, HC-SRO4, SG	Air	Monitoring	ENV
Xiaojung, L (2020) [18]	Multicopter	NE	SGPM	Air	Monitoring	ENV
Brinkman, J (2020) [19]	M600P	HXC	Anemometer	Air	Monitoring	ENV
Hugh, L (2020) [20]	Matrice DJI	HXC	MHC, COGI, SCH4	Air	Monitoring	ENV
Cagnazzo, C (2021) [21]	P4P y I2	NE	CT, CFPV, RGB	Water	Detection	ENV
Bukin, O (2021) [22]	NE	QDC	SOIA	Water	Monitoring	ENV
Barreto, J (2021) [23]	Mavic 2 Zoom	QDC	Camera	NE	Monitoring	ENV
Carabassa, V (2021) [24]	Atmos-7	Flexible Fixed Wings	CMSM	Soil: erosion	Monitoring	ENV
Lan, H (2021) [25]	Geodrone X4L	QDC	COV sensor	Air	Detection	ENV
de Facio, R (2021) [26]	Phantom 3	QDC	SGPM	Air	Detection	ENV
Burgues, J (2021) [27]	M600P	HXC	SEMOX	Air	Monitoring	ENV
Madokoro, H (2021) [28]	M600P	HXC	MS, Sensor PM2.5	Air	Monitoring	ENV
Tovar, A (2021) [29]	Condor, Dronetools	HXC	CM2, DLS	Water	Detection	ENV

Casallas, A (2021) [30]	NE	QDC	SMQ, SDL	Air	Monitoring	ENV
Ríos, R (2021) [31]	Dron Multirotor	QDC	M, CTL	Crops	Monitoring	ENV
Shin, Y (2021) [32]	NE	NE	SPM	Air	Monitoring	ENV
Cheng, W (2021) [33]	M600P	MTC	S03BC	Air	Monitoring	ENV

Note: Abbreviation: P4P: Phantom 4 pro, M600P: Matrice 600 Pro, I2: Inspire 2, DZX3V: DJI Zenmuse X3 visual, ZX3: Zenmuse X3, DZXTR: DJI Zenmuse XTR, YSSC: Yellow Scan Surveyor, C3D: cartography in 3D, PSIM: Parrot @Multispectral imaging sensor, S: Sequoia, SIMLI: Irradiance sensor for incident light measurement, C20: 20-megapixel camera red-green-blue, SF: Fluorescence Sensor, LDOCP: Continuous wave (cw) diode laser for petroleum, SGMP: sensors to PM10, PM2.5, SO₂, NO₂, CO, O₃, VOC, CO₂, XP: Xbee Pro 900HP, SG: Sensors to CO₂, CO, NH₃, SO₂, PM, O₃ y NO₂, SGPM: Sensors to NO₂, SO₂, O₃, CO, PM2.5 y PM10, MHC: Mirage HC, SCH4: Methane Sensor, CT: thermographic camera FLIR C3, CFPV: Camera FPV, SOIA: Optical sensor and artificial intelligence, CMSM: camera MicaSense RedEdge-M with a professional multispectral sensor, SEMOX: Electrochemical and metal oxide sensors, MS: Multi-sensor, CM2: Dual multispectral camera, DLS: Downward light sensor, SMQ: Sensor MQ, SDL: Digital laser sensor for PM1.0, PM10 y PM 2.5, M: Multi-spectral RGB, CTL: Thermal Lidar camera, SPM: sensor of PM 2.5 and PM 10, S03BC: sensor of O3 and BC, Environmental: ENV, Quadcopter: QDC, Hexacopter: HXC, One Propeller: OnP, Multicopter: MTC.

A. Drone

Nowadays, drone technology is very familiar and versatile. The drone can be equipped with different systems allowing it to perform different types of tasks. Nowadays drones are used in warfare, mine inspection, surface reconnaissance, and research, i.e., drones not only help in society, but they are also a threat. So, drones have their predetermined working capability, so it becomes important in today's world [35].

From the information gathered, it has been possible to identify a wide variety of unmanned aerial vehicles that are used for environmental monitoring, identification, quantification, and surveillance systems. Table II describes the characteristics of the drones most used for this purpose.

TABLE II. DESCRIPTION OF THE CHARACTERISTICS OF THE MOST COMMONLY USED DRONES

Drone	Characteristics
M600P	The Matrice 600 Pro has a payload of 6 kg and a flight time of 15 minutes and has a slow flight speed, therefore, these characteristics are essential for good monitoring of emission sources [05].
P4P	This unmanned aerial vehicle has a maximum flight time of 60 minutes and detects obstacles in 5 directions in mid-flight.
E-Drone	The E-Drone are automatic drones that are programmed to perform monitoring and detection at heights that are not above ground level [06].
P3	This drone has a transmission distance of 1 kilometer, therefore, it easily allows the analysis or monitoring in places that may be inaccessible or dangerous for humans [34].

Note: M600P: Matrice 600 Pro, P4P: Phantom 4 Pro, P3: Phantom 3

On the other hand, Table III shows the percentage of use of the drones where the Matrice 600 Pro drone is the one that has had more uses in the list of the mentioned investigations representing 19.20%, followed by the Phantom drone, it is worth mentioning that within this section three different Phantom models were included which were the P4, P4P, P3 and represented 15.40%. While in the other section, drones that were mentioned only once were considered and accounted for 50.00% of the total. Finally, drones that do not specify their names (NE) in the numbered research papers are detailed, accounting for 15.40% of the total.

TABLE III. PERCENTAGE OF DRONE USE

Drone	Amount	
	Frequency	Percentage
M600P	5	19,20%
Phantom	4	15,40%
Others	13	50,00%
NE	4	15,40%

Note: M600P: Matrice 600 Pro, NE: Not specified

B. Type of Drones

One of the most prominent features is the one used to define the difference between fixed-wing systems, multirotor systems, and other systems. There are hybrid systems, which are both multi-rotor and fixed-wing systems, ornithopters, and drones using turbofans. The technology used to maintain the drone's flight defines the type of drone, this characteristic is also the determining factor of the drone's shape and appearance. A second characteristic is the level of autonomy of the drone. The autonomy can vary from fully autonomous operation to one controlled by a remote pilot. Another notable characteristic is the difference in size between drones. The size can vary from drones the size of an insect to drones the size of a commercial aircraft. Weight is also an important characteristic. The weight of drones can vary from several grams to hundreds of kilograms. The last characteristic is the difference in power sources [36]. From one of the characteristics, all kinds of drones can be differentiated by the number of propellers, such as quadcopters (drones with four propellers), hexacopters (drones with six propellers), and fixed-wing drones, single-propeller drones, and multicopters.



Fig. 1. E-DRONE / Quadricopter. Source: Godall Rohi, O'tega Ejofodomi, Godswill Ofualagba, Autonomous monitoring, analysis, and countering of air pollution using environmental drones, Heliyon, [2020]

The drones that have more propellers are more used because they allow a better flight and therefore data collection at higher altitudes. As can be seen in Table IV, the quadcopter and the hexacopter are the most used drones, which represent more than

80.00%, while the rest of the drones (one propeller, multicopter, and flexible fixed wings) are below 5.00%.

TABLE IV. PERCENTAGES OF DRONE TYPES

Type of Drone	Amount	
	Frequency	Percentage
CP	12	54,54%
HC	6	31,81%
UH	1	4,55%
AFF	1	4,55%
MC	1	4,55%

Note: CP: Quadcopter, HC: Hexacopter, UH: One Propeller, AFF: Flexible Fixed Wings, MC: Multicopter

C. Sensors

Designing sensor systems for drone detection is often easier said than done; these systems include weather, visibility, false positives, and bird or aircraft detection. To have better information, alternative approaches have been implemented, such as the use of optical sensors. Optical sensors have been used for various types of object detection and have been successful in UAV detection [37].

Similarly, another alternative has been to use radio frequency (RF) scanners. Several studies have been conducted using RF scanners to detect Micro Doppler signatures to perform drone detection and classification [38].

This section reviews the devices or sensors that accompany the drone to perform the corresponding study. There are different types of sensors and each of them determines a type of pollutant based on concentrations as shown in Figure 2: (a): Gas Sensor, (b): NO₂ Sensor, (c): NH₃ Sensor.

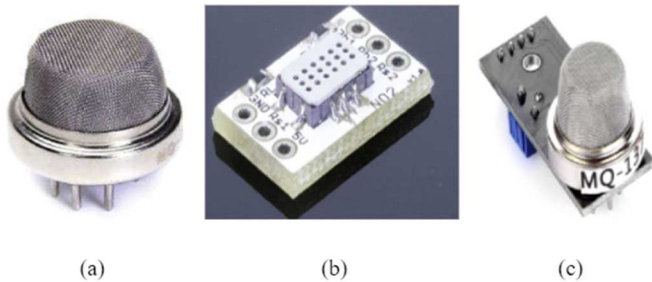


Fig. 2. These types of sensors were programmed to automatically recalibrate themselves every week by collecting statistical data in studies related to environmental sciences. Source: Godall Rohi, O'tega Ejofodomi, Godswill Ofualagba, Autonomous monitoring, analysis, and countering of air pollution using environmental drones, Heliyon, [2020]

Table V also mentions sensors that allow determining the concentrations of PM, CO, CO₂, NO, and other gases that negatively impact the environment.

TABLE V. PERCENTAGES OF SENSORS TYPES

Type of Sensors	Amount	
	Frequency	Percentage
Carbon	1	2.5%
SGMP	1	2.5%
SG	1	2.5%
MHC	1	2.5%

SCH4	1	2.5%
SEMOX	1	2.5%
MS	1	2.5%
SMQ	1	2.5%
SDL	1	2.5%
SPM	1	2.5%
S03BC	1	2.5%
SGPM	1	2.5%

Note SGMP: sensors to PM10, PM2.5, SO₂, NO₂, CO, O₃, VOC, CO₂, SG: Sensors to CO₂, CO, NH₃, SO₂, PM, O₃ y NO₂, SGPM: Sensors to NO₂, SO₂, O₃, CO, PM2.5 y PM10, MHC: Mirage HC, SCH4: Methane Sensor, SEMOX: Electrochemical and metal oxide sensors, MS: Multi-sensor, SMQ: Sensor MQ, SDL: Digital laser sensor for PM1.0, PM10 y PM 2.5, SPM: sensor of PM 2.5 and PM 10, S03BC: sensor of O₃ and BC.

Likewise, the sensors can also be RGB, thermal, lidar, and multispectral cameras as shown in Fig. 3, and are used for detecting, monitoring, and identifying the study areas.

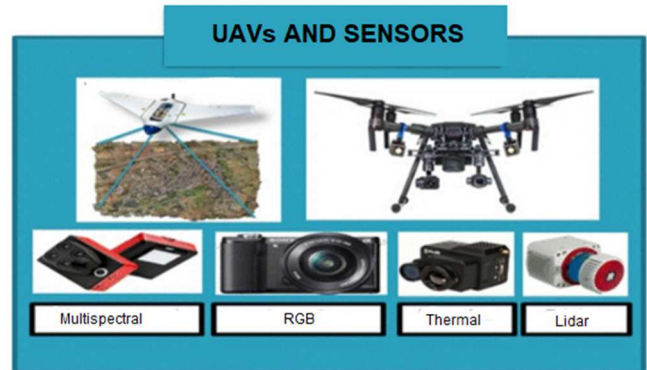


Fig. 3. Unmanned aerial vehicle and sensors. Source: R. Ríos, Hernández, Use of drones or Unmanned Aerial Vehicles in precision agriculture, Cuba: Agricultural Engineering Magazine, [2021]

D. Type of Contamination

Some of the main air pollutants are ozone (O₃), particulate matter (PM), nitrogen dioxide (NO₂), carbon dioxide (CO₂), sulfur dioxide (SO₂), and carbon monoxide (CO). Some of these pollutants occur naturally in the environment and others are generated through human activities [39].

PM pollution is a part of air pollution that is composed of extremely small particles and liquid droplets containing acids, organic chemicals, metals, and soil or dust particles. In addition, a high concentration of NO₂ can irritate the airways of the respiratory system and can aggravate and cause the development of respiratory diseases. Additionally, a high concentration of CO₂ is partly responsible for global warming. Exposure to ammonia (NH₃) can cause immediate burns to the nose, throat, and respiratory tract. Air pollution also has a detrimental effect on climate and crops [39].

In this field, the aim is to show in which medium the contamination (if any) develops. Consequently, Table VI shows three different media: air, water, and soil, which have 15, three, and two studies, respectively. In addition, four other studies were also found that are not based on contamination but the evaluation or monitoring of the state of the crops [9-28], irrigation programs [28], or land topography modeling [8].

TABLE VI. PERCENTAGE OF STUDIES BASED ON THE USE OF DRONES IN DIFFERENT ENVIRONMENTAL MATRICES

Matriz	Amount	
	Frequency	Percentage
Air	15	65,22%
Water	3	13,04%
Soil	2	8,70%
SC	4	17,39%

Note: SC: Without contamination refers to the use of the drone to evaluate the state of the crops or to support irrigation programs, therefore contaminants are not evaluated as in the other cases.

E. Type of action

The exponential increase in the public accessibility of drones has posed a major threat to overall security and confidentiality. Drone sales have been steadily increasing every year and are expected to be much higher in the future [40]. Currently, drones have many uses for different types of activities.

Both monitoring and detection are the types of action present in 48% and 20% respectively of the different research papers. On the other hand, in smaller proportion are two works on monitoring-detection (MD) and one work-related to detection-supervision (DS) which represent 8% and 4% respectively.

TABLE VII. PERCENTAGES OF STUDIES BASED ON THE DRONE ACTION TYPE

Type of Action	Amount	
	Frequency	Percentage
Monitoring	12	48,00%
Detection	5	20,00%
MD	2	8,00%
DS	1	4,00%

Note: MD: monitoring - detection, and DS: detection - supervision.

F. Type of Industry

There is research on the technical capabilities of drone technology, but much less on the practical application of drones in industry. At present, there is a gap between the technological developments of drones and the applications they can offer to the industry. This is due to the novelty of the technology and, in part, because drone development is for purposes other than manufacturing. Current industrial applications of drones are primarily outdoor. Their cost-effective applications in sectors such as agriculture, construction, infrastructure, energy, logistics, and mining take advantage of their ability to fly quickly and safely at high altitudes to locations that are difficult, dangerous, or costly to access. Manufacturing on the other hand, manufacturing operations is almost exclusively performed indoors [41].

Of the studies reviewed, there is evidence of a high representation of environmental issues, which accounts for 76% of the research, while works related to agriculture and ecology each account for 8% and, to a lesser extent, studies related to topography, which only account for 4%. Finally, there is one research work that does not specify (NE) the type of industry to which it is related and represents 4%.

TABLE VIII. PERCENTAGE OF STUDIES BASED ON INDUSTRY TYPE

Type of Industry	Amount	
	Frequency	Percentage
Environmental	19	76,00%
Agriculture	2	8,00%
Ecology	2	8,00%
Topography	1	4,00%
NE	1	4,00%

Nota: NE: Not specified.

IV. CONCLUSIÓN

It is concluded that drones are on the verge of being adopted for use in many industries. The most promising and cost-effective applications are those that aid in the inspection of hard-to-reach or hazardous areas, gas leak detection in large plants, and cycle counting in large warehouses. The findings in the results show that drones could have greater potential in process industries than in discrete manufacturing.

The advancement of technology has also allowed the improvement of drones and their components, especially for unmanned aerial vehicles to travel in areas that are difficult for humans to access. Therefore, their use in recent years has intensified in different areas and the environmental sciences sector has not been the exception. However, there are still scenarios in which studies have not yet been carried out, such as those related to the detection, quantification, or identification of flora as well as the identification of wild species employing the sound they emit. Finally, there is still room for researchers to conduct more studies related to this topic.

REFERENCES

[1] V. Kangunde, R. Jamisola, and E. Theophilus, "A review on drones controlled in real-time," *Int. J. Dynam. Control* vol. 9, pp. 1832–1846, 2021.

[2] J. Gómez, and A. Tascón, "A protocol for using unmanned aerial vehicles to inspect agro-industrial buildings," *Informes De La Construcción*, vol. 73 (564), e421, 2021.

[3] I. Jawhar, N. Mohamed, J. Al-Jaroodi, D. Agrawal, and S. Zhang, "Communication and networking of UAV-based systems: Classification and associated architectures," *Journal of Network and Computer Applications*, vol. 84, pp. 93-108, 2017.

[4] R. González, J. Uacán, I. Sánchez, R. Medina, F. Árcega, C. Zetina, and R. Casares, "Drones. Aplicaciones en ingeniería civil y geociencias," *Interciencia*, vol. 44 (6), pp. 326-331, 2019.

[5] Y. Zhang, and P. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63-72, 2021.

[6] J. Eslava, F. Martínez, A. Soto, E. Vera and D. Guevara, "Vehículos aéreos no tripulados como alternativa de solución a los retos de innovación en diferentes campos de aplicación: una revisión de la literatura," *Investigación e Innovación en Ingenierías*, vol. 9(1), pp. 149-166, 2021.

[7] M. Guevara, A. Meza, E. Esquivel, D. Arias, A. Tapia, and F. Masís, "Uso de vehículos aéreos no tripulados (VANTs) para el monitoreo y manejo de los recursos naturales: una síntesis," *Revista Tecnología En Marcha*, vol. 33(4), pp. 77–88, 2020.

[8] M. Alvarado, F. Gonzalez, A. Fletcher, and A. Doshi, "Towards the Development of a Low-Cost Airborne Sensing System to Monitor Dust Particles after Blasting at Open-Pit Mine Sites," *Sensors*, vol.15, pp. 19667-19687, 2015.

[9] E. Kuantama, R. Tarca, S. Dzitac, I. Dzitac, T. Vesselenyi, and I. Tarca, "The Design and Experimental Development of Air Scanning Using a Sniffer Quadcopter," *Sensors*, vol. 19(18), pp. 3849, 2019.

[10] J. Naughton, and W. McDonald, "Evaluating the Variability of Urban Land Surface Temperatures Using Drone Observations," *Remote Sensing*, vol. 11(14), pp. 1722, 2019.

[11] J. Resop, L. Lehmann, and W. Cully, "Drone Laser Scanning for Modeling Riverscape Topography and Vegetation: Comparison with Traditional Aerial Lidar," *Drones*, vol. 3(2), pp. 35, 2019.

- [12] D. Stavrakoudis, D. Katsantonis, K. Kadoglidou, A. Kalaitzidis, and I. Gitas, "Estimating Rice Agronomic Traits Using Drone-Collected Multispectral Imagery," *Remote Sensing*, vol. 11(5), pp. 545, 2019.
- [13] T. Buters, D. Belton, and A. Cross, "Seed and Seedling Detection Using Unmanned Aerial Vehicles and Automated Image Classification in the Monitoring of Ecological Recovery," *Drones*, vol. 3(3), pp. 53, 2019.
- [14] Z. Duan, Y. Li, J. Wang, G. Zhao, and S. Svanberg, "Aquatic environment monitoring using a drone-based fluorosensor," *Appl. Phys.*, vol. 125, pp. 108, 2019.
- [15] B. Oviedo, L. Huacón, and J. Rosales, "Sistema electrónico para la detección de niveles de monóxido de carbono (CO) en la Av. 7 de octubre de la ciudad de Quevedo, que facilite la toma de decisiones del departamento de medio ambiente del G.A.D. municipal," *Iberian Journal of Information Systems and Technologies*, pp 1-8, 2019.
- [16] A. Simo, S. Dzitac, F. Friguer, S. Musuroi, P. Andea, and D. Meianu, "Technical Solution for a Real-Time Air Quality Monitoring System," *International journal of computers communications & control*, vol. 15, 2020.
- [17] G. Rohi, O. Ejojodomi, and G. Ofualagba, "Autonomous monitoring, analysis, and countering of air pollution using environmental drones," *Heliyon*, vol. 6, 2020.
- [18] L. Xiaojun, W. Hongyuan, L. Ao, X. Di, and H. Zhengxuan, "Research of multi-rotor UAV atmospheric environment monitoring system based on 4G network," *E3S Web of Conferences*, vol. 165, 2020.
- [19] J. Brinkman, B. Davis, and C. Johnson, "Post-movement stabilization time for the downwash region of a 6-rotor UAV for remote gas monitoring," *Heliyon*, vol. 6, 2020.
- [20] H. Li, M. Mundia, M. Reeder, and N. Pekney, "Gathering pipeline methane emissions in utica shale using an unmanned aerial vehicle and ground-based mobile sampling," *Atmosphere*, vol. 11 (7), 2020.
- [21] C. Cagnazzo, E. Potente, H. Regnaud, S. Rosato, G. Mastronuzzi, "UAV/UGV System for meso-macro pollutants identification in the beach environment," *Rend. Online Soc. Geol. It.*, vol. 55, pp. 29 - 35, 2021.
- [22] O. Bukin, D. Proshchenko, D. Korovetskiy, A. Chekhlenok, V. Yurchik, and I. Bukin, "Development of the Artificial Intelligence and Optical Sensing Methods for Oil Pollution Monitoring of the Sea by Drones," *Applied Sciences*, vol. 11, pp. 3642, 2021.
- [23] J. Barreto, L. Cajiaba, J. Teixeira, L. Nascimento, A. Giacomo, N. Barcelos, T. Fettermann, and A. Martins, "Drone-Monitoring: Improving the Detectability of Threatened Marine Megafauna," *Drones*, vol. 5, pp. 14, 2021.
- [24] V. Carabassa, P. Montero, J. Alcañiz, and J. Padró, "Soil Erosion Monitoring in Quarry Restoration Using Drones," *Minerals*, vol. 11, pp. 949, 2021.
- [25] H. Lan, J. Ruiz, Y. Leleev, G. Demaria, M. Jussila, K. Hartonen, and M. Riekkola, "Quantitative analysis and spatial and temporal distribution of volatile organic compounds in atmospheric air by utilizing drone with miniaturized samplers," vol. 282, 2021.
- [26] R. De Fazio, L. Dinoi, M. De Vittorio, and P. Visconti, "A Sensor-Based Drone for Pollutants Detection in Eco-Friendly Cities: Hardware Design and Data Analysis Application," *Electronics*, vol. 11, pp. 52, 2021
- [27] J. Burgués, M. Esclapez, S. Doñate, L. Pastor, and S. Marco, "Aerial Mapping of Odorous Gases in a Wastewater Treatment Plant Using a Small Drone," *Detección remota*, vol. 13, pp. 1757, 2021.
- [28] H. Madokoro, O. Kiguchi, T. Nagayoshi, T. Chiba, M. Inoue, S. Chiyonobu, S. Nix, H. Woo, and K. Sato, "Development of Drone-Mounted Multiple Sensing System with Advanced Mobility for In Situ Atmospheric Measurement: A Case Study Focusing on PM2.5 Local Distribution," *Sensors*, vol. 21, pp. 4881, 2021.
- [29] A. Tovar, A. Román, D. Roque, and D. Navarro, "Applications of unmanned aerial vehicles in Antarctic environmental research," *Sci Rep*, vol. 11, 21717, 2021.
- [30] W. Frasser, A. Casallas, O. Alvarez, M. Ballen, P. Barragan, D. Duarte, H. Rocha, J. Buesaquillo, and E. Ortega, "Uso de aeronaves no tripuladas para el monitoreo de la calidad del aire," vol. (15), 2021.
- [31] H. Ríos, "Uso de los Drones o Vehículos Aéreos no Tripulados en la Agricultura de Precisión," *Revista Ingeniería Agrícola*, vol. 11, pp. 75-84, 2021.
- [32] S. Yu, C. Chang, and C. Ma, "Simulation, and measurement of air quality in the traffic congestion area," *Sustain Environ Res*, vol. 31, pp. 26, 2021.
- [33] C. Wu, B. Liu, D. Wu, H. Yang, X. Mao, J. Tan, Y. Liang, J. Sun, R. Xia, J. Sun, G. He, M. Li, T. Deng, Z. Zhou, and Y. Li, "Vertical profiling of black carbon and ozone using a multicopter unmanned aerial vehicle (UAV) in urban Shenzhen of South China," *Science of the Total Environment*, vol. 801, pp. 689, 2021.
- [34] R. De Fazio, L. Dinoi, M. De Vittorio, and P. Visconti, "A Sensor-Based Drone for Pollutants Detection in Eco-Friendly Cities: Hardware Design and Data Analysis Application," *Electronics*, vol. 11, pp. 52, 2021.
- [35] B. Vergouw, H. Nagel, G. Bondt, and B. Custers, "Drone technology: Types, payloads, applications, frequency spectrum issues, and future developments," *En Information Technology and Law Series*, The Hague: T.M.C. Asser Press, vol. 27, pp. 21–45, 2016.
- [36] C. Aker, and S. Kalkan, "Using deep networks for drone detection." In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–6, 2017.
- [37] S. Samaras, E. Diamantidou, D. Ataloglou, N. Sakellariou, A. Vafeiadis, V. Magouliantitis, A. Lalas, A. Dimou, D. Zarpalas, and K. Votis, P. Daras, and D. Tzovaras, "Deep learning on multi-sensor data for counter UAV applications—a systematic review," *Sensors*, vol. 19, pp. 4837, 2019.
- [38] J. Anderson, J. Thundiyil, and A. Stolbach, "Clearing the air: a review of the effects of particulate matter air pollution on human health." *A. J. Med. Toxicol*, vol. 8, pp. 166-175, 2010.
- [39] M. Fernández, "La Profesión de Piloto de Drones en el ámbito del Patrimonio Cultural y la Arqueología: ciencia y divulgación desde el aire." *En Las Profesiones del Patrimonio Cultural*, pp. 75-80, 2018.
- [40] L. Escolada, L. Manganiello, M. Lopez, and C. Vega, "Los sensores químicos y su utilidad en el control de gases contaminantes," *Revista Ingeniería UC*, vol. 19, pp. 74-88, 2012
- [41] J. Gallardo, M. Pompa, C. Aguirre, P. López, A. Meléndez, "Drones: tecnología con futuro promisorio en la gestión forestal." Vol. 11, pp. 52-55, 2020.
- [42] J. Cornejo et al., "Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review," 2021 IEEE International IOT, Electronics, and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-9, doi: 10.1109/IEMTRONICS52119.2021.9422660.
- [43] D. Huamanchahua et al., "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [44] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo and J. Macuri-Vasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [45] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [46] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.
- [47] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE), 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.

Nursery with Automation and Control Systems to Produce White Chuño (Tunta)

Alem Huayta Uribe
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
 72758250@continental.edu.pe

Jalber Brayan Macuri Vasquez
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
 76611724@continental.edu.pe

Hitan Orlando Cordova Sanchez
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
 70811607@continental.edu.pe

Deyby Huamanchahua
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
 dhuamanchahua@continental.edu.pe

Abstract—For the production of white chuño or also called tunta, a suitable environment is needed concerning temperature since they must be produced in very cold temperatures. To create a greenhouse in an HMI interface in the LabVIEW program maintaining the production phases (first freezing, leaching in water, second freezing, and the peeling), where a belt with rollers that classifies the potatoes is used, also a temperature sensor that will help to measure this parameter in real-time, following it enters the tanks with turbines for the leaching phase and then the presses help the removal of the peels and excess water. In conclusion, the system allows to control in real-time the internal temperature of the greenhouse replicating the temperature that occurs in the Andes, the HMI interface helps with the monitoring and control of the processes that are being carried out, on the other hand, the results showed the potential that the system must produce white chuño.

Keywords—White Chuño (Tunta), frozen, greenhouse, leaching, automation

I. INTRODUCTION

Potatoes were domesticated in southern Peru approximately 10 000 years ago and since then have played a major role in human occupation in the highlands of the Andes [5]. The production of white chuño is traditionally done at very high elevations of the central Andes in the altiplano, this process is carried out in places where the temperatures are very cold, where during the nights the potatoes are frozen and the days are warm to have an optimal thawing, and make the bitter potatoes release their poisonous glycoalkaloids and can be consumed without danger to the organism [1-3], the sunlight is a very big enemy for the production because it makes them lose their most precious characteristic which is white color, turning it brown, thus losing its value and quality in the market. [2,4].

For the production of white chuño in an artisanal way, there are three processes to be carried out, once the bitter potatoes have been chosen and classified according to size, first, we freeze them for 3 to 4 nights at temperatures ranging from -4°C to -15°C. Secondly, we leach them in rivers for 21 to 30 days, Second, leaching is carried out in rivers for 21 to 30 days, third, a second freezing process is carried out for 1 to 2 nights at temperatures ranging from -4°C to -15°C, and finally, the peels are removed from the potatoes, which have already been converted into chuño and stored in jute or wool sacks [1,2].

Studies carried out propose, as in [5], the study of the concentrations of proteins, iron, zinc, and calcium, in the

processing of chuño, obtained results of the concentration of proteins and zinc, which has a higher content of calcium, while the concentration of iron does not change, it is also noted that the water used to prepare white chuño increases the concentration of calcium. In [7] liquid chromatography showed the presence of epicatechin, chlorogenic acid, gallic acid, syringaldehyde, and protocatechuic acid in potato samples as in chuño, the results suggest that the antioxidant capacity and phenolic compounds cannot be eliminated during the process, therefore chuño can be used as an antioxidant in diet.

Having a great motivation to develop this project is to have a chuño with a very favorable color which is white. The proposed study aims to create a greenhouse for the production of white chuño taking into account the phases of production in an artisanal way taking it to an automated system, using a Programmable Logic Control (PLC) to control the phases and all this presented in an HMI interface.

II. MATERIALS AND METHODS

For the development of the proposed product, the VDI-2206 methodology developed by "The German Engineers Association" [8], which consists of a standard for the development of mechatronic projects, which generates the connection of several disciplines and their operation, is considered.

The simulation of the automation and control process of white chuño processing will be carried out in LabVIEW software. To carry out the proposed system for the preparation of white chuño, the phases mentioned by the National Institute of Quality (INACAL) in its Peruvian Technical Standard NTP 011.401:2020 PAPA were considered. Tunta. Good artisanal processing practices. 2nd Edition. Where it mentions a series of phases, parameters, and care that should be considered in the processing of white chuño, which are presented in Table I [9].

Table I. White Chuño Production Phases.

Phase	Details	Parameters
Selection and/or classification	The process to clean the potatoes, to later select them according to shape and size.	None
First freezing	The potatoes are subjected to	(-4°C to -

	temperatures below zero, for approximately 3 to 4 nights.	15°C)
Leaching in water	The frozen potatoes are submerged in water for approximately 21 to 30 days, this phase is done in rivers.	None
Frozen second	The potatoes removed from the water are again subjected to sub-zero temperatures for a period of 1 or 2 nights.	(-4°C to -15°C)
Shelling	To carry out this stage, we take advantage of early mornings, since the tunta is wet and the peel is semi-detached.	None

The proposed mechatronic system has all the phases that can be seen in Table I, in the first phase there is a conveyor belt that carries the potatoes to a conveyor belt with rollers, the latter has the function of classifying the potatoes according to their size since the rollers of the belt are separated from each other at strategic distances, to drop the potatoes on a platform based on their proportion. Thus, obtaining the first tubers, which are the largest, second, which are the medium, and third, which are the small ones [2].

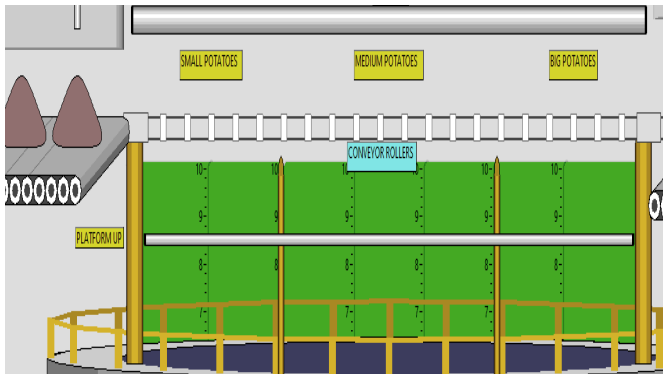


Fig. 1. Potato Size Selection Phases.

In the second phase, the greenhouse replicates natural frosts with an artificial freezing system, where the temperature is graduated according to the requirements shown in Table I, for which the proposed system has temperature sensors to monitor in real-time and constant for a given period, which is automatically and/or manually deactivated.

The third phase, which is leaching in water, merges with phase five, which is peeling because the greenhouse has a series of tanks where the potatoes are soaked. The tanks have a series of turbines that simulate washing machines that allow the water to move and make the potatoes pink and peel slowly throughout the leaching process.

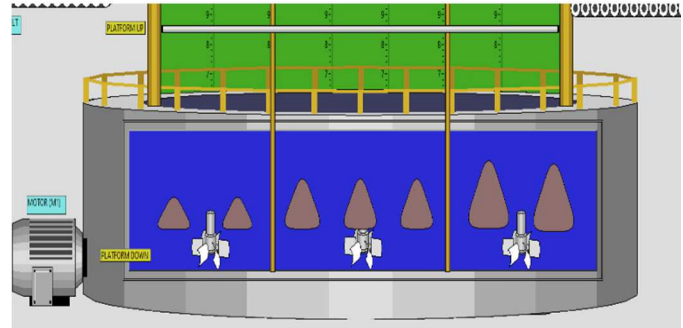


Fig. 2. Leaching and Peeling Stage.

In the fourth phase, which is the last stage, the second freezing is carried out in the same way as in the first freezing, but before proceeding with it, when the submerged platform rises again, the excess water is removed with a press that simulates the handmade footprints. Finally, the final product is removed by conveyor belts for storage and distribution.

On the other hand, the proposed automation system has two main control panels that allow its operation. One of them is directed to control the environmental parameters of the greenhouse and the other one is directed to control the leaching process. Fig. 3 shows that the control panel has an emergency stop system, led indicators that allow visualizing which system is active, as well as switches that allow turning on or off a system manually.

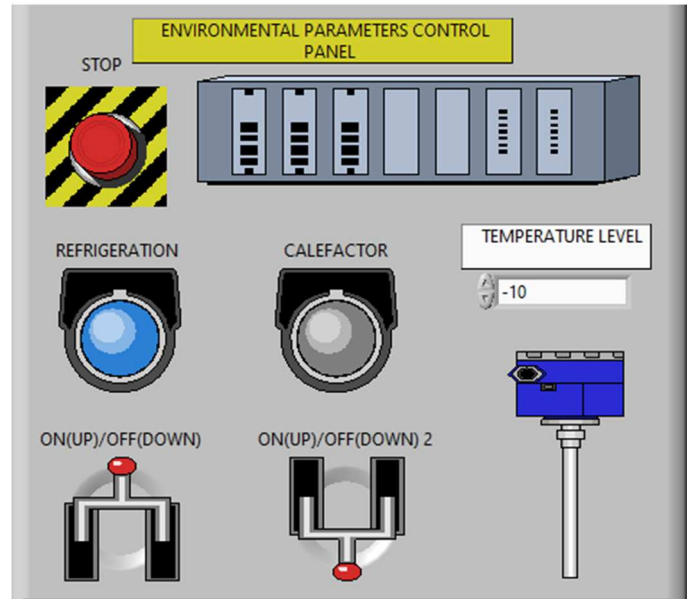


Fig. 3. Environmental Parameters Control Panel.

Fig. 4 shows the control panel to control and monitor the leaching process, it has an emergency stop button which is very important in any industrial process, led indicators to visualize different functions such as the motor, the location of the platform where the potatoes are located and finally the panel has push buttons to raise and lower the platform where the tubers (potatoes) are located manually if necessary, all this to

ensure an optimal process in case the mechatronic systems may fail for any error either mechanical and/or software.

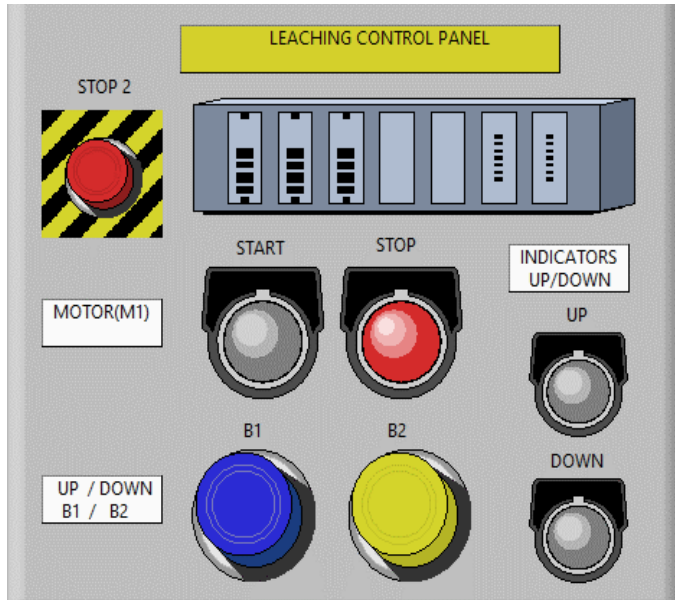


Fig. 4. Leaching Process Control Panel.

III. RESULTS

Fig. 5 shows the behavior of the temperature inside the greenhouse, obtaining as a result that the ambient temperature of the greenhouse is a replica of the real ambient temperature that occurs in the Andes, all this is achieved artificially and thanks to the mechatronic system proposed, ideal temperatures are achieved to produce white chuño since natural temperatures are increasing and climate changes are also negatively affecting the food supply [10]. Under this concept, the use of intelligent greenhouses is necessary to obtain higher quality products [11].

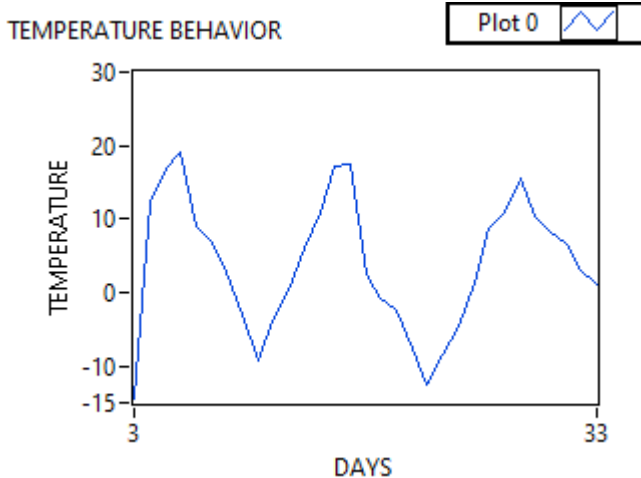


Fig. 5. Temperature Behavior Inside The Greenhouse.

Fig. 6 shows, as a result, the programming done for the simulation, which uses a block diagram consisting of a graphical visual programming language that simplifies the integration of hardware for engineering applications [12]. This

programming can be translated into a ladder language which is a high-level language. Ladder Logic is the most widely used language for programming PLCs throughout the industry [12].

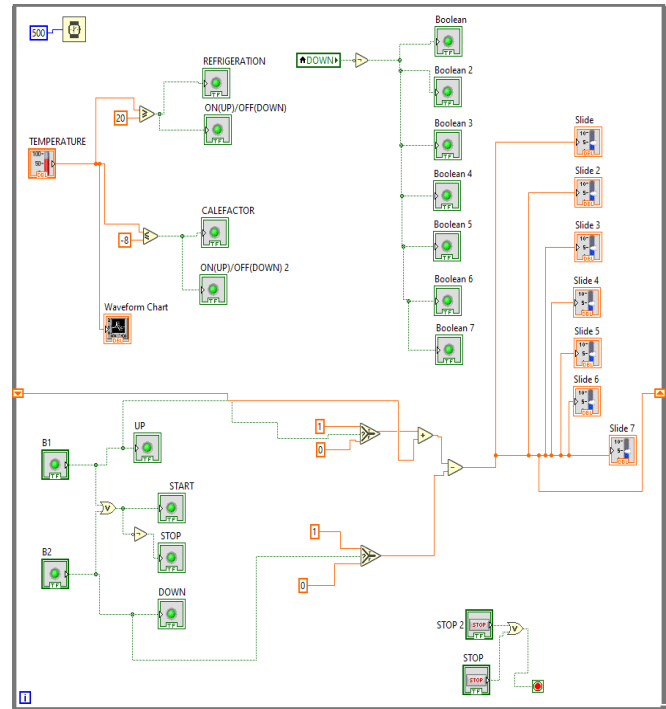


Fig. 6. Programming Block Diagram Developed in LabVIEW.

Fig. 7 shows the mechatronic system in its entirety, including the five phases shown in Table I. In the first phase, the potato enters through a conveyor belt that leads to the second conveyor and sorting belt (roller belt) where the potatoes are separated according to their size. At the end of the selection process, the greenhouse refrigeration system is turned on, bringing the ambient temperature down to -15°C for a period of approximately 12 hours before turning on the heating system and bringing the ambient temperature in the greenhouse to temperatures ranging from 18°C to 30°C, simulating the real ambient temperature that occurs in the Andes during the day and night. At the end of this phase, the potatoes are lowered into the tank to enter the leaching process, because the platform where they fall when selected works as an elevator, having the ability to raise and lower the potatoes automatically or manually when necessary, proceeding to drive the motor that allows the movement of the turbines. Additionally, in the tank, there will be a water cleaning and filtration system to be able to use it in a new process without the need to waste this fluid.

To enter the last phase, which is the second freezing, the pneumatic press will suppress the potatoes by making rapid oscillations that simulate the handmade steps that are made with bare feet to remove excess water [2], the use of the system is to ensure the safety of the food. Once this process is finished, freezing begins and at the end of the period the white chuño is ready, obtaining the characteristic color, the product is removed by employing independent conveyor belts for each size of chuño (small, medium, and large).

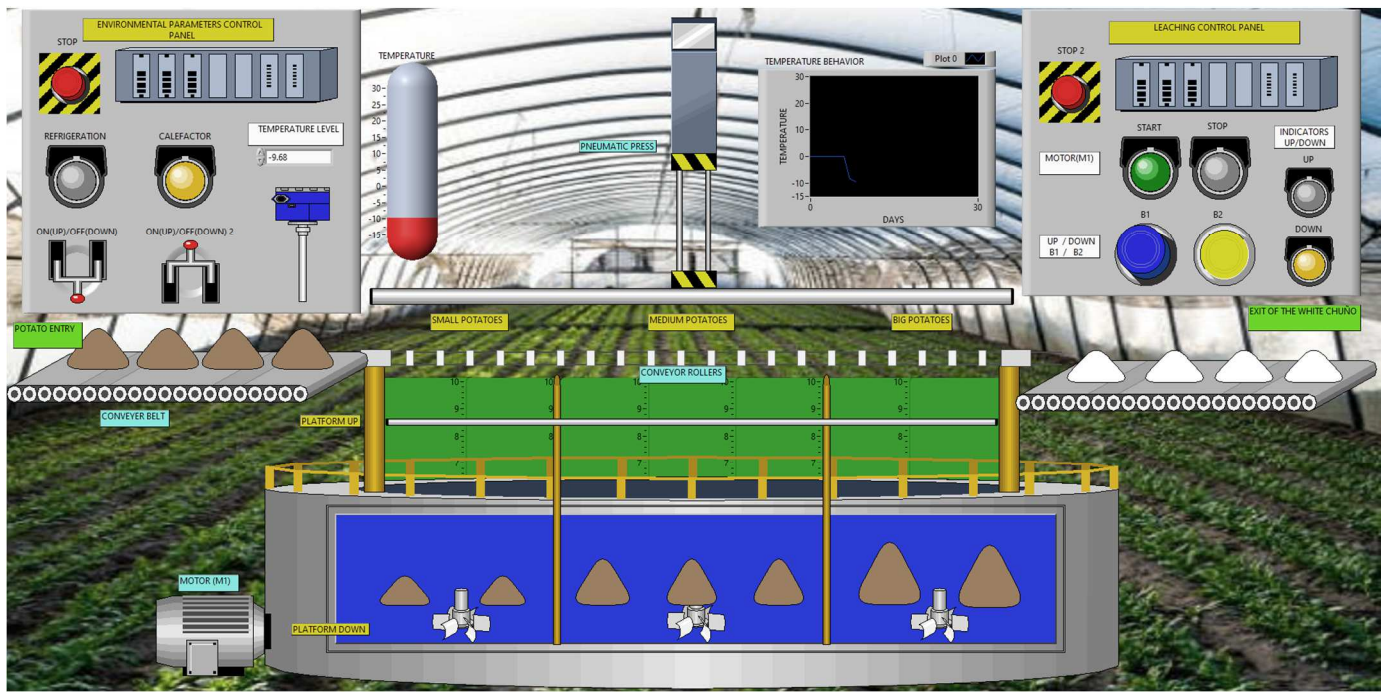


Fig. 7. Complete simulation designed in LabVIEW software, HMI environment.

Table II shows because of the approximate times that each phase of the production of white chuño takes in the nursery. As it is an intelligent automation system, it does not depend on external environmental factors, making the production of the white chuño more efficient and higher quality, since the chuño is not directly exposed to the sun, it obtains and preserves its characteristic white color, which is a very important aspect in determining the quality of the tunta.

Table II. Table of Results of Times and Parameters

Phase	Approximate Time	Parameter
Selection and/or classification	30 minutes	potato(small-medium-big)
First freezing	8 hours per day	(-10°C) Average Temperature
Leaching in water	600 continuous hours	(1°C-5°C) Average Temperature of water
Frozen second	8 hours per day	(-10°C) Average Temperature

Respecting the times and phases of production that are carried out in an artisanal manner for the production of white chuño, as well as taking into account the technical standard developed by INACAL for the production of tunta, results in the automation and control system being a sustainable technology since it preserves the artisanal production phases and complies with the standards mentioned in NTP 011.401:2020 [9].

IV. DISCUSSION

In the work done by Hernández-Morales, Carlos A.; Luna-Rivera, J. M. and Pérez-Jimenez Rafael designed a monitoring system for agricultural applications based on IoT applied in a commercial size greenhouse, in which they obtained, as a result, the temperature predictions with an error of 1 °C [10]. This system they developed is very good, but the mechatronic system developed in this paper is specifically focused on the production of white chuño, for which all environmental parameters, automation, and control systems are based on this product, being a greenhouse specialized in the production of white chuño.

In the work of A. Mellit, M. Benghanem, O. Herrak, and A. Messalaoui present the design of a remote monitoring system for intelligent greenhouses, which allows the creation of an optimal artificial environment inside the greenhouse, for which they use as the brain of the prototype an Arduino mega 2560 and a mobile application that allows the control of the different environmental parameters [11]. Unlike the work, a PLC is used as the brain of the prototype which allows automation and monitoring work much more complex compared to an Arduino, and for monitoring an HMI system is used which is much more efficient and practical to control the greenhouse, since you can see all the parameters and control it manually from the same interface if necessary.

V. CONCLUSIONS

The results obtained show that it is feasible to implement an automation and control system for the production of white chuño. For which the system works according to the recommendations and parameters given by INACAL.

The mechatronic system with automation and control concepts allows for regulating the internal temperature of the greenhouse in real-time, which allows replicating the

temperature that occurs in the Andes. The proposed system will also help improve the productivity of community members and companies engaged in the production of white chuño.

The novelty of the proposed mechatronic system involves real-time monitoring and control of the entire white chuño production process, as well as the environmental parameters required to produce it. All these thanks to the Human-Machine Interface (HMI), which facilitates the visualization of all processes and the stage of production. On the other hand, the system also provides very important data that helps to improve productivity, as well as to guarantee the innocuousness of the consumption of the white chuño, as well as the white color, which is a very important aspect that is damaged by too much exposure to the sun, which is avoided by being produced in a closed environment, which leads to a better quality of the product.

VI. REFERENCES

- [1] K. Yoshikawa and F. Apaza, "Unfrozen state by the supercooling of chuño for traditional agriculture in altiplano Andes", *Environmental and Sustainability Indicators*, vol. 8, núm. 100063, p. 100063, 2020.
- [2] M. Mamani, "El chuño: preparación, uso, almacenamiento", *Tecnología andina*, IEP. Ravines, R. (Ed.), 1978, p. 227-239.
- [3] S. de Haan et al., "Traditional processing of black and white chuño in the Peruvian Andes: Regional variants and effect on the mineral content of native potato cultivars", *Econ. Bot.*, vol. 64, núm. 3, pp. 217–234, 2010.
- [4] M. A. Melton, M. E. Biver, and R. Panjarjian, "Differentiating Chuño Blanco and Chuño negro in archaeological samples based on starch metrics and morphological attributes", *J. Archaeol. Sci. Rep.*, vol. 34, núm. 102650, p. 102650, 2020.
- [5] M. A. Hardigan et al., "Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato", *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, núm. 46, pp. E9999–E10008, 2017.
- [6] G. Burgos, S. de Haan, E. Salas, and M. Bonierbale, "Protein, iron, zinc and calcium concentrations of potatoes following traditional processing as 'chuño'", *J. Food Compost. Anal.*, vol. 22, núm. 6, pp. 617–619, 2009.
- [7] J. M. Peñarrieta, T. Salluca, L. Tejada, J. A. Alvarado, and B. Bergenstahl, "Changes in phenolic antioxidants during chuño production (traditional Andean freeze and sun-dried potato)", *J. Food Compost. Anal.*, vol. 24, núm. 4–5, pp. 580–587, 2011.
- [8] VDI-Richtlinien, "Design methodology for Mechatronic Systems. 118," 2004.
- [9] "Inacal aprobó norma técnica para impulsar buenas prácticas de procesamiento artesanal de la tunta", Gob.pe. [En línea]. Disponible en: <https://www.gob.pe/institucion/inacal/noticias/343057-inacal-aprobo-norma-tecnica-para-impulsar-buenas-practicas-de-procesamiento-artesanal-de-la-tunta>. [Consulted: 23-mar-2022].
- [10] C. A. Hernández-Morales, J. M. Luna-Rivera, and R. Pérez-Jiménez, "Design and deployment of a practical IoT-based monitoring system for protected cultivations", *Comput. Commun.*, vol. 186, pp. 51–64, 2022.
- [11] A. Mellit, M. Benghanem, O. Herrak, and A. Messalaoui, "Design of a novel remote monitoring system for smart greenhouses using the Internet of things and deep convolutional neural networks", *Energies*, vol. 14, núm. 16, p. 5045, 2021.
- [12] "¿Qué es LabVIEW?" Engineer Ambitiously - NI. <https://www.ni.com/es-cr/shop/labview.html> (Accessed May 13, 2022).
- [13] F. Fronchetti et al., "Language impact on productivity for industrial end users: A case study from Programmable Logic Controllers", *Journal of Computer Languages*, vol. 69, núm. 101087, p. 101087, 2022.
- [14] I. L. Q. Mosquera, J. E. R. Fierro, J. R. O. Zacarias, J. B. Montero, S. A. C. Quijano and D. Huamanchahua, "Design of an Automated System for Cattle-Feed Dispensing in Cattle-Cows," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0671-0675, doi: 10.1109/UEMCON53757.2021.9666491.
- [15] S. I. Del Carpio Ramirez, J. R. O. Zacarias, J. B. M. Vazquez, S. A. C. Quijano and D. Huamanchahua, "Comparison Analysis of FIR, ARX, ARMAX by Least-Squares Estimation of the Temperature Variations of a Pasteurization Process," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0699-0704, doi: 10.1109/UEMCON53757.2021.9666649.
- [16] A. H. Uribe, J. Brayan Macuri Vasquez, A. C. Miranda Yauri, and D. Huamanchahua, "Control and Monitoring System of Hydraulic Parameters for Rainbow Trout Culture," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0780-0784, doi: 10.1109/UEMCON53757.2021.9666512.
- [17] J. P. A. Misajel, S. A. C. Quijano, D. M. C. Esteban, S. R. T. Rojas, D. Huamanchahua and R. A. M. Grados, "Design of a Prototype for Water Desalination Plant using Flexible, Low-Cost Titanium Dioxide Nanoparticles," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0286-0289, doi: 10.1109/UEMCON53757.2021.9666586.

Biological Signals for the Control of Robotic Devices in Rehabilitation: An Innovative Review

Deyby Huamanchahua
Department of Electrical and Mechatronic Engineering
Universidad Ingeniería y Tecnología - UTEC
Lima, Perú
dhuamanchahua@utec.edu.pe

José Asencios-Chávez
Facultad de Ingeniería
Universidad Tecnológica del Perú
Lima, Perú
u18103183@utp.edu.pe

Luis Alberto Huamán-Lévano
Mechatronic Engineering
Universidad Continental
Huancayo, Perú
72043692@continental.edu.pe

Nicole Caballero-Canchanya
School of Biomedical Engineering
Universidad Nacional Mayor de San Marcos
Lima, Perú
nicole.caballero@unmsm.edu.pe

Abstract— Biological signals have been extensively studied for disease diagnosis, treatment, and biomedical research to know the patient's health status; however, the potential of these signals goes beyond diagnosis, to the point that they can be used for more precise control of robotic devices in rehabilitation. The purpose of this article is to document a systematic review of the applications of biological signals such as EMG, EEG, and EOG for the development of these devices. The objective is to provide the researcher with a matrix that integrates the types of signals, sensors, and related components for the implementation of robotic devices in rehabilitation. As a method, different databases and specialized search engines were used to collect the information from 2019 to 2021. All the investigations found were selected, 27 being selected. Finally, it is concluded that there are different applications of biological signals for manipulation, of prostheses, orthoses, and robotic devices, which allows continuing the development of these that allow a better rehabilitation to the patient.

Keywords— Sensors, Biological signal, Controller type, TRL, Robotic device

I. INTRODUCTION

Robotics has evolved and today it can not only be seen within the industrial aspect but also in the health area, specifically in the rehabilitation of patients who have temporarily or permanently lost their motor, cognitive or social skills due to a disease. Thus, these guided robotic devices for rehabilitation must be safe and effective, so that the acquisition of biological signals from the patient will allow a more accurate approach when establishing a specific rehabilitation for each disease, since "the bioelectrical signals of the human being contain abundant information, such as perceptual information and physiological states" [1]

Therefore, biological signals are signals produced by the human being, which can be physical, chemical, electrical, optical, or electromagnetic, and whose main function is to provide various types of information and provide the doctor with a possible diagnosis according to the analysis of

waveforms. However, these signals can be used to control robotic prototypes that help in the rehabilitation of patients who, due to a stroke with hemiparesis, spinal cord injuries, neurological disabilities, and, among other diseases, have affected their motor and cognitive or social abilities.

Then, for the control of a robotic rehabilitation prototype, three types of biological signals are essential to carry out an effective rehabilitation in a patient's body, first of all, the EMG signals, are the technique of obtaining electrical signals produced during muscle contractions [2], secondly, the EEG signal is a safe neurophysiological tool that records brain activities at low cost [3] and the EOG signals detect eye movements between the cornea and the patient retina [4].

Therefore, each biological signal has a specific type of control, for example, "Electromyography (EMG)-based control schemes have shown their superiority in human-robot cooperation because movement intention can be well estimated by EMG signals" [5], however, EEG signals, through the development of entity-relationship diagrams, can define specific tasks because "reflecting sensorimotor brain activity through ERD/ERS components are extracted to identify different tasks" [6] and, on the other hand, EOG signals are widely used in human-computer interfaces because they can accurately classify the movements of electrically active nerves at the back of the eyeball [7].

Then, when comparing the devices for the acquisition of the patient's biological signals, clear use of Ag/AgCl electrodes was obtained in each one of them, since these electrodes are beneficial because they are economical and, when properly applied, deviations can be reduced [8]. However, the sensors used by each of the robotic prototypes range from the force, pressure, and torque sensors for the movement of the actuators to the various EMG, EEG, and EOG sensors, most of which are used because they have integrated filtering and amplification circuit for the patient's biological signals that range from + 1.5mV and are easy to use [9].

For these reasons, the main objective of this work is to provide a research matrix in the comparison of biological signals obtained from a patient, within which it will be determined which of them is the most influential today in the control of robotic devices for rehabilitation.

II. METHODOLOGY

To carry out this research, a search query was performed in the Scopus database, IEEE Explore, Springer link, ScienceDirect, journal articles such as Frontiers in Neurorobotics as well as collaboration and dissemination platforms such as ResearchGate. Keywords such as robotic

device, exoskeleton, rehabilitation, biological signals, and control were used. This process resulted in an average of 38 publications spanning 2018 to 2021.

III. REVIEW OF PUBLICATIONS

To simplify and structure the information. The robotic rehabilitation devices were classified according to Type of loss, Robotic Device, sensor, Biological Signal, Biological Signal Acquisition Device, and state of development of the robotic device [prototype or study].

The review describes the different biological signals for the control or manipulation of robotic rehabilitation devices.

TABLE I. BIOLOGICAL SIGNALS FOR CONTROL OR MANIPULATION OF ROBOTIC REHABILITATION DEVICES

Reference	Type of loss	Robotic device	Sensors	Biological signal	Acquisition device	TRL
Sheng, B. (2019) [10]	LMU	SPT-ROB	N.E.	EMG	E	3
Li, X. (2019) [1]	MD	VR	EMGS	EMG	E	7
Zhang, J. (2019) [6]	AC	S-ROB	MYO & NUAMPS	EOG, EEG y EMG	E	5
Wu, Q. (2019) [11]	DMO	EXO	EMGS, FS, IS	EMG	E	4
Karácsony, T. (2019) [12]	AC	MI-BCI-VR	EEGS	EEG	E	4
Hajj, N. (2019) [13]	SP	EXO	EMGS, OPTS	EMG	N.E.	3
Pancholi, S. (2019) [14]	LMU	PROS	EMGS	EMG	E	3
Xu, J. (2019) [15]	LML	R-EXO	TS, EMGS, FS	EMG	E	4
Wang, Y. (2019) [9]	LMU	S-EXO	EMGS, PS	EMG	E	2
Son, C. (2021) [16]	AC	EXO	FS, EMGS & PS	EMG	N.E.	9
Huang, Y. (2021) [17]	MA	ROB	N.E.	EMG	E	8
Casey, A. (2021) [18]	ND	ROB	EMGS & EEGS	EEG y EMG	Open BCI & E	7
Yang, Z. (2021) [19]	AC	EXO	Sensor MTx, IS	EMG	BE	7
Takashi, T. (2021) [20]	OH	R-EXO	MS, EMGS	EMG	E	7
Soma, Y. (2021) [21]	T	H-ROB	N.E.	EMG	E	9
Chen, Y. (2021) [22]	AC	S-EXO	PS, AS, EMGS	EMG	Open BCI & E	5
Zhuang, Y. (2021) [5]	LML	EXO	TS	EMG	BE	6
Pérez, R. (2021) [23]	LML	EXO	EMGS	EMG	N.E.	6
Mora, E. (2020) [24]	LML	R-EXO	EMGS	EMG	E	5
McDonald, C. (2020) [25]	SCI	EXO	EMGS	EMG	N.E.	5
Bouteraa, Y. (2020) [26]	LMU	R-EXO	e-Health-S, PS, EMGS	EMG	E	5
Zhao, L. (2020) [27]	N.E.	EXO	FS, EMGS	EMG	E	4
Wang, W. (2018) [28]	AC	EXO	EMGS	EMG	E	4
Suppiah, R. (2020) [29]	AC	ROB	EMGS	EMG	E	5
Ferdiansyah, F. (2020) [30]	LMU	ORTH	EEGS & EMGS	EEG & EMG	E	3
Martínez, J. (2020) [7]	N.E.	N.E.	N.E.	EOG	Open BCI & EH	4
Kuntal, K. (2020) [8]	T	WC	N.E.	EOG	E	4
Rodrigues, C. (2020) [31]	SCI	EXO	N.E.	EMG	BE	5
Mokri, C. (2022) [2]	LML	ROB	EMGS & FS	EMG	E	6
Nann, E. (2021) [4]	AC	EXO	N.E.	EMG & EOG	E	7
Yamamoto, I. (2018) [32]	AC	ROB	EMGS, TS, ATS & POS	EMG	E	7
Castiblanco, J. (2021) [33]	AC	EXO	EMGS	EMG	E	5
Bouteraa, Y. (2018) [34]	AC	EXO	EMGS	EMG	E	3
Cisnal, A. (2021) [35]	AC	EXO	EMGS	EMG	E	4
Asokan, A. (2019) [36]	AC	EXO	EMGS	EMG	E	5
Guo, J. (2018) [37]	SCI	ROB	EMGS, PS & IS	EMG	E	4

Malik, A. (2019) [38]	AC	ROB	EMGS	EMG	E	4
Tiboni, M. (2018) [39]	LMU	EXO	EMGS	EMG	E	4

Note: Abbreviation: AC: Stroke, LMU: Upper Limb Movement Loss, MD: Muscular Impairment, BMD: Motor Dysfunction, SP: Spasticity, LML: Lower Limb Movement Loss, MA: Muscle Activation, ND: Neurological Disabilities, OH: Oral Hypofunction, SCI: Spinal Cord Injury, T: Tetraplegia, PROS: Prosthetics, VR: Virtual Reality, MI-BCI-VR: Brain Computer Interface with Virtual Reality, S-ROB: Soft Robot, ROB: Robotic System, H-ROB: Hybrid Robot, SPT-ROB: Robot based on Standardized Performance Tests, EXO: Exoskeletons, R-EXO: Robotic Exoskeleton, S-EXO: Soft Exoskeleton, WC: Wheelchair, ORTH: Orthosis, EMGS: EMG Sensor, MYO: Myo Armband, NUAMPS: Neuroscan NuAmps Express System, FS: Force Sensor, IS: Inertial Sensor, EEGS: EEG Sensor, OPTS: Optoelectric System, TS: Torque Sensor, PS: Pressure Sensor, MS: Magnetic Sensor, AS: Air Sensor, e-Health-S: e-Health Sensor Shield V2.0, ATS: Actuation Sensor, POS: Position Sensor, EMG: Elect Signal romyogram, EEG: Electroencephalogram signal, EOG: Electrooculogram signal, E: Electrodes, BE: Bipolar electrodes, TRL 1: Basic Research, TRL 2: Technology Formulation, TRL 3: Applied research, TRL 4: Small-scale development, TRL 5: Real-scale development, TRL 6: Valid system in simulated environment, TRL 7: Real-environment validated system, N.E.: Unspecified

A. Type of loss

This section shows the main types of losses that affect the correct mobility of the upper and lower extremities: cerebrovascular accident (CA), spinal cord injury (SCI), loss of movement of upper extremities (LMU), and lower extremities (LML), tetraplegia (T). Table II shows the characteristics of the above-mentioned types of loss.

Diagnosis	Feature
AC (Stroke)	It occurs when blood flow stops in one part of the brain. Its prevalence in the USA increases with advancing age.[40]
SCI (Spinal cord injury)	Spinal cord injury usually causes loss of mobility, strength, and sensation below the affected area.
T (Tetraplegia)	Paralysis is caused by disease or injury.

It is important to consider the type of loss that affects the normal mobility of the patient's limb to develop the robotic device that best contributes to the patient's rehabilitation. Considering the type of loss and the functionality that the robotic device should have, the most appropriate device can be designed. As can be seen in Table III, most of the studies focus on stroke with 36,84%. This is followed by loss of movement of the upper extremities with 15,79% and the lower extremities with 13,16%. To a lesser extent, we find special spinal cord injuries with 7,89%, tetraplegia with 5,26%, and the rest below 5%.

TABLE III. THE AMOUNT BY TYPE OF LOSS

Diagnosis	Amount	
	Frequency	Percentage
AC	14	36,84%
LMU	6	15,79%
LML	5	13,16%
SCI	3	7,89%
T	2	5,26%
MD	1	2,63%
SP	1	2,63%
MA	1	2,63%
ND	1	2,63%
DMO	1	2,63%
OH	1	2,63%
N.E.	2	5,26%

Note: CA: Stroke, LMU: Loss of upper limb movement, MD: Muscle deficiency, BMD: Motor dysfunction, SP: Spasticity, LML: Loss of lower limb movement, MA: Muscle activation, ND: Neurological disabilities, OH: Oral hypofunction, SCI: Spinal cord injury, T: Tetraplegia,

B. Robotic device

This area covers the main robotic devices handled with the control of biological signals: exoskeletons (EXO), robotic systems (ROB), and prosthetics (PROS). Exoskeletons were found to be divided into three groups, exoskeletons, robotic exoskeletons, and soft exoskeletons, each of them with different characteristics, the last one being.

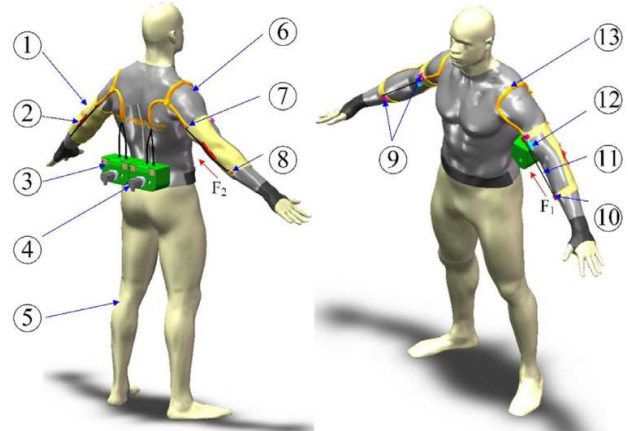


Fig. 1 General mechanical structure of the proposed soft exoskeleton robot for elbow motion assistance. (1-Soft envelope; 2-Guiding mechanism; 3-Actuator unit; 4-Servomotor; 5-Human; 6-Base layer; 7-Sheath support; 8-Anchor point; 9-Inertial measurement unit; 10-Force sensor; 11-Inner tendon; 12-EMG sensor; 13-Outer sheath). Source: Q. Wu, B. Chen, and H. Wu, 201CNeural-network-enhanced torque estimation control of a soft wearable exoskeleton for elbow assistance,” Mechatronics, 2019.

As well as identifying the type of loss that affects the patient's mobility, it is also important to consider the type of robotic device to be used for patient rehabilitation. As shown in Table IV, most of the studies concentrate on developing exoskeletons with 42,11%. In addition, 10,53% are robotic exoskeletons (R-EXO) and 5,26% are soft exoskeletons (S-EXO). Exoskeletons are followed by robotic systems with 18,42% and the rest below 5%.

TABLE IV. TYPES OF ROBOTIC DEVICES DEVELOPED

Robotic Device	Amount	
	Frequency	Percentage
EXO	16	42,11%
R-EXO	4	10,53%
ROB	7	18,42%
S-EXO	2	5,26%
PROS	1	2,63%
S-ROB	1	2,63%
H-ROB	1	2,63%
WC	1	2,63%
ORTH	1	2,63%
VR	1	2,63%

MI-BCI-VR	1	2,63%
SPT-ROB	1	2,63%
N.E.	1	2,63%

Note: PROS: Prosthesis, VR: Virtual Reality, MI-BCI-VR: Brain-Computer Interface with Virtual Reality, S-ROB: Soft Robot, ROB: Robotic System, H-ROB: Hybrid Robot, SPT-ROB: Robot based on Standardized Performance Testing, EXO: Exoskeletons, R-EXO: Robotic Exoskeleton, S-EXO: Soft Exoskeleton, WC: Wheelchair, ORTH: Orthosis, ORTH: Orthosis

OPTS	1	2,63%
MTxS	1	2,63%
e-Health-S	1	2,63%
ATS	1	2,63%
POS	1	2,63%
N.E.	6	15,79%

Note: EMGS: EMG Sensor, MYO: Myo Armband, NUAMPS: Neuroscan NuAmps Express System, FS: Force Sensor, IS: Inertial Sensor, EEGS: EEG Sensor, OPTS: Optoelectrical System, TS: Torque Sensor, PS: Pressure Sensor, MS: Magnetic Sensor, AS: Air Sensor, e-Health-S: e-Health Sensor Shield V2.0

C. Sensors

In this area, a review of the sensors that are most used in the development of robotic devices is made. Each sensor is used to measure a certain variable. According to Table V, the most used sensor for the development of robotic devices in rehabilitation are the EMG sensors (EMGS) with 60,53%. It is important to consider that this percentage represents the studies that use this sensor alone and together with other sensors. Of this total, 17,65% of the studies used the MyoWare sensor. [11], [23], [29], for EMG signal processing, while another 17,86% used the Open BCI sensor for EMG signal processing. Figure 2 shows the electromyography process using the Open BCI device for processing.

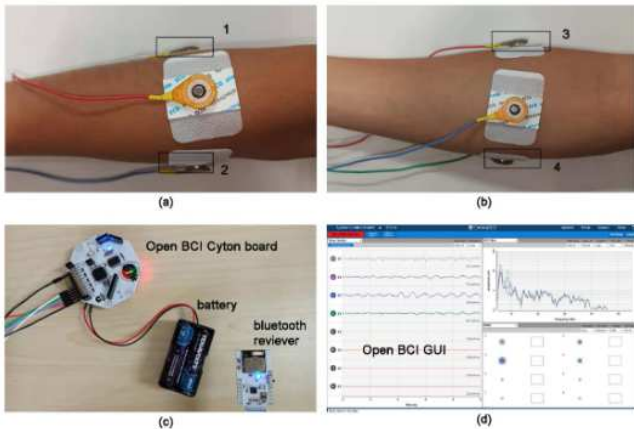


Fig. 2 The process of surface electromyography (sEMG) collection. (a) The electrode positions of 1 and 2. (b) The electrode positions of 3 and 4. (c) The hardware for sEMG collection. (d) The software of sEMG collection. Source: Chen, Y.; Yang, Z.; Wen, Y.A Soft Exoskeleton Glove for Hand Bilateral Training via Surface EMG. 2021

The other most used sensors are the force sensor (FS) and pressure sensor (PS) both with 13,16%, the EEG sensor (EEGS) with 7,89%, and the inertial and torque sensor with 5,26% each. The rest is below 5%.

TABLE V. PERCENTAGE OF SENSORS USED

Sensors	Amount	
	Frequency	Percentage
EMGS	23	60,53%
FS	5	13,16%
PS	5	13,16%
EEGS	3	7,89%
IS	2	5,26%
TS	2	5,26%
MS	1	2,63%
AS	1	2,63%
MYO	1	2,63%
NUAMOS	1	2,63%

D. Biological signal

This section reviews the biological signals used for the control of robotic devices. The human body generates more than 20 biosignals that are used for different medical purposes; according to the research reviewed, it was identified that the signals that were mostly used for this type of application were electromyography (EMG) signals, either superficial (sEMG) or intramuscular (iEMG), electroencephalography (EEG) and electrooculography (EOG) signals. In Table VI the most used signal in the studies is EMG signals with 81,58%. This reflects that the robotic devices developed are mostly employing the EMG signal to be controlled or manipulated by the patient. In addition, 5,26% employed only the EOG signal, and a lower percentage of 2,63% mentioned that they employed only the EEG signal.

The rest of the studies employed two or more signals for manipulation of the robotic devices with a percentage of 11%. However, what is striking is that the EMG signal is almost always present.

TABLE VI. PERCENTAGE OF STUDIES ACCORDING TO A BIOLOGICAL SIGNAL USED

Biological signal	Amount	
	Frequency	Percentage
EMG	31	81,58%
EEG y EMG	2	5,26%
EOG	2	5,26%
EEG	1	2,63%
EMG Y EOG	1	2,63%
EOG, EMG y EMG	1	2,63%

Note: EMG: Electromyography signal, Intramuscular electromyography signal, EEG: Electroencephalography signal, EOG: Electrooculography signal.

E. Signal acquisition device

In this part, a review of biological signal acquisition devices is made. This device is the first contact to capture the signal and pass it for processing. Table VII shows the various devices used in the investigations reviewed for the acquisition of the biological signal. The 86,84% of the investigations reviewed employ electrodes as the biological signal acquisition device, taking advantage of the advantages it offers such as reproducibility, stability, and ease of preparation. Electrodes are devices that convert the bioelectrical signal into electrical potentials to be later treated. Three main types of electrodes were identified, being dry electrodes with a percentage of 90,91% the most used, followed by bipolar electrodes with a percentage of 9,09%, and to a lesser extent wet electrodes with a percentage of 2,63%. Similarly, the electrodes were Ag/AgCl electrodes since these provide greater stability.

TABLE VII. BIOLOGICAL SIGNAL ACQUISITION DEVICE

Device	Amount	
	Frequency	Percentage
Electrodes	33	86,84%
Dry electrodes	30	90,91%
Bipolar electrodes	3	9,09%
Wet electrodes	1	2,63%
Not specified	4	10,53%

Doll	2	5,26%
Mandible	1	2,63%
Foot	1	2,63%
Knee	2	5,26%
N.E.	6	15,79%

F. Development Status

Each research managed to determine different objectives to collaborate in the development of a robotic device that uses biological signals for the control of rehabilitation reaching different states such as functional, prototypes, designs, and studies. The Technology Maturity Levels manual helps to identify the state of technology in the research, in which the reader will know how to associate the 9 levels. In Table VIII, it is observed that TRL 5 (Components validated in relevant environment) out of 38 research obtained 23.68%, followed by TRL 4 (Components validated in the laboratory) with 28.95% and TRL 7 (Prototype validated in a real environment) with 15.79%. This means that most of the studies reviewed are in the prototype and technology development stage.

TABLE VIII. PERCENTAGES OF STUDIES BASED ON TLR

Technology Readiness Levels (TRL)	Amount	
	Frequency	Percentage
2	1	2,63%
3	5	13,16%
4	11	28,95%
5	9	23,68%
6	3	7,89%
7	6	15,79%
8	1	2,63%
9	2	5,26%

G. Rehabilitation body part

In this area, a review is made of the main parts of the body where the robotic device was used for rehabilitation. According to Table IX, the body part where rehabilitation with a robotic device is most applied is the upper extremity with 31.58%. For this purpose, more robotic devices of the exoskeleton type are used, to support the patient to perform daily activities with the least limitations. This is followed by the hand, with 18.42%. On the other hand, the third body part with the highest use of robotic devices to complement its rehabilitation is the lower extremity.

TABLE IX. PERCENTAGES OF STUDIES BASED ON TLR

Body Part	Amount	
	Frequency	Percentage
Upper Limb	12	31,58%
Hand	7	18,42%
Elbow	2	5,26%
Lower Limb	5	13,16%

IV. CONCLUSION

With the growing population and shortage of specialists, many researchers have devised robotic devices to aid in the rehabilitation process. The article reviews various research on designs and prototypes of robotic devices controlled by biological signals for rehabilitation. It is anticipated that these robotic devices will move closer and closer to ideal performance in the future.

This research recognized the details in the development of the different designs and prototypes of robotic devices for rehabilitation intending to be able to develop a scientific article on a prototype of robotic systems controlled by electrooculography signals that do not have the weaknesses found in previous research, in addition to following the latest advances and developments in the subject. Likewise, we found that the EMG signal is the most used for the control of robotic devices and this same signal serves as a support for the use of EEG and EOG signals.

REFERENCES

- [1] X. Li, Z. Zhou, W. Liu, and M. Ji, "Wireless sEMG-based identification in a virtual reality environment," *Microelectronics Reliability*, vol. 98, pp. 78–85, Jul. 2019, doi: 10.1016/J.MICROREL.2019.04.007.
- [2] C. Mokri, M. Bamdad, and V. Abolghasemi, "Muscle force estimation from lower limb EMG signals using novel optimized machine learning techniques," *Medical and Biological Engineering and Computing*, vol. 60, no. 3, pp. 683–699, Mar. 2022, doi: 10.1007/S11517-021-02466-Z/TABLES/9.
- [3] N. K. Al-Qazzaz, Z. A. A. Alyasseri, K. H. Abdulkareem, N. S. Ali, M. N. Al-Mhiqani, and C. Guger, "EEG feature fusion for motor imagery: A new robust framework towards stroke patients' rehabilitation," *Computers in Biology and Medicine*, vol. 137, p. 104799, Oct. 2021, doi: 10.1016/J.COMPBIOMED.2021.104799.
- [4] M. Nann *et al.*, "Restoring Activities of Daily Living Using an EEG/EOG-Controlled Semiautonomous and Mobile Whole-Arm Exoskeleton in Chronic Stroke," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2314–2321, Jun. 2021, doi: 10.1109/JSYST.2020.3021485.
- [5] Y. Zhuang, Y. Leng, J. Zhou, R. Song, L. Li, and S. W. Su, "Voluntary Control of an Ankle Joint Exoskeleton by Able-Bodied Individuals and Stroke Survivors Using EMG-Based Admittance Control Scheme," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 2, pp. 695–705, Feb. 2021, doi: 10.1109/TBME.2020.3012296.
- [6] J. Zhang, B. Wang, C. Zhang, Y. Xiao, and M. Y. Wang, "An EEG/EMG/EOG-based multimodal human-machine interface to real-time control of a soft robot hand," *Frontiers*

- in *Neurorobotics*, vol. 13, p. 7, Mar. 2019, doi: 10.3389/FNBOT.2019.00007/BIBTEX.
- [7] J. Martínez-Cerveró *et al.*, “Open software/hardware platform for human-computer interface based on electrooculography (EOG) signal classification,” *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, doi: 10.3390/S20092443.
- [8] K. Kuntal, I. Banerjee, and P. P. Lakshmi, “Design of Wheelchair based on Electrooculography,” *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, pp. 632–636, Jul. 2020, doi: 10.1109/ICCSP48568.2020.9182157.
- [9] Y. Wang and Q. Xu, “Design of a new wrist rehabilitation robot based on soft fluidic muscle,” *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, vol. 2019-July, pp. 595–600, Jul. 2019, doi: 10.1109/AIM.2019.8868626.
- [10] B. Sheng, L. Tang, O. M. Moosman, C. Deng, S. Xie, and Y. Zhang, “Development of a biological signal-based evaluator for robot-assisted upper-limb rehabilitation: a pilot study,” *Australasian Physical & Engineering Sciences in Medicine 2019 42:3*, vol. 42, no. 3, pp. 789–801, Aug. 2019, doi: 10.1007/S13246-019-00783-0.
- [11] Q. Wu, B. Chen, and H. Wu, “Neural-network-enhanced torque estimation control of a soft wearable exoskeleton for elbow assistance,” *Mechatronics*, vol. 63, p. 102279, Nov. 2019, doi: 10.1016/J.MECHATRONICS.2019.102279.
- [12] T. Karácsy, J. P. Hansen, H. K. Iversen, and S. Puthusserypady, “Brain computer interface for neuro-rehabilitation with deep learning classification and virtual reality feedback,” *ACM International Conference Proceeding Series*, Mar. 2019, doi: 10.1145/3311823.3311864.
- [13] N. al Hajj, J. Charafeddine, M. Khalil, and S. Al-Fayad, “Using Bio-Kinematic signals for Rehabilitation Exoskeleton Control,” *International Conference on Advances in Biomedical Engineering, ICABME*, vol. 2019-October, Oct. 2019, doi: 10.1109/ICABME47164.2019.8940290.
- [14] S. Pancholi, P. Jain, A. Varghese, and A. M. Joshi, “A Novel Time-Domain based Feature for EMG-PR Prosthetic and Rehabilitation Application,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 5084–5087, Jul. 2019, doi: 10.1109/EMBC.2019.8857399.
- [15] J. Xu *et al.*, “A Multi-Mode Rehabilitation Robot with Magnetorheological Actuators Based on Human Motion Intention Estimation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 2216–2228, Oct. 2019, doi: 10.1109/TNSRE.2019.2937000.
- [16] C. Son *et al.*, “The effect of pelvic movements of a gait training system for stroke patients: a single blind, randomized, parallel study,” *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, Dec. 2021, doi: 10.1186/S12984-021-00964-7.
- [17] Y. Huang, R. Song, A. Argha, B. G. Celler, A. v. Savkin, and S. W. Su, “Human motion intent description based on bumpless switching mechanism for rehabilitation robot,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 673–682, 2021, doi: 10.1109/TNSRE.2021.3066592.
- [18] A. Casey, H. Azhar, M. Grzes, and M. Sakel, “BCI controlled robotic arm as assistance to the rehabilitation of neurologically disabled patients,” *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 5, pp. 525–537, 2021, doi: 10.1080/17483107.2019.1683239.
- [19] Z. Yang, S. Guo, Y. Liu, H. Hirata, and T. Tamiya, “An intention-based online bilateral training system for upper limb motor rehabilitation,” *Microsystem Technologies*, vol. 27, no. 1, pp. 211–222, Jan. 2021, doi: 10.1007/S00542-020-04939-X.
- [20] T. Kameda, M. Sakamoto, and K. Terada, “Semi-powered exoskeleton that regulates the muscular activity of jaw movement for oral functional rehabilitation/training,” *Dental Materials Journal*, vol. 40, no. 1, pp. 101–109, 2021, doi: 10.4012/DMJ.2019-400.
- [21] Y. Soma *et al.*, “Hybrid assistive limb functional treatment for a patient with chronic incomplete cervical spinal cord injury,” *International Medical Case Reports Journal*, vol. 14, pp. 413–420, 2021, doi: 10.2147/IMCRJ.S306558.
- [22] Y. Chen, Z. Yang, and Y. Wen, “A soft exoskeleton glove for hand bilateral training via surface EMG,” *Sensors (Switzerland)*, vol. 21, no. 2, pp. 1–18, Jan. 2021, doi: 10.3390/S21020578.
- [23] R. Pérez-San Lázaro, I. Salgado, and I. Chairez, “Adaptive sliding-mode controller of a lower limb mobile exoskeleton for active rehabilitation,” *ISA Transactions*, vol. 109, pp. 218–228, Mar. 2021, doi: 10.1016/J.ISATRA.2020.10.008.
- [24] E. Mora-Tola, J. Loja-Duchi, A. Ordóñez-Torres, A. Vázquez-Rodas, F. Astudillo-Salinas, and L. I. Minchala, “Robotic knee exoskeleton prototype to assist patients in gait rehabilitation,” *IEEE Latin America Transactions*, vol. 18, no. 9, pp. 1503–1510, Sep. 2020, doi: 10.1109/TLA.2020.9381791.
- [25] C. G. McDonald, J. L. Sullivan, T. A. Dennis, and M. K. O’Malley, “A Myoelectric Control Interface for Upper-Limb Robotic Rehabilitation following Spinal Cord Injury,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 978–987, Apr. 2020, doi: 10.1109/TNSRE.2020.2979743.
- [26] Y. Bouteraa, I. ben Abdallah, and A. Elmogy, “Design and control of an exoskeleton robot with EMG-driven electrical stimulation for upper limb rehabilitation,” *Industrial Robot*, vol. 47, no. 4, pp. 489–501, Jun. 2020, doi: 10.1108/IR-02-2020-0041.
- [27] L. Zhao, C. Xie, and R. Song, “Design and Validation of a Wearable Hand Exoskeleton System,” *ICARM 2020 - 2020 5th IEEE International Conference on Advanced Robotics and Mechatronics*, pp. 559–563, Dec. 2020, doi: 10.1109/ICARM49381.2020.9195326.
- [28] W. Wang, L. Qin, X. Yuan, X. Ming, T. Sun, and Y. Liu, “Bionic control of exoskeleton robot based on motion intention for rehabilitation training,” <https://doi.org/10.1080/01691864.2019.1621774>, vol. 33, no. 12, pp. 590–601, Jun. 2019, doi: 10.1080/01691864.2019.1621774.
- [29] R. Suppiah, A. Sharma, N. Kim, K. Abidi, and A. Alkaff, “An electromyography-aided robotics hand for rehabilitation - A proof-of-concept study,” *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2020-November, pp. 361–366, Nov. 2020, doi: 10.1109/TENCON50793.2020.9293940.
- [30] F. Adila Ferdiansyah, P. Prajitno, and S. Kusuma Wijaya, “EEG-EMG based bio-robotics elbow orthotics control,” *Journal of Physics: Conference Series*, vol. 1528, no. 1, Jun. 2020, doi: 10.1088/1742-6596/1528/1/012033.
- [31] C. Rodrigues *et al.*, “Comparison of Intramuscular and Surface Electromyography Recordings towards the Control of Wearable Robots for Incomplete Spinal Cord Injury Rehabilitation,” *Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechanics*, vol. 2020-November, pp. 564–569, Nov. 2020, doi: 10.1109/BIOROB49111.2020.9224361.
- [32] I. Yamamoto, M. Matsui, T. Higashi, N. Iso, K. Hachisuka, and A. Hachisuka, “Wrist rehabilitation robot system and its effectiveness for patients,” *Sensors and Materials*, vol. 30, no. 8, pp. 1825–1830, 2018, doi: 10.18494/SAM.2018.1901.
- [33] J. C. Castiblanco, I. F. Mondragon, C. Alvarado-Rojas, and J. D. Colorado, “Assist-As-Needed Exoskeleton for Hand Joint Rehabilitation Based on Muscle Effort Detection,” *Sensors (Basel)*, vol. 21, no. 13, Jul. 2021, doi: 10.3390/S21134372.
- [34] Y. Bouteraa, “Mechatronic design of a biofeedback based-hand exoskeleton for physical rehabilitation,” *2018 15th International Multi-Conference on Systems, Signals, and Devices, SSD 2018*, pp. 1023–1027, Dec. 2018, doi: 10.1109/SSD.2018.8570393.
- [35] A. Cisnal, J. Perez-Turiel, J. C. Fraile, D. Sierra, and E. de La Fuente, “RobHand: A Hand Exoskeleton with Real-Time EMG-Driven

- Embedded Control. Quantifying Hand Gesture Recognition Delays for Bilateral Rehabilitation," *IEEE Access*, vol. 9, pp. 137809–137823, 2021, doi: 10.1109/ACCESS.2021.3118281.
- [36] A. Asokan and M. Vigneshwar, "Design and Control of an EMG-based Low-cost Exoskeleton for Stroke Rehabilitation," *2019 5th Indian Control Conference, ICC 2019 - Proceedings*, pp. 478–483, May 2019, doi: 10.1109/INDIANCC.2019.8715555.
- [37] J. Guo *et al.*, "A soft robotic exo-sheath using fabric EMG sensing for hand rehabilitation and assistance," *2018 IEEE International Conference on Soft Robotics (RoboSoft)*, pp. 497–503, Apr. 2018, doi: 10.1109/ROBOSOFT.2018.8405375.
- [38] A. M. M. Ali, K. Kadir, M. M. Billah, Z. Yosuf, and Z. Janin, "Development of Prismatic robotic arms for rehabilitation by using Electromyogram (EMG)," *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application, ICSIMA 2018*, Apr. 2019, doi: 10.1109/ICSIMA.2018.8688770.
- [39] M. Tiboni, A. Borboni, R. Faglia, and N. Pellegrini, "Robotics rehabilitation of the elbow based on surface electromyography signals," <https://doi.org/10.1177/1687814018754590>, vol. 10, no. 2, pp. 1–14, Feb. 2018, doi: 10.1177/1687814018754590.
- [40] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.
- [41] J. Cornejo *et al.*, "Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-9, doi: 10.1109/IEMTRONICS52119.2021.9422660.
- [42] D. Huamanchahua *et al.*, "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [43] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo, and J. Macuri-Vasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [44] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [45] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.
- [46] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE), 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.

Human Cinematic Capture and Movement System Through Kinect: A Detailed and Innovative Review

Deyby Huamanchahua
 Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y Tecnología
 - UTEC
 Lima, Peru
dhuamanchahua@utec.edu.pe

Jhon Ortiz-Zacarias
 Department of Mechatronics
 Engineering
 Universidad Continental
 Huancayo, Perú
71689110@continental.edu.pe

Yossef Rojas-Tapara
 Department of Mechatronics
 Engineering
 Universidad Continental
 Huancayo, Perú
72448224@continental.edu.pe

Yerson Taza-Aquino
 Department of Mechatronics
 Engineering
 Universidad Continental
 Huancayo, Perú
74239368@continental.edu.pe

Jhon Quispe-Quispe
 Department of Mechatronics
 Engineering
 Universidad Continental
 Huancayo, Perú
73378400@continental.edu.pe

Abstract—There are many techniques focused on capturing the kinematic movement of the human body, all these techniques help in the creation of robotic devices to achieve efficient rehabilitation, many studies provide information that helps the researcher, but these are scattered, each proposing different characteristics, solutions, and methods to obtain body movements. Based on this, the objective is to provide the researcher with a structured matrix that integrates different studies related to the capture of the kinematic movement of the human body, this will allow him to evaluate alternatives where he can choose, analyze the study or characteristics that best contribute to his research topic. As methods, specialized search engines were used that helped structure the matrix. Currently, there is little information on capturing human kinematic movements, through the interaction of different sensors so it was also incorporated into the matrix, also in the proposals is considered captured movements, recognition method, sensor, treatment, and application. In conclusion, the article will provide information to people who want to create robotic devices such as exoskeletons, prostheses, or different suits that help the rehabilitation of the human body. Where it is important to have accuracy in obtaining body movement patterns.

Keywords—Kinect, Kinematic movement, Motion capture, Exoskeletons.

I. INTRODUCTION

The measurement of human kinematics has been a major source of research in areas related to medicine, sports science, and biomedical engineering [1]. An additional use consists of inverse dynamics principles to estimate the forces acting within joints and muscles [2]. Systems for motion capture consist of special hardware and data processing software [3]. Applications include studies to optimize sports performance [5], in medicine for orthopedic analysis or to verify rehabilitation processes, and in 3D animation for applications in film and game production [3]. Among the standardized procedures for detailed body motion capture are stationary and portable optical systems [9], which consist of a network of cameras, usually infrared, video cameras, sensors, and markers (some systems do not require markers) that convert real information into digital data for processing in virtual environments calibrated with each other, in a confined space or laboratory [10]. However, limitations include marker

occlusions when performing certain movements or activities. On the other hand, placement of the markers on the skin requires a significant amount of time and the knowledge of an expert [9]. These are developed from Microsoft Kinect controls and open-source frameworks. Kinect is a motion-sensing input device developed by PrimeSense and Microsoft for the Xbox 360 video game console and later adopted for Windows PCs [3].

IMU (Inertial Measurement Units) inertial measurement systems combine accelerometers, gyroscopes, and magnetometers (compasses) and can solve some of the issues mentioned above. IMUs can be used directly on the body on specific segments to deduce the kinematics of the joints, without suffering from the phenomenon of occlusion, the data can be processed, and the systems are mobile, allowing measurements in the field. Compared to optical systems, the placement of the sensors is less time-consuming and does not require expertise in the process. Despite this, IMUs have certain disadvantages, such as the limitations of direct measurements, a feature that is present in optical systems [9]. In our country, several entities offer interested organizations access to motion capture technologies, useful for the development of film, television, and video game projects, clinical and biomechanical studies, or for pedagogical purposes.

The reason for this review work is to provide detailed information for research who want to develop motion capture methodologies through Kinect. Therefore, this work aims to develop a structured matrix where the inputs of sensor, capture, movement, recognition methods, treatment and application are detailed.

II. METHODOLOGY

For the development of this research, a search was carried out in the database of "Scopus", "IEEE Explore", "Springer link", "Hindawi", "Google Scholar", "Academia Journals", "NCBI", articles from journals such as "BMC", "BMJ", "JMIR" "Indonesian Journal of Electrical Engineering and Computer Science", Learn Tech Lib, MDPI, medRxiv, Taylor and Francis online, as well as collaboration and dissemination platforms such as "Wiley Online Library", and "Semantic Scholar" were also accessed. The publication period of the

selected articles covers from January 2018 to December 2022. The objective of this review is to analyze publications that consider captured movements, sensor types, method of recognition, treatment, and application, the search terms used were: "capture of body movements", and "human kinematic movements through Kinect". A total of 68 articles in English were collected. The result of the review, analysis, interpretation, and integration of the information of each research is presented in the text shown.

III. REVIEW OF PUBLICATIONS

To simplify and structure the information, the capture system and human kinematic movement through Kinect were classified according to the sensors implemented, the approach to the patients' posture, the application of the Kinect recognition method, the treatment applied to patients, and finally, the application of the Kinect in patients. The entire classification is shown in Table I.

TABLE I. KINECT HUMAN KINEMATIC MOTION AND CAPTURE SYSTEM

Reference	Sensor	Captured Motions	Method of Recognition	Treatment	Application
Rawal, A [1]	NE	MS	OTK	MES	NE
Schlagenhauf, F [2]	KV2	MS	VN	NE	MBA
Shair, E [3]	NE	PSB	NE	NE	NE
Yahya, M [4]	RC	NE	RC	NE	NE
Yan, Z [5]	RC	SH	L2D	NE	NE
Amine, E [6]	RC	PH	RC	PH	NE
Hou, J [7]	RC	DO	RC	NE	PPI
Gao, X [8]	NE	CC	NE	NE	NE
Mentiplar, B [9]	KV2	T-C-R	KV2	NE	DRL
Grooten, W [10]	KCT	PH	SQC	NE	MPI
Ma, M [11]	KV2	MS	VN	NE	DN
Bortolini, M [12]	NE	PH	MAS	NE	EPA
Nor, F [13]	KCT	PH	KDJ-KDM	NE	PH
Pachoulakis, I [14]	KCT	MS	U3D	NE	PH
Russo, R [15]	CS	MS	CMI	NE	CTM
Bakhti, K [16]	KV2	MS	NE	DPU	SBC
Kotsifaki, A [17]	KV2	MI	CMI	ECM	NE
Zhang, M [18]	KCT	PH	NE	NE	DDJ
Yang, K [19]	KV2	MS	DTW	RMS	RRV
Teufl, W [20]	IMU	NE	OTK	NE	NE
Huang, C [21]	KV2	SH	OPS	NE	CCA
Bouteraa, Y [22]	KCT - EMG	DN	E3D	DN	EJN
Mubin, O [23]	NE	DM	NE	DN	AVA
Lestari, P [24]	RC	SH	RC-KDM	NE	SHM
Hu, T [25]	KCT-SDP	SH	SVM	NE	RIH
Fuyun, L [26]	KCT	SH	IDS	PH	PH
Jawed, U [27]	KCT	PH	IDS	PH	RTN
Saito, N [28]	KCT	SH	RC	NE	CVM
Ji, X [29]	KCT	M	KCT	NE	AVA
Matos, A [30]	NE	NE	CRA	PH	RTN
Ahn, S [31]	GO	NE	NE	NE	MAR
Ayed, I [32]	NE	NE	KEY-KGX	RMA	DDJ
Heng, S [33]	KCT	CA	KCT	RFA	CKV
Wang, L [34]	KCT	MS	IDS	DN	RMS
Torres, R [35]	KCT	SH	MAB	PH	AIP
Lafayette, T [36]	KV2	MI	MAB-RC	RMA	ACB
Lafayette, T [37]	KV2	DE	KV2- MAB	RMA	ECM
Postolache, G [38]	KCT	MS	KDM	RMA	DDJ
Sarsfield, J [39]	KCT	MS	KCT	DN	DN
Butnariu, S [40]	KCT- SI	MI	NE	DM	SBE
Scano, A [41]	RC	MS	KV2-VN	NE	RTN
Liu, C [42]	RC	PH	KV2	NE	AIP
Tommaso, L [43]	RC	H	KCT	RMA	RTN
Ressman, J [44]	RC	MI	KCT	TMD	SUP

Colombel, J [45]	RC	CC	KV2	NE	RTN
Ma, Y [46]	RC	MS	KV2	NE	EAN
Steinebach, T [47]	RC-SI	MS	KV2-CL7	TME	TPO
Çubukçua, B [48]	RC	H	KV2	NE	RTN
Vilas, M [49]	RC	SH	KV2	TTR	MTD
Sabo, A [50]	RC	PH	KV2	PKN	MTD
Guoliang, L [51]	RC	PH	KCT	HA	RTN
Kavian, M [52]	RC	M	KDM	PAM	RTN
Chen, J [53]	RC	MS	KV2	RMA	RTN
Marin, J [54]	AGM	CC	LC9	AC	RTN
Li, Z [55]	CS	CC	CPM	TME	ARR
Millán, M [56]	AGM	CC	TAS	PKN	MAS
Qiu, S [57]	AGM	MI	DPI	AC	ACB
Hafer, J [58]	AGM	R	VAC	OTS	MTF
Balasubramanian, S [59]	AGM	MS	MSD	HPA	VMG
Chen, Y [60]	AGM	H	BFI	CAA	RTN
Bilesan, A [61]	KV2	PH	KV2	NE	ACB
Meng, X [62]	KCT	SH	KCT	RMA	AVA
Zhang, Z [63]	KCT	SH	KCT	PKN	AIP
Hocking, D [64]	KCT	CC	KCT	RMA	ACB
Jun, Q [65]	KCT	OD	NE	NE	MTD
Sadeghzadehyazdi, N [66]	KCT	SH	MAS	AC	MTD
Xiao, B [67]	KCT	MS	KCT	RMA	RTN
Çubukçu, B [68]	KV2	H	KV2	NE	RTN

Note: RC=RGB-D, CAMERA KV2=Kinect® V2, KCT=Kinect, CS=Smartphone camera, UMI=Inertial Measurement Unit, MS=upper limb, MI=lower limb, PH=human posture, SH=human tracking (full body), EMG=electromyography sensor, SDP=depth sensor, GO=goniometer, SI=inertial sensors, AGM=Accelerometer Gyroscope magnetometer, DN=neurological disabilities (cerebrovascular), DM=motor disabilities, DE=limb discrepancy, PSB=upper arm, DO=object detection, CC=body, T=Trunk, C=hip, R=knee, M=hand, CA=face, H=shoulder, OTK=OptiTrack™, VN=Vicon Nexus, L2D=LiDAR 2D, SQC=Qinematic software™, MAS= Motion Analysis System, KSI=Kinect Skeleton, Joint KDM=Kinect Depth Map, KDJ= Kinect skeleton Joint, U3D=Unity3D, CMI=Capture of Infrared Motion Infrared motion capture, DTW=Modified Dynamic Time Warping Algorithm, OPS=Open Pose (Open Pose) image extraction, E3D=Extraction of EMG images of 3D hand exoskeleton, SVM=Support Vector Machine, IDS=digital images, CRA=camera, KEY=Kinect EyeToy system, KGX=Kinect GestureTek IREX, MAB=MatLab, CL7=Captiv L700, LC9=Logitech C920, CPM=CPM based on REBA, TAS=Trigno Avanti Sensors, DPI=In-house development, VAC=Vicon Opal APDM Inc, MSD=MPU9250 SEN - 14001 board, BFI=BoostFix, MES=upper extremity movement, DPU=diagnose PANU, ECM=lower limb kinematic assessment, RMS=upper limb rehabilitation, RMA=Motor rehabilitation, RFA=facial rehabilitation, TMD=sports medicine test, TME=musculoskeletal disorders, TTR=Transthyretin familial amyloid polyneuropathy(TTR- FAP), PKN=Parkinson's, HA=hemiplegia, PAM=Loss of range of motion in hand joints, AC=Analysis of walking, OTS=Osteoarthritis, HPA=Hemiparesis, CAA=Capuslitas adhesive, MBA=movement with low error in frequency and amplitude of angles, PPI=Precise predictions of instances, DRL=LCA risk detection, MPI=Measures balance, posture and lateral tilt, EPA=Accurate assessment of operator movements and absolute position, CTM=Compare three techniques to measure motion, SBC=Low cost solution, DDJ=Game development, RRV=rehabilitation combined with virtual reality technology, CCA=Correction of left to right confusion and auto occlusion problems in joints, E3N=Game based training for people with neurological disabilities, AVA=Revision of rehabilitation technology with VR/AR, SHM=intelligent human segmentation system, RIH=recognition of human interaction, RTN=rehabilitation, CVM=understanding visitor experiences in museums, MAR=joint movement, CKV=comparison in performance of Kinect V1/V2, AIP=analysis of incorrect posture, ACB=kinematic analysis of individuals with different biotypes, SBE=segmentation of the human body based on Euler angles, SUP=Sitting with one leg, EAN=Evaluation, TPO=Preventive treatment, MTD=Monitoring and data processing, ARR=Rapid and real-time assistance, VMG=Validation of gross movements, NE=Not Specified.

A. Types of sensors

In the research, different solutions for human kinematic motion and capture systems through Kinect were developed to capture the original motion of the patient. They are shown in Fig. 1, The Kinect sensor, which is a three-dimensional distance image sensor, was also used for object detection [28]. For this reason, the Kinect sensor is known as a camera capable of layered results in three dimensions of orientation. In this study, we detected a circle as the first step in object detection. Within this group, partial captures and total captures including limbs can be evidenced without the use of external sensors, for which a high precision work is performed were identified. Table II shows that the Kinect stands out with a percentage of 27.94%, the RGB-D camera was also used with a percentage of 25.00%, and finally Kinect® V2 with a percentage of 16.18%. However, the least used sensors have a percentage of 1.47%, as shown in Table II.

KV2	11	16.18%
AGM	6	8.82%
CS	2	2.94%
KCT - EMG	1	1.47%
KCT- SI	1	1.47%
NE	7	10.29%

Note RGB-D camera (RC), Kinect® V2 (KV2), Kinect (KCT), Smartphone camera (CS), Accelerometer Gyroscope Magnetometer (AGM), Goniometer (GO), Inertial Measurement Unit (IMU), Electromyography Sensor (EMG), Inertial Sensors (SI), Depth Sensor (SDP), depth sensor (SDP).

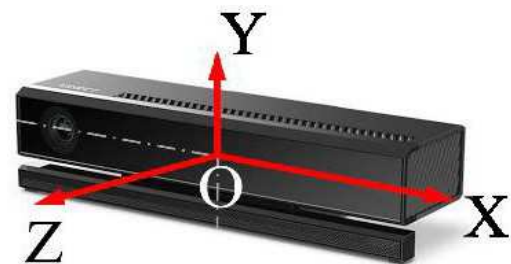


Fig. 1. Kinect sensor coordinates: Nanami Saito, Fusako Kusunoki, Shigenori Inagaki, Hiroshi Mizoguchi [2019]

TABLE II. TYPES OF SENSORS

Sensors	Amount	
	Frequency	Percentage
KCT	19	27.94%
RC	17	25.00%

B. Captured Motions

In this area a review is made of the parts where the approaches are applied through the capture system and human kinematic movement through Kinect, according to the reviewed research, applications, side, and front view were identified with a direct approach with the help of Kinect and RGB-D image processing through programs such as MatLab and the same Windows operating system and the application that stands out the most in the studies are movements in MS with 23.53%; it reflects that the prototypes focus on the realization of human kinematic movement and capture system that focuses on the upper limb.

In addition, 16.18% indicates that it is applied in SH this means that it was based on human tracking (full body) and the 3D coordinates of the anatomical landmarks identified from the skeletal model of the Kinect v2 system during functional tasks are simultaneously recorded with the 3DMC system. Finally, 14.71% focused on the development of movement prototypes in PH, which set the coordinates of the local segments, including Thorax λ and Arm η . Each of the segments is based on the global coordinate as illustrated in Fig. 2 [46].

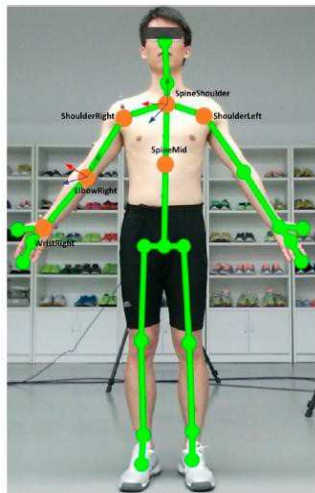


Fig. 2. Full-body human tracking approach: Ye Ma, Dongwei Liu, and Laisi Cai [2020].

The rest of the applications do not exceed 1.47% in Table III. It can also be seen that the most important supported movement is the knee since the mobility of the lower extremities depends on this part of the body.

TABLE III. PERCENTAGES OF STUDIES BASED ON APPROACHES

Approach	Amount	
	Frequency	Percentage
MS	16	23.53%
SH	11	16.18%
PH	10	14.71%
CC	6	8.82%
MI	5	7.35%
H	4	5.88%
M	2	2.94%
DO	2	2.94%
DE	1	1.47%

NE 5 7.35%

Note: Upper limb (MS), human posture (PH), human tracking (full body) (SH), Full Body (CC), lower limb (MI), Shoulder (H), hand (M), knee (R), face (CA), Object Detection (OD), Trunk (T), neurological disabilities (cerebrovascular) (DN), limb discrepancy (DE), hip (C), motor disabilities (DM), upper arm (PSB).

C. Method of recognition

In this area, a review is made of the recognition method of the kinematic capture system and human movement through Kinect. In Table IV, it can be observed that Kinect and Kinect® V2 stand out with a percentage of 14.71%, these devices were used for the better management of recognition in patients, also the RGB-D camera sensor has a percentage of 5.88%. On the other hand, most of the applications of the sensors do not exceed the percentage of 1.47%. The skeleton model is built using the Shotton skeleton detector to locate the different parts of the body and accurately model an articulated structure of connected segments. The obtained 3D skeleton, consisting of 25 joints, represents the human posture as illustrated in Fig. 3 [6].

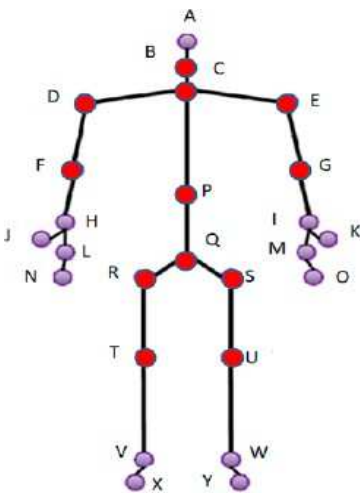


Fig. 3. The 3D joints defining the skeleton: El-Amine Elforaici Mohamed, Chaaraoui Ismail, Bouachir Wassim, Ouakrim, Youssef, Mezghani Neila [2018]

As it is visualized in the orthosis of 4 DoF, it is given to know the ranges of movements of this orthosis, as it is observed the first joint comes to make of the knee of the movement of extensions from 0° to 0° and the flexion from 0° to 120°, and the other joints come to make of the ankle, as the second joint of the movement is pronation from 0° to 10° and supination from 0° to 3°, the third joint of the movement is dorsiflexion from 0° to 25° and plantarflexion from 0° to 45° and finally, the fourth joint of the movement is adduction from 0° to 40° and adduction from 0° to 30° [69].

TABLE IV. PERCENTAGE OF RECOGNITION METHOD

Method of recognition	Amount	
	Frequency	Percentage
KCT	8	13.33%
KV2	6	10.00%
RC	4	6.67%
IDS	3	5.00%
CMI	2	3.33%
KDM	2	3.33%

MAS	2	3.33%
OTK	2	3.33%
VN	1	1.67%
MSD	7	11.67%
NE	8	11.76%

Note: Kinect® V2 (KV2), Kinect (KCT), Rgb-d camara (RC), imágenes digitales (IDS), Kinect Depth Map (KDM), captura de movimiento infrarrojo(CMI), Vicon Nexus (VN), OptiTrack™ (OTK), Kinect EyeToy system (KEY), Kinect GestureTek IREX(KGX), MPU9250 SEN - 14001 board(MSD), Sistema de análisis de movimiento(MAS), extracción de imágenes EMG de exoesqueleto de mano 3D (E3D), Support Vector Machine (SVM), Extracción imágenes (Open Pose) (OPS), MatLab (MAB), CPM basado en REBA (CPM), software Qinematic™ (SQ), Logitech C920 (LC9), cámara(CRA), Trigno Avanti Sensors (TAS), Unity3D (U3D), Desarrollo propio (DPI), BoostFix (BFI), Captiv L700 (CL7), LiDAR 2D (L2D), Algoritmo de deformación del tiempo dinámico modificado (DTW), Kinect Skeleton Joint (KDJ), In-house development (DPI), Modified Dynamic Time Warping Algorithm (DTW), BoostFix (BFI), Captiv L700 (CL7), LiDAR 2D (L2D), MatLab(MAB), Unity3D(U3D).

D. Treatment

In this area, there is a review of the type of treatment for the human kinematic capture and movement system through Kinect, Table V shows the various treatments used for the capture and movement of patients. 13.24% of the reviewed research is based on motor rehabilitation, on the other hand, we have the human posture, with 7.35%, also the most used pull was the limb discrepancy with 5.88%, and then we get the musculoskeletal disorders pulls, gait analysis and Gait analysis and Parkinson's with 4. 41% and finally the rest of the treatments do not exceed 1.47% since they are not so used in patients and Fig. 4, you can see the positioning and the choice of the appropriate angle for the measurement, for which different measures were chosen and processed with ANOVA which are reported along with the degrees of freedom, the statistic, and the statistical significance. Statistical significance was defined by a value of <0.05 [15].



Fig. 4. Addressing increased accuracy of shoulder measurement: Russell R. Russo, Matthew B. Burn, Sabir K. Ismaily, Brayden J. Gerrie, Shuyang Han, Jerry Alexander, Christopher Lenherr, Philip C. Noble, Joshua D. Harris, Patrick C. McCulloch [2017].

TABLE V. TREATMENT PERCENTAGE

Treatment	Amount	
	Frequency	Percentage
RMA	9	13.24%
PH	5	7.35%
DN	4	5.88%
AC	3	4.41%
PKN	3	4.41%
TME	2	2.94%
CAA	1	1.47%
NE	29	42.65%

Note: Postura humana (PH), Rehabilitación motora (RMA), discapacidades neurológicas (cerebrovascular) (DN), Trastornos musculo-esqueléticos (TME), Análisis del caminar (AC), Parkinson(PKN), Capsulitas adhesiva (CAA), hemiplejia(HA), Osteoartritis (OTS), rehabilitación facial (RFA), Perdida del rango de movimiento en las articulaciones de las manos (PAM), evaluación cinemática de miembros inferiores (ECM), discapacidades motoras (DM),

diagnostic PANU, (DPU), Hemiparesia (HPA), movimiento en la extremidad superior(MES), test de medicina deportiva (TMD), Poli neuropatía amiloidea familiar transtiberina(TTR-FAP) (TTR), rehabilitación de miembros superiores(RMS), motor disabilities (DM), diagnose PANU (DPU), Hemiparesis (HPA), movimiento en la extremidad superior(MES), upper limb Rehabilitation (RMS), sports medicine test (TMD), Poli neuropatía amiloidea familiar transtiberina(TTR-FAP) (TTR).

E. Application

Within these applications, it was possible to identify certain advantages and disadvantages which lead to a deeper development of this system, to achieve quality standards and professional development of moving image tracking. The experimental tests validate the real integration of HCI and HRI in one rehabilitation system through the cooperation between game-based software and the developed wireless robotic exoskeleton. A video is provided with this paper to show the rehabilitation training using the designed interactive game (Supplementary Fig 5) [22]. In Table VI, they were used for the rehabilitation of patients that stand out with a percentage of 17.65%, also shown is the development of games, kinematic analysis of individuals with different biotypes, and data processing follow-up with a percentage of 5.88%, finally, the rest of applications that do not exceed 1.47% are not so frequently used in patients.

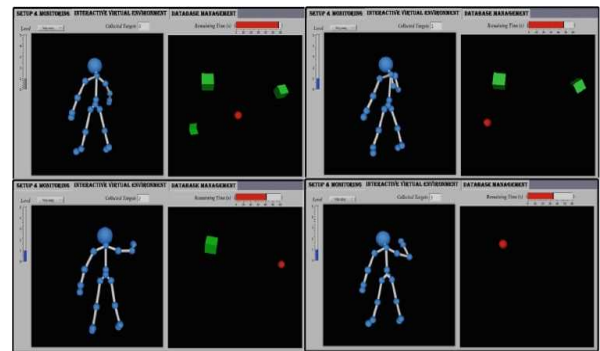


Fig. 5. Interactive virtual environment for rehabilitation with interactive games first stage: Yassine Bouteraa, Ismail Ben Abdallah, Ahmed M. Elmogy [2018].

TABLE VI. APPLICATION PERCENTAGES

Application	Amount	
	Frequency	Percentage
RTN	12	17.65%
ACB	4	5.88%
MTD	4	5.88%
PH	3	4.41%
AIP	3	4.41%
DDJ	3	4.41%
AVA	3	4.41%
DN	2	2.94%
MAR	1	1.47%
EAN	1	1.47%
NE	8	11.76%

Note: Rehabilitation (RTN), game development (DDJ), human posture (PH), monitoring and data processing (MTD), rehabilitation technology review with VR, AR (AVA), kinematic analysis of individuals with different biotypes (ACB), analysis of incorrect posture, neurological disabilities (AIP), (cerebrovascular) (DN), evaluation (EAN), rehabilitation (RTN), Rapid and real-time assistance (ARR), LCA risk detection (DRL), lower limb kinematic assessment (ECM), Game based training for people with neurological disabilities (EJN), accurate assessment of operator movements and absolute position (EPA), motion Analysis System (MAS), movement with low error in frequency and amplitude of angles (MBA), compare three techniques to measure motion (CTM), understanding visitor experiences in museums (CVM), rehabilitation of upper limbs (RMS), rehabilitation combined with virtual reality technology (RRV), low cost solution (SBC), segmentation of the human body based on Euler angles (SBE), intelligent human segmentation system (SHM), sitting with one leg (SUP),preventive treatment (TPO), validation of gross movements(VMG).

IV. CONCLUSION

Due to the increase in the number of users who request precision in the capture of body kinematic movements, for the creation of robotic devices such as exoskeletons of lower and upper extremities, prostheses, body rehabilitation suits, etc. Numerous research groups propose the capture of the kinematic movements of the body through the RGB sensor, this electronic device senses the colors where the images are processed, in this way the body movements are obtained. In the same way, the proposal of motion capture is through the Kinect, this sensor has a three-dimensional reading achieving the capture of all body movements.

The techniques used for the quantification of the gait parameters corresponding to the angles of the human body based on the Kinect and the Microsoft SDK consider that it is necessary to make a direct capture for greater precision and from these patterns propose a tool oriented to specialists to help in the diagnostic processes and determine the progress, success or failure of the therapies applied for rehabilitation and thus make decisions regarding them.

The applications of motion capture are oriented in greater quantity for patients in the rehabilitation of their limbs according to the review carried out, it is for that reason that accuracy is fundamental for a rapid improvement of the patient. The article reviewed various investigations on the capture of kinematic movements. However, there is still room for improvement and the development of more technologies or combinations of sensors to obtain accurate body kinematic movements.

REFERENCES

[1] A. Rawal, A. Chehata, T. Horberry, M. Shumack, C. Chen, and L. Bonato, "Defining the upper extremity range of motion for safe automobile driving," *Clin. Biomech.*, vol. 54, pp. 78–85, May 2018, doi: 10.1016/j.clinbiomech.2018.03.009.

[2] F. Schlagenhaut, S. Sreeram, and W. Singhoose, "Comparison of Kinect and Vicon Motion Capture of Upper-Body Joint Angle Tracking," in *IEEE International Conference on Control and Automation, ICCA*, Aug. 2018, vol. 2018-June, pp. 674–679, doi: 10.1109/ICCA.2018.8444349.

[3] E. F. Shair, S. Ahmad, A. R. Abdullah, M. Marhaban, and S. B. M. Tamrin, "Selection of Spectrogram's Best Window Size in EMG Signal during Core Lifting Task," undefined, 2018.

[4] M. Yahya, J. A. Shah, K. A. Kadir, Z. M. Yusof, S. Khan, and A. Warsi, "Motion capture sensing techniques used in human upper limb motion: a review," *Sensor Review*, vol. 39, no. 4, Emerald Group Publishing Ltd., pp. 504–511, Jul. 15, 2019, doi: 10.1108/SR-10-2018-0270.

[5] "Multisensor Online Transfer Learning for 3D LiDAR-based Human Classification with a Mobile Robot | DeepAI," <https://deepai.org/publication/multisensor-online-transfer-learning-for-3d-lidar-based-human-classification-with-a-mobile-robot> (accessed May 27, 2021).

[6] M. E. A. Elforaici, I. Chaaraoui, W. Bouachir, Y. Ouakrim, and N. Mezghani, "Posture recognition using an RGB-D camera: exploring 3D body modeling and deep learning approaches," *2018 IEEE Life Sci. Conf. LSC 2018*, pp. 69–72, Sep. 2018, Accessed: May 27, 2021. [Online]. Available: <http://arxiv.org/abs/1810.00308>.

[7] J. Hou, A. Dai, and M. Niebner, "3D-SIS: 3D semantic instance segmentation of RGB-D scans," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 4416–4425, doi: 10.1109/CVPR.2019.00455.

[8] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, "Optimized Skeleton-based Action Recognition via Sparsified Graph Regression," *arXiv*, p. arXiv:1811.12013, Nov. 2018, Accessed: May 27, 2021. [Online]. Available: <http://arxiv.org/abs/1811.12013>.

[9] B. F. Mentiplay, K. Hasanki, L. G. Perraton, Y. H. Pua, P. C. Charlton, and R. A. Clark, "Three-dimensional assessment of squats and drop jumps using the Microsoft Xbox One Kinect: Reliability and validity," *J. Sports Sci.*, vol. 36, no. 19, pp. 2202–2209, Oct. 2018, doi: 10.1080/02640414.2018.1445439.

[10] W. J. A. Grooten, L. Sandberg, J. Ressenman, N. Diamantoglou, E. Johansson, and E. Rasmussen-Barr, "Reliability and validity of a novel Kinect-based software program for measuring posture, balance, and side-bending," *BMC Musculoskelet. Disord.*, vol. 19, no. 1, pp. 1–13, Jan. 2018, doi: 10.1186/s12891-017-1927-0.

[11] M. Ma, R. Proffitt, and M. Skubic, "Validation of a Kinect V2 based rehabilitation game," *PLoS One*, vol. 13, no. 8, p. e0202338, Aug. 2018, doi: 10.1371/journal.pone.0202338.

[12] M. Bortolini, M. Gamberi, F. Pilati, and A. Regattieri, "Automatic assessment of the ergonomic risk for manual manufacturing and assembly activities through optical motion capture technology," in *Procedia CIRP*, Jan. 2018, vol. 72, pp. 81–86, doi: 10.1016/j.procir.2018.03.198.

[13] F. A. N. Rashid, N. S. Suriani, and A. Nazari, "Kinect-based physiotherapy and assessment: A comprehensive review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 3, Institute of Advanced Engineering and Science, pp. 1176–1187, Sep. 01, 2018, doi: 10.11591/ijeecs.v11.i3.pp1176-1187.

[14] I. Pachoulakis, N. Papadopoulos, and A. Analyti, "Kinect-based exergames tailored to Parkinson patients," *Int. J. Comput. Games Technol.*, vol. 2018, 2018, doi: 10.1155/2018/2618271.

[15] R. R. Russo et al., "Is digital photography an accurate and precise method for measuring range of motion of the shoulder and elbow?" *J. Orthop. Sci.*, vol. 23, no. 2, pp. 310–315, Mar. 2018, doi: 10.1016/j.jos.2017.11.016.

[16] K. K. A. Bakhti, I. Laffont, M. Muthalib, J. Froger, and D. Mottet, "Kinect-based assessment of proximal arm non-use after a stroke," *J. Neuroeng. Rehabil.*, vol. 15, no. 1, pp. 1–12, Nov. 2018, doi: 10.1186/s12984-018-0451-2.

[17] A. Kotsifaki, R. Whiteley, and C. Hansen, "Dual Kinect v2 system can capture lower limb kinematics reasonably well in a clinical setting: Concurrent validity of a dual camera markerless motion capture system in professional football players," *BMJ Open Sport Exerc. Med.*, vol. 4, no. 1, p. 441, Jul. 2018, doi: 10.1136/bmjsem-2018-000441.

[18] M. Zhang et al., "Recent Developments in Game-Based Virtual Reality Educational Laboratories...," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 1, pp. 138–159, 2018.

[19] K. Yang, L. Peng, L. Tong, R. Liu, and B. Liu, "An Assessment Method for Upper Limb Rehabilitation Training Using Kinect," in *8th Annual IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2018*, Apr. 2019, pp. 949–953, doi: 10.1109/CYBER.2018.8688256.

[20] W. Teufl, M. Miezal, B. Taetz, M. Fröhlich, and G. Bleser, "Validity, test-retest reliability and long-term stability of magnetometer free inertial sensor based 3D joint kinematics," *Sensors (Switzerland)*, vol. 18, no. 7, p. 1980, Jul. 2018, doi: 10.3390/s18071980.

[21] C. C. Huang and M. H. Nguyen, "Robust 3D skeleton tracking based on openpose and a probabilistic tracking framework," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2019, vol. 2019-October, pp. 4107–4112, doi: 10.1109/SMC.2019.8913977.

[22] Y. Bouteraa, I. Ben Abdallah, and A. M. Elmogy, "Training of Hand Rehabilitation Using Low Cost Exoskeleton and Vision-Based Game Interface," *J. Intell. Robot. Syst. Theory Appl.*, vol. 96, no. 1, pp. 31–47, Oct. 2019, doi: 10.1007/s10846-018-0966-6.

[23] O. Mubin, F. Alhajjar, N. Jishtu, B. Alsinglawi, and A. Al Mahmud, "Exoskeletons with virtual reality, augmented reality, and gamification for stroke patients' rehabilitation: Systematic review," *JMIR Rehabilitation and Assistive Technologies*, vol. 6, no. 2, JMIR Publications Inc., Jul. 01, 2019, doi: 10.2196/12010.

[24] P. Lestari and H. P. Schade, "Human Detection from RGB Depth Image using Active contour and Grow-cut Segmentation," in *2019 International Conference on Computer, Control, Informatics, and its Applications: Emerging Trends in Big Data and Artificial Intelligence, IC3INA 2019*, Oct. 2019, pp. 70–75, doi: 10.1109/IC3INA48034.2019.8949571.

[25] T. Hu, X. Zhu, S. Wang, and L. Duan, "Human interaction recognition using spatial-temporal salient feature," *Multimed. Tools Appl.*, vol. 78, no. 20, pp. 28715–28735, Oct. 2019, doi: 10.1007/s11042-018-6074-6.

[26] L. Fuyun, F. Jianping, and F. Mousong, "A Natural Human-Computer Interaction Method in Virtual Roaming," in *Proceedings - 2019 15th International Conference on Computational Intelligence and Security, CIS 2019*, Dec. 2019, pp. 411–414, doi: 10.1109/CIS.2019.00096.

[27] U. Jawed, A. Mazhar, F. Altaf, A. Rehman, S. Shams, and A. Asghar, "Rehabilitation Posture Correction Using Neural Network," Dec. 2019, doi: 10.1109/ICEEST48626.2019.8981676.

[28] N. Saito, F. Kusunoki, S. Inagaki, and H. Mizoguchi, "Novel application of an RGB-D camera for face-direction measurements and object detection: Towards understanding museum visitors' experiences," in *Proceedings of the International Conference on Sensing Technology, ICST*, Dec. 2019, vol. 2019-December, doi: 10.1109/ICST46873.2019.9047675.

[29] X. Ji, Z. Wang, and X. Zhang, "Design of human machine interactive system based on hand gesture recognition," in *Proceedings - 2019 6th International Conference on Information Science and Control Engineering, ICISCE 2019*, Dec. 2019, pp. 250–253, doi: 10.1109/ICISCE48695.2019.00057.

[30] A. C. Matos, T. Azevedo Terroso, L. Corte-Real, and P. Carvalho, "Stereo vision system for human motion analysis in a rehabilitation context," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 7, no. 5–6, pp. 707–723, Nov. 2019, doi: 10.1080/21681163.2018.1542346.

[31] S. Y. Ahn, H. Ko, J. O. Yoon, S. U. Cho, J. H. Park, and K. H. Cho, "Determining the reliability of a new method for measuring joint range of motion through a randomized controlled trial," *Ann. Rehabil. Med.*, vol. 43, no. 6, pp. 707–719, Dec. 2019, doi: 10.5535/arm.2019.43.6.707.

[32] I. Ayed, A. Ghazel, A. Jaume-i-Capó, G. Moyà-Alcover, J. Varona, and P. Martínez-Bueso, "Vision-based serious games and virtual reality systems for motor rehabilitation: A review geared toward a research methodology," *International Journal of Medical Informatics*, vol. 131, Elsevier Ireland Ltd, p. 103909, Nov. 01, 2019, doi: 10.1016/j.ijmedinf.2019.06.016.

[33] S. G. Heng, R. Samad, M. Mustafa, N. R. H. Abdullah, and D. Pebrianti, "Analysis of performance between Kinect V1 and Kinect V2 for various facial part movements," in 2019

- IEEE 9th International Conference on System Engineering and Technology, ICSET 2019 - Proceeding, Oct. 2019, pp. 17–22, doi: 10.1109/ICSEngT.2019.8906419.
- [34] L. Wang, J. Liu, and J. Lan, "Feature Evaluation of Upper Limb Exercise Rehabilitation Interactive System Based on Kinect," *IEEE Access*, vol. 7, pp. 165985–165996, 2019, doi: 10.1109/ACCESS.2019.2953228.
- [35] R. Torres, M. Huerta, R. Clotet, and G. Sagbay, "Body tracking method of symptoms of Parkinson's disease using projection of patterns with Kinect technology," in *IFMBE Proceedings*, 2019, vol. 68, no. 3, pp. 221–226, doi: 10.1007/978-981-10-9023-3_40.
- [36] T. B. D. G. Lafayette, L. F. Colaco, J. M. X. N. Teixeira, C. R. De Vasconcelos, and A. E. F. Da Gama, "Comparison of RGB and HSV color spaces for motion capture and analysis of individuals with limb discrepancy," in *Proceedings - 2019 21st Symposium on Virtual and Augmented Reality, SVR 2019*, Oct. 2019, pp. 178–185, doi: 10.1109/SVR.2019.00042.
- [37] T. B. de G. Lafayette, J. M. X. N. Teixeira, and A. E. F. Da Gama, "Hybrid solution for motion capture with Kinect v2 to different biotypes recognition," in *IFMBE Proceedings*, 2019, vol. 70, no. 1, pp. 249–259, doi: 10.1007/978-981-13-2119-1_39.
- [38] G. Postolache et al., "Serious Games Based on Kinect and Leap Motion Controller for Upper Limbs Physical Rehabilitation," in *Smart Sensors, Measurement, and Instrumentation*, vol. 29, Springer International Publishing, 2019, pp. 147–177.
- [39] J. Sarsfield et al., "Clinical assessment of depth sensor based pose estimation algorithms for technology supervised rehabilitation applications," *Int. J. Med. Inform.*, vol. 121, pp. 30–38, Jan. 2019, doi: 10.1016/j.ijmedinf.2018.11.001.
- [40] S. Butnariu, C. Antonya, and P. Ursu, "Medical recovery system based on inertial sensors," in *Advances in Intelligent Systems and Computing*, 2019, vol. 939, pp. 395–405, doi: 10.1007/978-3-030-16681-6_39.
- [41] A. Scano, R. M. Mira, P. Cerveri, L. M. Tosatti, and M. Sacco, "Analysis of upper-limb and trunk kinematic variability: Accuracy and reliability of an RGB-D sensor," *Multimodal Technol. Interact.*, vol. 4, no. 2, p. 14, Jun. 2020, doi: 10.3390/mti4020014.
- [42] C. H. Liu, P. Lee, Y. L. Chen, C. W. Yen, and C. W. Yu, "Study of postural stability features by using Kinect depth sensors to assess body joint coordination patterns," *Sensors (Switzerland)*, vol. 20, no. 5, p. 1291, Mar. 2020, doi: 10.3390/s20051291.
- [43] L. T. De Paolis and V. De Luca, "The performance of Kinect in assessing the shoulder joint mobility," in *IEEE Medical Measurements and Applications, MeMeA 2020 - Conference Proceedings*, Jun. 2020, pp. 1–6, doi: 10.1109/MeMeA49120.2020.9137213.
- [44] J. Ressman, E. Rasmussen-Barr, and W. J. A. Grooten, "Reliability and validity of a novel Kinect-based software program for measuring a single leg squat," *BMC Sports Sci. Med. Rehabil.*, vol. 12, no. 1, pp. 1–12, May 2020, doi: 10.1186/s13102-020-00179-8.
- [45] J. Colombel, V. Bonnet, D. Danej, R. Dumas, A. Seilles, and F. Charpillet, "Physically consistent whole-body kinematics assessment based on an RGB-D sensor. Application to simple rehabilitation exercises," *Sensors (Switzerland)*, vol. 20, no. 10, May 2020, doi: 10.3390/s20102848.
- [46] Y. Ma, D. Liu, and L. Cai, "Deep learning-based upper limb functional assessment using a single Kinect v2 sensor," *Sensors (Switzerland)*, vol. 20, no. 7, p. 1903, Apr. 2020, doi: 10.3390/s20071903.
- [47] T. Steinebach, E. H. Grosse, C. H. Glock, J. Wakula, and A. Lunin, "Accuracy evaluation of two markerless motion capture systems for measurement of upper extremities: Kinect V2 and Captiv," *Hum. Factors Ergon. Manuf.*, vol. 30, no. 4, pp. 291–302, Jul. 2020, doi: 10.1002/hfm.20840.
- [48] B. Çubukçu, U. Yüzgeç, R. Zileli, and A. Zileli, "Reliability and validity analyze of Kinect V2 based measurement system for shoulder motions," *Med. Eng. Phys.*, vol. 76, pp. 20–31, Feb. 2020, doi: 10.1016/j.medengphy.2019.10.017.
- [49] M. D. C. Vilas-Boas et al., "Validation of a single RGB-D camera for gait assessment of polyneuropathy patients," *Sensors (Switzerland)*, vol. 19, no. 22, Nov. 2019, doi: 10.3390/s19224929.
- [50] A. Sabo, S. Mehdizadeh, K. D. Ng, A. Iaboni, and B. Taati, "Assessment of Parkinsonian gait in older adults with dementia via human pose tracking in video data," *J. Neuroeng. Rehabil.*, vol. 17, no. 1, pp. 1–10, Jul. 2020, doi: 10.1186/s12984-020-00728-9.
- [51] G. Luo, Y. Zhu, R. Wang, Y. Tong, W. Lu, and H. Wang, "Random forest-based classification and analysis of hemiplegia gait using low-cost depth cameras," *Med. Biol. Eng. Comput.*, vol. 58, no. 2, pp. 373–382, Feb. 2020, doi: 10.1007/s11517-019-02079-7.
- [52] M. Kavian and A. Nadian-Ghomsheh, "Monitoring Wrist and Fingers Range of Motion using Leap Motion Camera for Physical Rehabilitation," in *Iranian Conference on Machine Vision and Image Processing, MVIP*, Feb. 2020, vol. 2020-February, doi: 10.1109/MVIP49855.2020.9116876.
- [53] J. F. Chen, C. C. Wang, E. H. K. Wu, and C. F. Chou, "Simultaneous Heterogeneous Sensor Localization, Joint Tracking, and Upper Extremity Modeling for Stroke Rehabilitation," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3570–3581, Sep. 2020, doi: 10.1109/JSYST.2020.2963842.
- [54] J. Marin, T. Blanco, J. de la Torre, and J. J. Marin, "Gait analysis in a box: A system based on magnetometer-free IMUs or clusters of optical markers with automatic event detection," *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–27, Jun. 2020, doi: 10.3390/s20123338.
- [55] Z. Li, R. Zhang, C. H. Lee, and Y. C. Lee, "An evaluation of posture recognition based on intelligent rapid entire body assessment system for determining musculoskeletal disorders," *Sensors (Switzerland)*, vol. 20, no. 16, pp. 1–21, Aug. 2020, doi: 10.3390/s20164414.
- [56] M. Millán and H. Cantú, "Wearable device for automatic detection and monitoring of freezing in Parkinson's disease," *SHS Web Conf.*, vol. 77, p. 05001, 2020, doi: 10.1051/shsconf/20207705001.
- [57] S. Qiu et al., "Towards wearable-inertial-sensor-based gait posture evaluation for subjects with unbalanced gaits," *Sensors (Switzerland)*, vol. 20, no. 4, p. 1193, Feb. 2020, doi: 10.3390/s20041193.
- [58] J. F. Hafer, S. G. Provenzano, K. L. Kern, C. E. Agresta, J. A. Grant, and R. F. Zernicke, "Measuring markers of aging and knee osteoarthritis gait using inertial measurement units," *J. Biomech.*, vol. 99, p. 109567, Jan. 2020, doi: 10.1016/j.jbiomech.2019.109567.
- [59] A. David et al., "Quantification of the relative arm-use in patients with hemiparesis using inertial measurement units," *medRxiv*, p. 2020.06.09.20121996, Jun. 2020, doi: 10.1101/2020.06.09.20121996.
- [60] Y. P. Chen, C. Y. Lin, M. J. Tsai, T. Y. Chuang, and O. K. S. Lee, "Wearable motion sensor device to facilitate rehabilitation in patients with shoulder adhesive capsulitis: Pilot study to assess feasibility," *J. Med. Internet Res.*, vol. 22, no. 7, p. e17032, Jul. 2020, doi: 10.2196/17032.
- [61] A. Bilesan, S. Komizunai, T. Tsujita, and A. Konno, "Improved 3D Human Motion Capture Using Kinect Skeleton and Depth Sensor," *Journal of Robotics and Mechatronics*, vol. 33, no. 6, pp. 1408–1422, Dec. 2021, doi: 10.20965/JRM.2021.P1408.
- [62] T. Huang et al., "Multi-sensor recognition of human posture," *Journal of Physics: Conference Series*, vol. 2137, no. 1, p. 012038, Dec. 2021, doi: 10.1088/1742-6596/2137/1/012038.
- [63] Z. Zhang et al., "Automated and accurate assessment for postural abnormalities in patients with Parkinson's disease based on Kinect and machine learning," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/S12984-021-00959-4/TABLES/5.
- [64] D. R. Hocking et al., "Feasibility of a virtual reality-based exercise intervention and low-cost motion tracking method for estimation of motor proficiency in youth with autism spectrum disorder," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, pp. 1–13, Dec. 2022, doi: 10.1186/S12984-021-00978-1/TABLES/4.
- [65] Q.-J. Xing, Y.-Y. Shen, R. Cao, S.-X. Zong, S.-X. Zhao, and Y.-F. Shen, "Functional movement screen dataset collected with two Azure Kinect depth sensors," *Scientific Data* 2022 9:1, vol. 9, no. 1, pp. 1–17, Mar. 2022, doi: 10.1038/s41597-022-01188-7.
- [66] N. Sadeghzadehyazdi, T. Batabyal, and S. T. Acton, "Modeling spatiotemporal patterns of gait anomaly with a CNN-LSTM deep neural network," *Expert Systems with Applications*, vol. 185, p. 115582, Dec. 2021, doi: 10.1016/j.eswa.2021.115582.
- [67] B. Xiao et al., "Design of a virtual reality rehabilitation system for upper limbs that inhibits compensatory movement," *Medicine in Novel Technology and Devices*, vol. 13, p. 100110, Mar. 2022, doi: 10.1016/J.MEDNTD.2021.100110.
- [68] B. Çubukçu, U. Yüzgeç, A. Zileli, and R. Zileli, "Kinect-based integrated physiotherapy mentor application for shoulder damage," *Future Generation Computer Systems*, vol. 122, pp. 105–116, Sep. 2021, doi: 10.1016/J.FUTURE.2021.04.003.
- [69] J. R. Ortiz-Zacarias, Y. S. Valenzuela-Lino, J. Asto-Evangelista, and D. Huamanchahua, "Kinematic Position and Orientation Analysis of a 4 DoF Orthosis for Knee and Ankle Rehabilitation," 2021 6th International Conference on Robotics and Automation Engineering, ICRAE 2021, pp. 141–146, 2021, doi: 10.1109/ICRAE53653.2021.9657817.
- [70] J. Cornejo et al., "Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–9, doi: 10.1109/IEMTRONICS52119.2021.9422660.
- [71] D. Huamanchahua et al., "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [72] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo and J. Macuri-Vasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [73] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [74] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.
- [75] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE), 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.

Transtibial Electromechanical Prosthesis Based on a Parallel Robot: A Innovate Review

Deyby Huamanchahua
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología-UTEC
 Lima, Perú*

Diego Osoros-Aguilar
*Facultad de Ingeniería Electrónica y
 Eléctrica
 Universidad Nacional Mayor de San
 Marcos, Av. Venezuela Cdra 34 S/N,
 Ciudad Universitaria
 Lima 01, Perú
 diego.osoros@unmsm.edu.pe*

Victor André León-Sales
*Facultad de Ingeniería Electrónica y
 Eléctrica
 Universidad Nacional Mayor de San
 Marcos, Av. Venezuela Cdra 34 S/N,
 Ciudad Universitaria
 Lima 01, Perú
 victorandre.leon@unmsm.edu.pe*

Yadhira S. Valenzuela-Lino
*Department of Mechatronics
 Engineering
 Universidad Continental
 Huancayo, Perú
 73105932@continental.edu.pe*

Harold Huallanca-Escalera
*Department of Mechatronics
 Engineering
 Universidad Peruana de Ciencias
 Aplicadas-UPC
 Lima, Perú
 u201822201@upc.edu.pe*

Abstract— In recent years, the development of prosthetic limbs has increased due to the various studies that have been carried out around Biomechanics. These technological advances have allowed many amputee patients to return to their daily activities and, in some cases, to recover their social life. However, there are still limitations for some types of prostheses such as passive prostheses, mainly those intended for people with transtibial amputations since they do not perform the physiological function of the joints. Therefore, in-depth research on transtibial prostheses has been carried out. The search period spans from the year 2018 to the year 2021. Information was considered in a simplified and well-structured manner on the design of a transtibial electromechanical prosthesis. Therefore, we will analyze its main features related to its design and application in the patient.

Keywords— *Prosthesis, Foot, Ankle, Transtibial, Parallel (keywords)*

I. INTRODUCTION

In the last decades, the development of prostheses for upper and/or lower limbs has been increasing, since due to several biomechanical studies it is possible to allow people with limb amputations to return to a satisfactory social and working life [1]. However, there are passive prostheses with limited functionality, specifically those designed for people with transtibial amputations since they do not perform the physiological function of the joints [1].

Also, because of these prostheses, these individuals often suffer from asymmetries and reduced locomotor performance [2]-[45]; in addition to presenting greater instability during walking on both regular and irregular terrain. Additionally, passive, and quasi-passive prostheses are not sufficient to facilitate forward propulsion of the push-off phase, which causes amputees with this type of prosthesis to use between 10% and 30% more metabolic energy when walking [3]-[46].

In this context, several studies of electromechanical prostheses for the ankle and foot in which actuators, sensors, and controllers are used [4]-[47], were carried out to mimic the biomechanics of the ankle. On the other hand, to improve energy efficiency, elastic components were included especially in the actuators to regulate the stiffness [5] or tune the position and force tracking performance [6]-[48].

On the other hand, lower limb prostheses do not have a one-size-fits-all design, as it is important to consider the anthropometry of the users because leg shape, muscle stiffness, and gait patterns are diverse [7]-[49]; therefore, a possible solution to this parameter selection dilemma is to use a machine learning approach to choose a control system based on the user's performance [8]-[50].

The objective of this article is to systematically review the development and design trends of electromechanical lower limb prostheses and to analyze the various sensors, input signals, mechanisms, treatment, training modes, rehabilitation movements, degrees of freedom, material, actuators, and stage of development of the prostheses. Finally, conclusions will be drawn based on the recent trends shown by the prostheses.

II. METHODOLOGY

For the preparation of the following review article, a detailed database search was performed, taken from repositories and journals such as the following: IEEEExplore, Scopus, Google Scholar, Sage Journals, Hindawi, Mary Ann Liebert, Inc. publisher, MDPI, Springer Link, BMC, ScienceDirect, Research Square. The search period spans from the year 2018 to the year 2021, and to improve the information search, articles in the English language were prioritized and keywords such as Prosthesis + Foot + Ankle + Transtibial + Parallel were used. In the end, a database of 35 scientific articles was obtained to then classify the important subtopics and improve the understanding of the articles and meet the objectives of writing this work, it should be noted that there were about 15 scientific articles of interest to this work which could not be accessed.

III. REVIEW OF PUBLICATIONS

The information was considered in a simplified and well-structured manner regarding the design of a transtibial electromechanical prosthesis. The reviewed publications of the last 4 years on the topic are shown in detail in Table I, considering the most relevant aspects of the publications.

TABLE I. TRANSTIBIAL ELECTROMECHANICAL PROSTHESIS ON A PARALLEL ROBOT UCU

Reference	Input signals	Mechanism	Treatment	Training mode	Rehabilitation Movement	Degrees of freedom (DoF)	Development status of the prosthesis
Chiu, V [2]	PM	P	A	AW	PD	3	3
Gao, F [3]	PM	P	A	PS	PD	3	7
Culver, S [9]	EMG	S	HKA	PS	PD	6	5
Martin, J [10]	PM	S	HKA	PS	RT	1	4
Xu, D [11]	PM	S	HKA	PS	PD	6	5
Mai, J [12]	PM	S	HKA	PS	PD	9	5
Shultz, A [13]	PM	S	HKA	PS	PD	6	5
Shepherd, M [14]	PM	S	A	PS	PD	3	3
Sun, X [15]	PM	S	HKA	PS	PD	2	4
Kumar, D [16]	PM	S	A	PS	PD	4	3
Heremans, F [17]	PM	S	A	PS	PD	NE	4
Frossard, L [18]	EMG	S	HKA	AW	PD	NE	4
Zheng, E [19]	EMG	S	A	PS	PD	5	4
Sathsara, A [20]	EMG	P	A	PS	PD	3	5
Ferreira, C [21]	EMG	P	HKA	AW	NE	NE	4
Agboola-Dobson, A [22]	EMG	NE	HKA	NE	D	2	6
Heremans, F [23]	PM	P	A	NE	PD	1	4
Bartlett, H [24]	MCS	S	A	PS	PD	2	6
Li, J [25]	PM	P	A	PS	PD	3	4
Diteesawat, R [26]	PM	S	HKA	PS	PD-A	3	2
Jin, X [27]	EMG	S	HKA	PS	PD	NE	2
Wen, T [28]	PM	S	A	PS	PD	3	7
Yanggang, F [29]	PM	S	A	PS	PD	3	7
Beomonte, P [30]	PM	S	A	PS	RT	3	7
Moltedo, M [31]	EMG	S	A	AW	PD	1	7
Geeroms, J [32]	PM	P	HKA	AW	PD	NE	2
Maun, J [33]	MCS	S	HKA	PS	RT	6	6
Um, H [34]	PM	NE	A	NE	PD	1	3
Rogers, E [35]	EMG	NE	A	NE	PF	2	6
Bartlett, H [36]	MCS	S	A	PS	PD	2	6
Xu, X [37]	MCS	NE	A	NE	NE	2	3
Guo, S [38]	NE	H	HKA	PS	NE	3	4
Prost, V [39]	MCS	NE	A	PS	PD	1	4
Lecomte, C [40]	MCS	S	A	PS	PD	1	6
Iqbal, N [41]	EMG	P	HKA	NE	NE	NE	3

Note: Abbreviation: EMG: Electromyography, PM: Preset Movements, MCS: Motion Capture System, NE: Not specified; S: Serial, P: Parallel, H: Hybrid, NE: Not specified ; HKA: Hip, Knee and Ankle, A: Ankle; PS: Passive, AW: Assisted Walking, NE: Not specified; RT: Rotational and Translational, PD: Plantarflexion and Dorsiflexion, A: Abduction, PD-A: Plantarflexion, Dorsiflexion and Abduction, PF: Plantarflexion, NE: Not specified ;TRL 1: Basic research, TRL 2: Technology formulation, TRL 3: Applied research, TRL 4: Small scale development, TRL 5: Full scale development, TRL 6: System validated in simulated environment, TRL 7: System validated in real environment, NE: Not specified.

A. Input signals

In this group, the input signals used in the different studies reviewed are mentioned. As shown in Fig. 1 for data collection (input), a motion capture system (MCS) is used to be processed and used in lower limb prosthesis research.

In Table II. Among the most used signals in the study and development of lower limb prostheses are the Preset

Movements (PM) with 51.43% of the developed article, followed by Electromyography (EMG) with 28.57%, and finally, we have the Motion Capture Systems (MCS) with 17.14%. In addition, within this analysis, we have a 2.86% that represents the articles that do not specify data in this field.

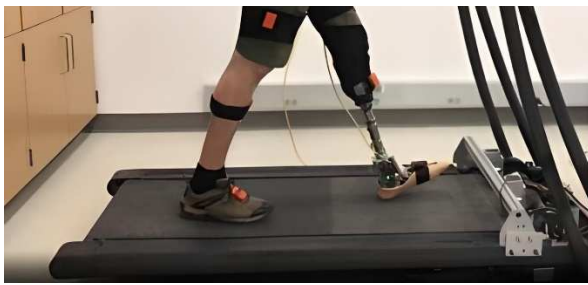


Fig. 1. Photograph of a transtibial amputee walking with an experimental prosthesis using a motion capture system (shown in orange) [25].

TABLE II. PERCENTAGES OF STUDIES BASED ON AN INPUT SIGNAL

Input Signal	Amount	
	Frequency	Percentage
PM	18	51,43%
EMG	10	28,57%
MCS	6	17,14%
NE	1	2,86%

Note: EMG: Electromyography, PM: Preset Movements, MCS: Motion Capture System, NE: Not specified.

B. Mechanism

The mechanisms allow establishing classification criteria according to the kinematic structure, which makes it possible to differentiate them in case they are serial or parallel mechanisms, each of them having certain advantages and disadvantages according to the specified task [42].

Among the main virtues of a serial mechanism are the low rigidity and the wide working space in which it performs. However, it has limitations in its precision, for which the parallel mechanism is beneficial for its high performance in these tasks despite its great rigidity. In addition, there are hybrid models, which are in the development phase, which seek to complement both mechanisms concerning better performance and accuracy. In Table III. According to the reviewed studies of the present research, it was observed that 60% of researchers have preferred to use a serial mechanism (S) and 22.85% have chosen the parallel mechanism (P). In addition, a 2.85% preference for the hybrid model (H) was found. Finally, there is a 14.36% which represents the information not specified in the articles reviewed.

TABLE III. PERCENTAGES OF STUDIES BASED ON KINEMATIC MECHANISMS

Mechanism	Amount	
	Frequency	Percentage
S	21	60,00%
P	8	22,85%
H	1	2,85%
NE	5	14,30%

Note: S: Serie, P: Parallel, H: Hybrid, NE: Not specified.

C. Treatment

For each application of a prosthesis, a treatment is performed on the amputee patient. Therefore, from the reviews carried out, we found targeted treatments for both the leg and the foot. Thus, each prosthesis has a different application depending on the amputation of the patient.

In 42.85%, they applied the treatment to the leg in general, that is, considering the hip, knee, and ankle, and in 57.15% they focused on the treatment of the ankle. In this case, there

were no unspecified. This may be due to the need to be able to specify what type of treatment should be applied.

TABLE IV. PERCENTAGES OF TREATMENTS USED BY THE STUDIES

Treatment	Amount	
	Frequency	Percentage
A	20	57,15%
HKA	15	42,85%

Note: HKA: Hip, Knee and Ankle, A: Ankle.

D. Training Mode

In the process of developing a transtibial prosthesis, it is necessary to put the user into practice to better understand the adaptability of the prosthesis to the user. Thus, in the articles reviewed, two main modes of training are presented: passive training and assisted walking.

In a higher percentage, passive training has been applied in the development of 24 transtibial prostheses. Likewise, but to a lesser extent (14.28%), they opted for assisted walking.

TABLE V. PERCENTAGES OF THE VARIOUS TRAINING MODES

Training mode	Amount	
	Frequency	Percentage
P	24	68,57%
AW	5	14,28%
NE	6	17,15%

Note P: Passive, AW: Assisted walking, NE: Not specified.

E. Rehabilitation Movement

In this section, a review of the movements of the lower limbs, where rehabilitation is applied, is developed. Likewise, the lower limbs present different movements shown in Fig. 2, according to the studies, plantar flexion, and dorsiflexion (PD), rotational and translational (RT), flexion, dorsiflexion, and abduction (PD-A), dorsiflexion (D) and plantarflexion (PF) movements were identified. On the other hand, in table VI with a percentage of 71.42% the most used rehabilitation movement was PD, this infers that the studies are focused on knee movements; in addition, with a percentage of 8.57%, movements belonging to the ankle (RT) were analyzed, likewise, there was a tie between the movements of flexion, dorsiflexion, and abduction (PD-A), dorsiflexion (D) and plantarflexion (PF) with a percentage of 2.86%. Finally, 11.43% of the studies did not specify the rehabilitation movements.

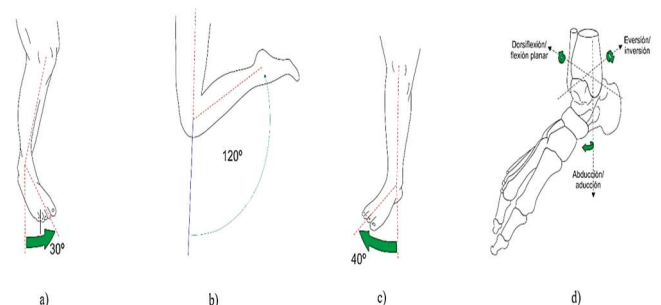


Fig. 2. Identification of exoskeleton joints. a) Internal rotation between 0° to 30°. b) Knee flexion and extension angles from 0° to 120°. c) External rotation between 0° to 40°. d) Foot movement occurs around 3 axes [43].

TABLE VI. PERCENTAGES OF REHABILITATION MOVEMENTS

Rehabilitation Movement	Amount	
	Frequency	Percentage
PD	25	71,42%
RT	3	8,57%
PD-A	1	2,86%
D	1	2,86%
PF	1	2,86%
NE	4	11,43%

Note PD: Plantarflexion and Dorsiflexion, RT: Rotational and translational, PD-A: Plantarflexion, Dorsiflexion, and Abduction, D: Dorsiflexion, PF: Plantarflexion, NE: Not specified.

F. Degrees of Freedom

The DoF can be understood as one of the variables necessary to define all the movements of a body in space. In this section, an analysis and review of the degrees of freedom of the foot, foot-ankle, and transfemoral prosthesis prototypes will be carried out. Fig. 3 shows four configuration systems in transfemoral prosthetic mechanism made up of 1 system of serial mechanism and 3 systems of hybrid mechanisms, in which a comparison is made to opt for a prosthetic mechanism as the reasonable configuration. And Fig. 4 shows the four main movements in a foot-ankle prosthesis.

According to the research carried out, Table I shows that the largest number of lower limb prostheses both foot, foot-ankle and transfemoral, representing 28.57%, have 3-DoF, followed by 17.14% of the prostheses have 1-DoF and 2-DoF, 11.43% of the prostheses have 6-DoF, and 2.86% of the prostheses investigated have 4-DoF, 5-DoF and 9-DoF; finally, 17.14% of the studies do not specify in detail this field.

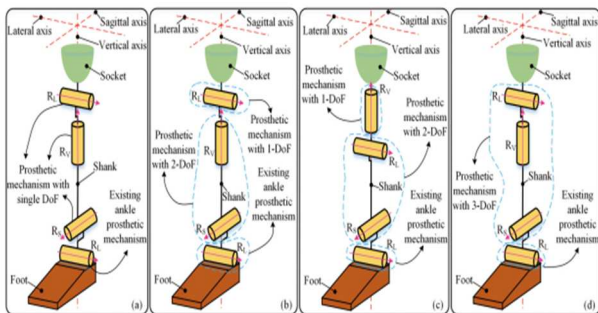


Fig. 3. Configuration systems in transfemoral prosthetic mechanisms. (a) Series mechanism. (b) The hybrid mechanism I: two 1-DoF mechanisms and one 2-DoF mechanism. (c) Hybrid mechanism II: two 1-DoF mechanisms and one 2-DoF mechanism. (d) Hybrid mechanism III: one 3-DoF mechanism and one 1-DoF mechanism [40].

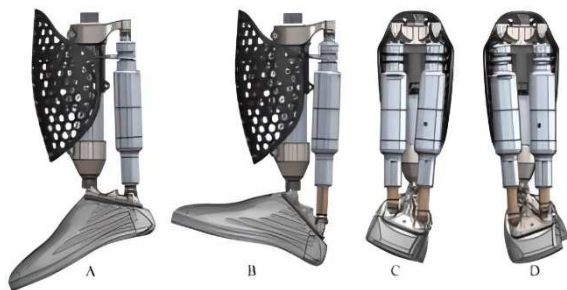


Fig. 4. Four main movements in the foot-ankle prosthesis: (a) Planar flexion upon contraction of both actuators. (b) Dorsiflexion occurs when both actuators are extended. (c) Eversion occurs with the following configuration: the left actuator contracts and the right actuator extends. (d) In inversion with the following configuration, the left actuator extends, and the right actuator contracts [37].

TABLE VII. PERCENTAGE OF DOF OF PROSTHESIS

Degrees of freedom	Amount	
	Frequency	Percentage
1	6	17,14%
2	6	17,14%
3	10	28,57%
4	1	2,86%
5	1	2,86%
6	4	11,43%
9	1	2,86%

G. Development status of the prosthesis

The Development of Technology Readiness Level (TRL) Metrics and Risk Measures manual [44] was important to identify the state of maturity of the research technology, which is made up of 9 levels. Table VIII shows that TRL 4 out of 35 investigations obtained 28.57%; that is to say that the great majority of studies validated the components in laboratory environments; on the other hand, there was a tie between TRL 3 and 6 with a percentage of 17.14%; therefore, it can be affirmed that the first study worked with the proof of concept and the second one elaborated the system or subsystem of a prototype in a relevant environment. On the other hand, TRL5 has a percentage of 14.29%, this figure infers that the research performed validation of components in relevant environments, as well as TRL 7 in which the researchers demonstrated the prototype in a real environment. Finally, TRL 2 has a percentage of 8.57%, from which we conclude that only 3 studies out of 35 used the formulated concept or technological application.

TABLE VIII. PERCENTAGES OF TRL-BASED STUDIES

Technology Readiness Levels (TRL)	Amount	
	Frequency	Percentage
2	3	8,57%
3	6	17,14%
4	10	28,57%
5	5	14,29%
6	6	17,14%
7	5	14,29%

H. Sensors

The present research mentions the use of various sensors used in electromechanical prostheses for lower limbs, the most used being mechanical sensors with a percentage of 37.14% of the state of the art; on the other hand, 37.14% did not specify the type of sensor used in the studies. Finally, 85% of the studies worked with electrical sensors.

TABLE IX. PERCENTAGES OF THE STUDIES ACCORDING TO THE SENSOR USED

Sensor	Amount	
	Frequency	Percentage
MS	19	54,29%
ES	3	8,57%
NE	13	37,14%

Note: MS: Mechanical sensors, ES: Electrical Sensors, NE: Not specified.

I. Actuators

The researchers established solutions to reproduce physiological movements with various actuators that do not damage the joints with sudden movements. Among the actuators, stepper actuators (APP) were identified with a percentage of 51.42%; in addition, the studies worked with servomotors (S) with a percentage of 22.86%; on the other hand, the studies used hydraulic actuators (AH) and linear actuators (AL) with a percentage of 2.86%. And finally, with 20%, the actuators used were not specified.

TABLE X. PERCENTAGES OF THE STUDIES ACCORDING TO THE ACTUATOR USED

Actuator	Amount	
	Frequency	Percentage
APP	18	51,42 %
S	8	22,86%
AH	1	2,86%
AL	1	2,86%
NE	7	20%

Note: APP: Stepper actuator, S: Servomotor, AH: Hydraulic actuator, AL: Linear actuator, NE: Not specified.

J. Materials

In this section, a review is made of the type of material most used in the prototyping and development of lower limb prostheses. Table # shows the different materials that have been used, in the database collected, for the fabrication of the subject under study. The most used material is carbon fiber with 27.03 %, well above the other materials, its hierarchy in the use of this material is reasonable due to its mechanical properties like steel and at the same time, it is light as wood or plastic. Next are Aluminum and Steel with 8.11 %, Nylon Filament with 5.71 %, Titanium, PLA, and Fiberglass with 2.70 %, and 43.24 % do not specify the type of material used. It should be noted that in several articles more than one material is used in the manufacture.

TABLE XI. PERCENTAGE IN THE TYPE OF MATERIAL

Material	Amount	
	Frequency	Percentage
Carbon fiber	10	27,03
Aluminum	3	8,11
Steel	3	8,11
Nylon Filament	2	5,41
Titanium	1	2,70
PLA	1	2,70
Fiberglass	1	2,70
NE	16	43,24

IV. CONCLUSION

Most of the studies presented in this review worked with preset movements and serial mechanisms for the design and development of transtibial prostheses. However, prostheses with EMG signals and parallel mechanisms progressed in a smaller amount due to the complexity of these, despite this, there is still room for improvement and innovation in the various rehabilitation methodologies. On the other hand, the treatments applied are mostly focused on the knee and ankle. In addition, the most used training mode is assisted walking.

Therefore, these characteristics highlighted in other research can be considered for the development of a transtibial prosthesis with a parallel mechanism considering the use of preset movements or EMG signals as input signals. Finally, different treatments and training modes should be considered to validate the prototype design and reach the maximum scale of the technology maturity state (TRL).

We also consider that one of the future technologies to be considered is neuro-prostheses because they are currently a technology with a lot of potentials, but most of them are invasive, i.e., they require surgery. Non-invasive neuro-prostheses do exist, but the devices on the market are not able to provide good efficiency in the day-to-day use of the wearer. However, a neuro-prosthetic leg can provide sensory feedback such as knee angle and touch, resulting in a significant improvement in amputee gait and thus a better quality of life for the disabled.

Finally, from the literature reviewed, the characteristics that impressed us and are considered important for the research are the Mechanism and the Treatment, i.e., the types of prostheses according to the levels of lower limb amputations. The first characteristic, it gives us very even information regarding the choice of serial and parallel mechanisms. The second characteristic shows that almost half of the prostheses developed are for leg amputees, i.e., considering the hip, knee, and ankle, and the other half focus on ankle treatment. Therefore, both information will be very important for the future design and development of an electromechanical transtibial prosthesis based on a parallel robot.

REFERENCES

- [1] S. Alleva, M. G. Antonelli, P. B. Zobel, and F. Durante, «Biomechanical Design and Prototyping of a Powered Ankle-Foot Prosthesis», *Mater.* 2020, Vol. 13, Page 5806, vol. 13, no. 24, p. 5806, Dec. 2020, doi: 10.3390/MA13245806.
- [2] V. L. Chiu, A. S. Voloshina, and S. H. Collins, «An Ankle-Foot Prosthesis Emulator Capable of Modulating Center of Pressure», *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 166–176, Jan. 2020, doi: 10.1109/TBME.2019.2910071.
- [3] F. Gao, Y. Liu, and W. H. Liao, «Implementation and testing of ankle-foot prosthesis with a new compensated controller», *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 4, pp. 1775–1784, Aug. 2020, doi: 10.1109/TMECH.2019.2928892.
- [4] Y. Feng and Q. Wang, «Combining push-off power and nonlinear damping behaviors for a lightweight motor-driven transtibial prosthesis», *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 6, pp. 2512–2523, Dec. 2017, doi: 10.1109/TMECH.2017.2766205.
- [5] H. Jin, D. Yang, H. Zhang, Z. Liu, and J. Zhao, «Flexible Actuator with Variable Stiffness and Its Decoupling Control Algorithm: Principle Prototype Design and Experimental Verification», *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 3, pp. 1279–1291, Jun. 2018, doi: 10.1109/TMECH.2018.2791499.
- [6] E. Sariyildiz, G. Chen, and H. Yu, «A Unified Robust Motion Controller Design for Series Elastic Actuators», *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 5, pp. 2229–2240, Oct. 2017, doi: 10.1109/TMECH.2017.2719682.
- [7] M. Kim and S. H. Collins, «Step-to-Step Ankle Inversion/Eversion Torque Modulation Can Reduce Effort Associated with Balance», *Front. Neurobot.*, vol. 11, no. NOV, Nov. 2017, doi: 10.3389/FNBOT.2017.00062.
- [8] T. C. Wen, M. Jacobson, X. Zhou, H. J. Chung, and M. Kim, «The personalization of stiffness for an ankle-foot prosthesis emulator using Human-in-the-loop optimization», *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 3431–3436, Oct. 2020, doi: 10.1109/IROS45743.2020.9341101.
- [9] S. Culver, H. Bartlett, A. Shultz, y M. Goldfarb, «A Stair Ascent and Descent Controller for a Powered Ankle Prosthesis», *IEEE Trans.*

- Neural Syst. Rehabil. Eng., vol. 26, n.o 5, pp. 993-1002, may 2018, doi: 10.1109/TNSRE.2018.2819508.S
- [10] J. H. Martin, L. Parra, L. E. Duran, A. Posada, y P. Meziat, «Ankle Design with Electromyographic Acquisition System for Transtibial Prosthesis», en 2018 International Conference on System Science and Engineering (ICSSE), New Taipei, jun. 2018, pp. 1-6. doi: 10.1109/ICSSE.2018.8520069.
- [11] D. Xu, Y. Feng, J. Mai, y Q. Wang, «Real-Time On-Board Recognition of Continuous Locomotion Modes for Amputees With Robotic Transtibial Prostheses», IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, n.o 10, pp. 2015-2025, oct. 2018, doi: 10.1109/TNSRE.2018.2870152.
- [12] J. Mai, D. Xu, H. Li, S. Zhang, J. Tan, y Q. Wang, «Implementing a SoC-FPGA Based Acceleration System for On-Board SVM Training for Robotic Transtibial Prostheses», en 2018 IEEE International Conference on Real-time Computing and Robotics (RCAR), Kandima, Maldives, ago. 2018, pp. 150-155. doi: 10.1109/RCAR.2018.8621732.
- [13] A. H. Shultz y M. Goldfarb, «A Unified Controller for Walking on Even and Uneven Terrain With a Powered Ankle Prosthesis», IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, n.o 4, pp. 788-797, abr. 2018, doi: 10.1109/TNSRE.2018.2810165.
- [14] M. K. Shepherd, A. F. Azocar, M. J. Major, y E. J. Rouse, «The Difference Threshold of Ankle-Foot Prosthesis Stiffness for Persons with Transtibial Amputation», en 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob), Enschede, ago. 2018, pp. 100-104. doi: 10.1109/BIOROB.2018.8488075.
- [15] X. Sun, F. Sugai, K. Okada, y M. Inaba, «Design, Control and Preliminary Test of Robotic Ankle Prosthesis», en 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, oct. 2018, pp. 2787-2793. doi: 10.1109/IROS.2018.8594498.
- [16] S. Sahoo, D. K. Pratihari, y S. Mukhopadhyay, «A novel energy efficient powered ankle prosthesis using four-bar controlled compliant actuator», Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 232, n.º 24, pp. 4664-4675, dic. 2018, doi: 10.1177/0954406217753461.
- [17] F. Heremans, B. Dehez, y R. Ronsse, «Design and Validation of a Lightweight Adaptive and Compliant Locking Mechanism for an Ankle Prosthesis», en 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob), Enschede, ago. 2018, pp. 94-99. doi: 10.1109/BIOROB.2018.8487209.
- [18] L. Frossard, B. Leech, y M. Pitkin, «Automated Characterization of Anthropomorphicity of Prosthetic Feet Fitted to Bone-Anchored Transtibial Prosthesis», IEEE Trans. Biomed. Eng., vol. 66, n.o 12, pp. 3402-3410, dic. 2019, doi: 10.1109/TBME.2019.2904713.
- [19] E. Zheng, Q. Wang, y H. Qiao, «Locomotion Mode Recognition With Robotic Transtibial Prosthesis in Inter-Session and Inter-Day Applications», IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, n.o 9, pp. 1836-1845, sep. 2019, doi: 10.1109/TNSRE.2019.2934525.
- [20] A. K. P. Sathara, K. N. D. Widanage, N. Sooriyaperakasam, R. K. P. S. Ranaweera, y R. A. R. C. Gopura, «A Hybrid Powering Mechanism for a Transtibial Robotic Prosthesis», en 2019 Moratuwa Engineering Research Conference (MERCOn), Moratuwa, Sri Lanka, jul. 2019, pp. 447-453. doi: 10.1109/MERCOn.2019.8818842.
- [21] C. Ferreira, F. Dzeladini, A. Ijspeert, L. P. Reis, y C. P. Santos, «Development of a simulated transtibial amputee model», en 2019 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Porto, Portugal, abr. 2019, pp. 1-6. doi: 10.1109/ICARSC.2019.8733636.
- [22] A. Agboola-Dobson, G. Wei, y L. Ren, «Biologically Inspired Design and Development of a Variable Stiffness Powered Ankle-Foot Prosthesis», Journal of Mechanisms and Robotics, vol. 11, n.o 4, p. 041012, ago. 2019, doi: 10.1115/1.4043603.
- [23] F. Heremans, S. Vijayakumar, M. Bouri, B. Dehez, y R. Ronsse, «Bio-inspired design and validation of the Efficient Lockable Spring Ankle (ELSA) prosthesis», en 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, ON, Canada, jun. 2019, pp. 411-416. doi: 10.1109/ICORR.2019.8779421.
- [24] L. B. Harrison, B. E. Lawson, y M. Goldfarb, «Design, Control, and Preliminary Assessment of a Multifunctional Semipowered Ankle Prosthesis», IEEE/ASME Trans. Mechatron., vol. 24, n.º 4, pp. 1532-1540, ago. 2019, doi: 10.1109/TMECH.2019.2918685.
- [25] M. Dong, Y. Kong, J. Li, y W. Fan, «Kinematic Calibration of a Parallel 2-UPS/RRR Ankle Rehabilitation Robot», Journal of Healthcare Engineering, vol. 2020, pp. 1-12, sep. 2020, doi: 10.1155/2020/3053629.
- [26] R. S. Diteesawat, T. Helps, M. Taghavi, y J. Rossiter, «Characteristic Analysis and Design Optimization of Bubble Artificial Muscles», Soft Robotics, vol. 0, n.o 0, pp. 186-199, abr. 2020, doi: 10.1089/soro.2019.0157.
- [27] X. Jin, J. Guo, Z. Li, y R. Wang, «Motion Prediction of Human Wearing Powered Exoskeleton», Mathematical Problems in Engineering, vol. 2020, pp. 1-8, dic. 2020, doi: 10.1155/2020/8899880.
- [28] K. Myunghee, M. Jacobson, X. Zhou, y H.-J. Chung, «The personalization of stiffness for an ankle-foot prosthesis emulator using Human-in-the-loop optimization», en 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, oct. 2020, pp. 3431-3436. doi: 10.1109/IROS45743.2020.9341101.
- [29] F. Yanggang, J. Mai, S. K. Agrawal, y Q. Wang, «Energy Regeneration From Electromagnetic Induction by Human Dynamics for Lower Extremity Robotic Prostheses», IEEE Trans. Robot., vol. 36, n.º 5, pp. 1442-1451, oct. 2020, doi: 10.1109/TRO.2020.2991969.
- [30] P. Beomonte Zobel, S. Alleva, M. G. Antonelli, y F. Durante, «Biomechanical Design and Prototyping of a Powered Ankle-Foot Prosthesis», Materials, vol. 13, n.º 24, p. 5806, dic. 2020, doi: 10.3390/ma13245806.
- [31] M. Moltedo et al., «Walking with a powered ankle-foot orthosis: the effects of actuation timing and stiffness level on healthy users», J NeuroEngineering Rehabil, vol. 17, n.o 1, p. 98, dic. 2020, doi: 10.1186/s12984-020-00723-0.
- [32] J. Geeroms, L. Flynn, V. Ducastel, B. Vanderborght, y D. Lefeber, «On the use of (lockable) parallel elasticity in active prosthetic ankles», en 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, oct. 2020, pp. 3383-3388. doi: 10.1109/IROS45743.2020.9341679.
- [33] J. A. Maun, S. A. Gard, M. J. Major, y K. Z. Takahashi, «Reducing stiffness of shock-absorbing pylon amplifies prosthesis energy loss and redistributes joint mechanical work during walking», J NeuroEngineering Rehabil, vol. 18, n.o 1, p. 143, dic. 2021, doi: 10.1186/s12984-021-00939-8.
- [34] H.-J. Um, H.-S. Kim, W. Hong, H.-S. Kim, y P. Hur, «Design of 3D printable prosthetic foot to implement nonlinear stiffness behavior of human toe joint based on finite element analysis», Sci Rep, vol. 11, n.o 1, p. 19780, dic. 2021, doi: 10.1038/s41598-021-98839-3.
- [35] E. A. Rogers, M. E. Carney, S. H. Yeon, T. R. Clites, D. Solav, y H. M. Herr, «An Ankle-Foot Prosthesis for Rock Climbing Augmentation», IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 41-51, 2021, doi: 10.1109/TNSRE.2020.3033474.
- [36] B. Harrison L., S. T. King, M. Goldfarb, y B. E. Lawson, «A Semi-Powered Ankle Prosthesis and Unified Controller for Level and Sloped Walking», IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 320-329, 2021, doi: 10.1109/TNSRE.2021.3049194.
- [37] X. Xu, X. Xu, Y. Liu, K. Zhong, y H. Zhang, «Design of bionic active-passive hybrid-driven prosthesis based on gait analysis and simulation of compound control method», BioMed Eng OnLine, vol. 20, n.o 1, p. 126, dic. 2021, doi: 10.1186/s12938-021-00962-9.
- [38] M. Song, S. Guo, A. S. Oliveira, X. Wang, y H. Qu, «Design method and verification of a hybrid prosthetic mechanism with energy-damper clutchable device for transfemoral amputees», Front. Mech. Eng., vol. 16, n.o 4, pp. 747-764, dic. 2021, doi: 10.1007/s11465-021-0644-4.
- [39] V. Prost, W. B. Johnson, J. A. Kent, M. J. Major, y A. G. Winter, «Biomechanical Evaluation of Prosthetic Feet Designed Using the Lower Leg Trajectory Error Framework», In Review, preprint, oct. 2021. doi: 10.21203/rs.3.rs-944164/v1.
- [40] C. Lecomte, A. L. Ármannsdóttir, F. Starker, H. Tryggvason, K. Briem, y S. Brynjólfsson, «Variable stiffness foot design and validation», Journal of Biomechanics, vol. 122, p. 110440, jun. 2021, doi: 10.1016/j.jbiomech.2021.110440.
- [41] N. Iqbal, T. Khan, M. Khan, T. Hussain, T. Hameed, y S. A. C. Bukhari, «Neuromechanical Signal-Based Parallel and Scalable Model for Lower Limb Movement Recognition», IEEE Sensors J., vol. 21, n.º 14, pp. 16213-16221, jul. 2021, doi: 10.1109/JSEN.2021.3076114.
- [42] Y. Zhang, G. Cui, and Z. Sun, «Mechanism design, position analysis and simulation of a new six degree-of-freedom serial-parallel manipulator», Proceeding 2009 IEEE 10th Int. Conf. Comput. Ind. Des. Concept. Des. E-Business, Creat. Des. Manuf. - CAID CD'2009, pp. 1005-1009, 2009, doi: 10.1109/CAIDCD.2009.5375422.

- [43] J. R. Ortiz-Zacarias, Y. S. Valenzuela-Lino, J. Asto-Evangelista, and D. Huamanchahua, «Kinematic Position and Orientation Analysis of a 4 DoF Orthosis for Knee and Ankle Rehabilitation», *2021 6th Int. Conf. Robot. Autom. Eng. ICRAE 2021*, pp. 141–146, 2021, doi: 10.1109/ICRAE53653.2021.9657817.
- [44] D. W. Engel, A. C. Dalton, K. Anderson, C. Sivaramakrishnan, and C. Lansing, «Development of Technology Readiness Level (TRL) Metrics and Risk Measures», 2012, Accessed: Apr. 04, 2022. [Online]. Available: <http://www.ntis.gov/ordering.htm>.
- [45] H. Houdijk, E. Pollmann, M. Groenewold, H. Wiggerts, and W. Polomski, «The energy cost for the step-to-step transition in amputee walking», *Gait & Posture*, vol. 30, no. 1, pp. 35–40, Jul. 2009, doi: 10.1016/J.GAITPOST.2009.02.009.
- [46] H. M. Herr and A. M. Grabowski, «Bionic ankle-foot prosthesis normalizes walking gait for persons with leg amputation», *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1728, pp. 457–464, 2012, doi: 10.1098/RSPB.2011.1194.
- [47] Q. Wang, K. Yuan, J. Zhu, and L. Wang, «Walk the walk: A lightweight active transtibial prosthesis», *IEEE Robotics and Automation Magazine*, vol. 22, no. 4, pp. 80–89, 2015, doi: 10.1109/MRA.2015.2408791.
- [48] B. Vanderborght et al., «Variable impedance actuators: A review», *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1601–1614, Dec. 2013, doi: 10.1016/J.ROBOT.2013.06.009.
- [49] M. Kim and S. H. Collins, «Once-Per-Step Control of Ankle Push-Off Work Improves Balance in a Three-Dimensional Simulation of Bipedal Walking», *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 406–418, Apr. 2017, doi: 10.1109/TRO.2016.2636297.
- [50] M. Kim et al., «Bayesian optimization of soft exosuits using a metabolic estimator stopping process», *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, pp. 9173–9179, May 2019, doi: 10.1109/ICRA.2019.8793817.
- [44] J. R. Ortiz-Zacarias, Y. S. Valenzuela-Lino, J. Asto-Evangelista, and D. Huamanchahua, «Kinematic Position and Orientation Analysis of a 4 DoF Orthosis for Knee and Ankle Rehabilitation», *2021 6th International Conference on Robotics and Automation Engineering, ICRAE 2021*, pp. 141–146, 2021, doi: 10.1109/ICRAE53653.2021.9657817.
- [45] J. Cornejo et al., "Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1-9, doi: 10.1109/IEMTRONICS52119.2021.9422660.
- [46] D. Huamanchahua et al., "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [47] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo and J. Macuri-Vasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [48] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [49] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.
- [50] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.

Knee and Ankle Exoskeletons for Motor Rehabilitation: A Technology Review

Deyby Huamanchahua
Department of Electrical and Mechatronics
Engineering
Universidad de Ingeniería y Tecnología -
UTECH
Lima, Perú
dhuamanchahua@utec.edu.pe

Cesar Luciano Otarola-Ruiz
Department of Bioengineering
Universidad de Ingeniería y Tecnología -
UTECH
Lima, Perú
cesar.otarola@utec.edu.pe

Ana Quispe-Piña
Department of Biomedical Engineering
Universidad Tecnológica del Perú
Lima, Perú
u18303812@utp.edu.pe

Elvis J. de la Torre-Velarde
Department of Mechatronics Engineering
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
u201822520@upc.edu.pe

Abstract— Exoskeletons have been around since 1890 when a passive device was developed to assist movement. Although the exoskeleton initially focused on the whole body, it has become increasingly focused on a more specific area, which makes rehabilitation work more feasible and, consequently, less costly. Therefore, the purpose of the article is to present a systematic review of trends in the technical area of knee and ankle exoskeletons for rehabilitation purposes. The objective is to provide the interested researcher with a structured matrix that integrates essential components and mechanisms in the development of knee and ankle exoskeletons. As a methodology, specialized databases in biomechanics were used that aided in the collection of research conducted between 2019 and 2022. The filtering process resulted in the selection of 37 investigations. Finally, it was concluded that innovations in exoskeletons contribute to patient rehabilitation and that there is still room for further improvement in the implementation of projects on a commercial scale and the development of new exoskeletons based on previous research.

Keywords— ankle exoskeleton, knee exoskeleton, rehabilitation, *TRL*

I. INTRODUCTION

The development of robotics is becoming more and more surprising, to such an extent that an exoskeleton previously considered bulky and heavy for the user, today feels as if it were part of his own body [1]. Exoskeletons are mechanical structures that have different applications such as increasing the user's strength, supporting body weight, and helping to lift heavy loads, among others. [2].

Passive exoskeletons are distinguished by their proven efficacy in preventing injuries caused by performing false movements or maintaining prolonged work postures [3].

In a study of patients with diseases related to lower limb muscle weakness, they indicated that knee-ankle-foot orthoses (KAFOs) offer stability and reduce the cost of standing and walking, but for better performance, they often need to be custom-made. More than 76% of polio survivors reported using

a custom-made KAFO of which 74% obtained greater foot stability and 69% obtained improved walking performance. [4]

Robotic-assisted gait training (HIL) has potential benefits over the process of interactively generating automated exoskeleton assistance patterns to reduce physiological cost function while the user is wearing the device [5]. However, compared to rehabilitation, HIL has not demonstrated better results in combining accelerometer and GPS features to assess community mobility in users [6].

Recently, this problem was addressed through PIGRO, which allows the design of exoskeletons for robotic neurorehabilitation training [7]. In addition to PIGRO, there are other advances in the exoskeleton actuation system such as T-FLEX, which was the first ankle exoskeleton to be experimentally validated in people with stroke. [8].

In this sense, the main objective of the article is to perform a systematic review of the latest advances in the development of knee and ankle exoskeletons for the rehabilitation of healthy users as well as disabled individuals. This analysis of exoskeleton characteristics is composed of sensors, mechanism, material, treatment, rehabilitation movements, and development status. Finally, from the study, conclusions will be developed.

II. METHODOLOGY

For the development of this research, a search query was made in the IEEE Xplore database, Search database, and Science Direct, and access was also gained to collaboration and dissemination platforms such as ResearchGate, National Center for Biotechnology Information (NCBI), Multidisciplinary Digital Publishing Institute (MDPI), SCITEPRESS, as well as access to articles presented in some congresses such as: "Academia Journals International Research Congress, Celaya, 2021". In the search for information, keywords such as an exoskeleton, rehabilitation, knee, and ankle were used. Articles published from 2019 – to 2022 were limited, obtaining 37 articles.

III. REVIEW OF PUBLICATIONS

Articles on Knee and ankle exoskeletons for rehabilitation were classified according to sensors, mechanisms, type of

material used for manufacturing, type of treatment, rehabilitation movements, and exoskeleton development status (TRL).

TABLE I. KNEE AND ANKLE EXOSKELETONS FOR REHABILITATION

Reference	Sensors	Mechanism	Material	Treatment	Rehabilitation movements	TRL
MacLean, M [1]	KS	NE	NE	AR	R	5
Deo, I [2]	NE	S	stainless steel 304	KAR	RT	3
Yu, S [3]	IMU y EMG.	NE	Carbon fiber	AR	R	4
Dae-Hoon, M [5]	DS	S	Resin - aluminum	KS	RT	5
Lonini, L [6]	PRS	S	NE	KAR	NE	3
Belforte, G [7]	PSS	NE	Steel-Aluminum	NR	RT	4
Divekar, N [8]	FS	S	NE	KAR	RT	4
Rezazadeh, S [9]	IMU	SP	NE	KAS	RT	8
Farkhatdinov, I [10]	EMG	P	NE	KS	R	4
Gomez, D [11]	IS	P	NE	AR	T	6
Luviano, D [12]	NE	S	Aluminum	AR	T	2
Lyu, M [13]	IMU y EMG.	S	Aluminum	NR	R	5
Freitas, B [14]	NE	S	Aluminum	KS	RT	3
Chen, J [15]	FS	S	Aluminum	KS	R	3
Borzuola, R [16]	NE	S	Aluminum	KAS	RT	2
Chen, C [17]	IMU	NE	Aluminum	KS	R	3
Zhou, T [18]	NE	S	Carbon fiber	KR	R	3
Grimmer, M [19]	FS	S	NE	AR	T	4
Lee, S [20]	NE	S	NE	KAR	RT	4
Wang, Z [21]	FS	S	Carbon fiber	KR	R	4
Hidayah, R [22]	FS	S	Nylon	KAS	R	4
Han, H [23]	FS	NE	Titanium	KAR	RT	3
Jammeli, I [24]	FS	S	NE	KR	R	4
Bougrinat, Y [25]	FS	NE	Carbon fiber	AS	T	5
Franks, P [26]	NE	S	Carbon fiber	KAR	RT	3
Sanz-Morère, C [27]	MS	S	NE	KAR	RT	6
McCain, E [28]	EMG	S	Carbon fiber	NR	T	4
Minchala, L [29]	EMG	S	Aluminum	KAR	RT	3
Mu, J [30]	FS	S	Carbon fiber	KAS	RT	3
Wang, Z [31]	IMU, EMG y FS	S	Nylon-Carbon fiber	KR	R	3
Liu, X [32]	IMU y PS	S	NE	KR	R	4
Seo, K [33]	IMU y PR	S	NE	KAR	T	4
Lee, T [34]	IMU, EMG y FS	S	NE	KAS	RT	4
Bryan, G [35]	FS	S	Aluminum – Titanium	KAS	RT	3
Zhao, W [36]	NE	S	NE	KR	R	2
Rosenberg, M [37]	EMG	S	NE	KS	RT	2
Mouzo, F [38]	FS	NE	NE	KR	R	3

Note: Abbreviation: FS: Force sensing, KS: Kinematic sensors, IMU: Inertial measurement unit, EMG: Electromyography sensor, MS: Measurement sensor, DS: Distance sensing, PRS: Portable sensors, PSS: Position sensors, PR: Pressure sensors, IS: Inertial sensor, S: Serials, P: Parallels, AR: Ankle rehabilitation, KR: knee rehabilitation, KAR: Knee and ankle rehabilitation, AS: Ankle support, KS: Knee support, KAS: Knee and ankle support, NR: Neurorehabilitation, R: Knee, T: Ankle, RT: Knee and ankle, TRL 1: Basic Research, TRL 2: Technology Formulation, TRL 3: Applied research, TRL 4: Small-scale development, TRL 5: Real-scale development, TRL 6: Valid system in a simulated environment, TRL 7: Validated system in a real environment, TRL 8: Validated system and certified for a real environment, NE: Unspecific.

A. Sensors

Several types of sensors were found in this group: force sensors (FS), kinematic sensors (KS), IMU sensors (IMU), EMG sensors (EMG), measurement sensors (MS), distance sensors (DS), portable sensors (PRS), position sensors (PSS), pressure sensors (PR) and inertial sensors (IS). Table II describes the characteristics of the mentioned types of sensors.

TABLE II. PERCENTAGE OF STUDIES BASED ON THE USE OF SENSORS

Sensors	Amount	
	Frequency	Percentage
FS	13	35,13%
IMU	8	21,62%
EMG	8	21,62%
DS	1	2,70%
PRS	1	2,70%
KS	1	2,70%
MS	1	2,70%
IS	1	2,70%
PSS	1	2,70%
PR	1	2,70%

Note: FS: Fonce sensing, IMU: IMU sensor, EMG: EMG Sensors, DS: Distance sensing, PRS: Portable sensors, KS: Kinematic sensors, MS: Measurement sensor, IS: Inertial sensor, PSS: Position sensors, PR: pressure sensors

It is important to consider the type of sensor we are going to use and be able to develop a suitable exoskeleton for rehabilitation. As can be seen in Table II, most of the studies used the force sensor with 35.13% [8], followed by IMU and EMG sensors with 21.62% [9] and the rest at 2.7%.

B. Mechanism

In this area, we find two general descriptions of the knee and ankle design: Serial (S) as shown in Fig. 1, and Parallel (P).

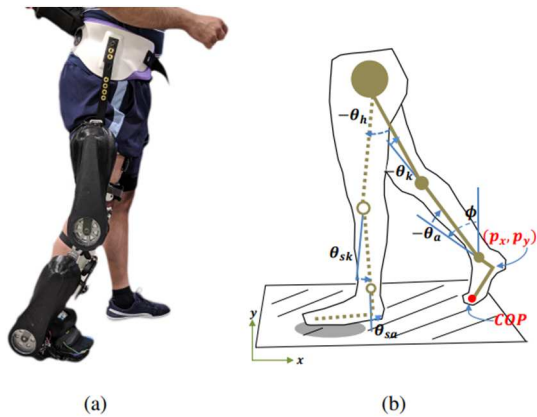


Fig. 1. (a) Comex exoskeleton worn by a healthy user. (b) Kinematic model of the body reproduced from. COP denotes the center of pressure, solid links denote the stance leg, and dashed links denote the swing leg. Source: Nikhil V. Divekar, Jianping Lin, Christopher Nesler, Sara Borboa and Robert D. Gregg, A Potential Energy Shaping Controller with Ground Reaction Force Feedback for a Multi-Activity Knee-Ankle Exoskeleton, USA: 8th IEEE International Conference on Biomedical Robotics and Biomechatronics, [2020]

In the construction of exoskeletons, the mechanism is considered a fundamental characteristic for the simulations and the obtaining of parameters that allow the evaluation of its

effectiveness. These mechanisms were grouped into two large groups, which are described below.

TABLE III. TYPES OF MECHANISM

Mechanism	Description
S (Serial)	Exoskeletons with a kinematic structure take the form of an open-loop chain.
P (Parallel)	Exoskeletons were developed to increase torque and joint work.

Table IV shows that most of the studies used a serial mechanism with 75.67% while the use of a parallel mechanism is 8.11% [9, 10, 11] This trend could be explained by the relative ease of implementation and mathematical modeling of a serial mechanism compared to one of a parallel nature.

TABLE IV. PERCENTAGES OF STUDIES BASED ON THE MECHANISM

Supported movements	Amount	
	Frequency	Percentage
S	28	75,67%
P	3	8,11%

Note: S: Serial, P: Parallel

C. Type of material

This section reviews the type of material typically used to develop exoskeletons. Table V shows in detail the different materials used in the process of developing robotic parts. The most used material is carbon fiber, 18.92% of the reviewed research uses this material which is the lightest and offers higher long-term performance. Another material of greater use in the elaboration of the components of the exoskeletons with 16.22% is aluminum, this material protects from oxidation and is very common in its use in the development of prototypes [12-17]., due to this, it exists as one of the most used materials in the development of exoskeletons. In the research carried out, structures of composite materials were found such as steel with aluminum in a percentage of 5.41%; aluminum with resin in a percentage of 2.7%; nylon with carbon fiber in a percentage of 2.7%; aluminum with titanium in a percentage of 2.7%. The structures composed of only one type of material in low percentage are steel 2.7%, nylon 2.7%, and titanium 2.7%. Finally, 43.24% of the research carried out does not specify the type of material used.

TABLE V. PERCENTAGES OF THE TYPE OF MATERIAL

Material	Amount	
	Frequency	Percentage
Carbon Fiber	7	18.92%
Aluminum	6	16.22%
Steel-Aluminum	2	5.41%
Aluminum-Resin	1	2.70%
Nylon-Carbon Fiber	1	2.70%
Aluminum-Titanium	1	2.70%
Stainless steel 304	1	2.70%
Nylon	1	2.70%
Titanium	1	2.70%
N.E.	16	43.24%

D. Treatment

This section analyzes the purpose of the exoskeleton and/or the type of treatment they offer. From the research conducted and collected in Table VI, it can be seen that 24.32% of the designs are focused on designed to cover the area of knee and ankle rehabilitation (KAR); 18.92% are only focused on knee rehabilitation (KR), 16.22% focus on the assistance or support in knee movements (KS); in the same way 16.22% focus on the assistance and support of the knee and ankle (KAS), when patients exercise heavy activities; 13.51% only focus on ankle rehabilitation (AR) for people with movement difficulties in this joint; 8.11% are focused on neurorehabilitation (NR) for people who have suffered a brain injury that prevents them from moving their lower extremities; finally, 2.7% of the exoskeleton is oriented towards assistance and support for ankle movements (AS).

TABLE VI. PERCENTAGES OF THE TYPE OF TREATMENT

Treatment	Amount	
	Frequency	Percentage
KAR	9	24.32%
KR	7	18.92%
KS	6	16.22%
KAS	6	16.22%
AR	5	13.51%
NR	3	8.11%
AS	1	2.70%

Note AR: Ankle rehabilitation, KR: knee rehabilitation, KAR: Knee and ankle rehabilitation, AS: Ankle support, KS: Knee support, KAS: Knee and ankle support, NR: Neurorehabilitation.

E. Rehabilitation movements

A review was carried out to identify the rehabilitation movements performed; firstly, it was identified which joints were responsible for the movements; thus, knee (R), ankle (T), and knee-ankle (RT) applications were identified. In addition, 24.32% of jobs were found to have additional applications at the hip; however, this was not considered in the study for 2 reasons. First, the actuation of the exoskeleton was dependent on the knee or ankle [18, 19] leaving hip rehabilitation in the background. In other articles, the actuation was not purely focused on the hip but had rehabilitation movements in both the knee and ankle simultaneously [7, 10], so these other two joints were prioritized for the paper.

In most of the studies, it was found that most rehabilitation movements focused on both joints (RT) with 43.24%; in 37.84% they focused exclusively on the knee (R) and, finally, in 16.22%, the work focused only on the ankle (T). These figures suggest that there is a tendency to design exoskeletons for joint rehabilitation of the knee and ankle.

More specifically, the movements performed on these two joints will be detailed. For a better understanding, a figure with the movements found in the review of articles is presented.

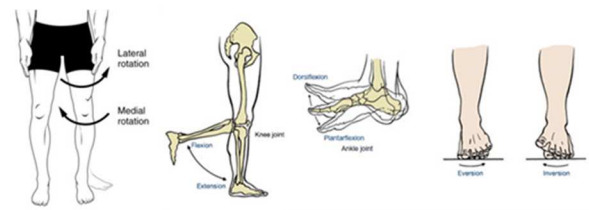


Fig. 2. Knee rotation, flexion, and extension. Ankle dorsiflexion, plantarflexion, eversion, and inversion. Sources: Howell, D., Knee Pain Wringing Out – Gait Deviation, United States: Damien Howell Physical Therapy, [2018]. Thompson, N., Muscles That Move the Leg, United States: American Council of Exercise, [2017].

In knee rehabilitation movements, we found that most of the research, 81.08%, focuses on knee flexion-extension (RF), this could be explained by the fact that it is one of the most common rehabilitation movements since it is used in multiple situations such as walking and sit-to-stand movement. On the other hand, only one article was found in which knee rotation (RR) is used.

Regarding ankle rehabilitation, 48.65% focus on plantarflexion (PT), followed by 37.84% on dorsiflexion (TD). These figures are interesting, as both movements are inverse, such as flexion and extension of the knee, but in this type of movement, both are usually worked separately. Finally, inversion and eversion (IT) movements represent 8.11% of items.

TABLE VII. PERCENTAGES OF STUDIES

Rehabilitation movements	Amount	
	Frequency	Percentage
RF	30	81.08%
RR	1	2,70%
TI	3	8,11%
TD	14	37,84%
TP	18	48,65%

Note: RF: Knee flexion-extension, RR: Knee rotation, TI: Ankle inversion-eversion, TD: Ankle dorsiflexion, TP: Ankle plantarflexion.

F. State of exoskeleton development

One of the aspects evaluated was the state of development of the project, so the level of progress of each exoskeleton concerning its stated objectives was seen. In this sense, the TRLs were grouped into 4 stages: Ideation, which corresponds to TRLs 1 and 2; Experimentation, which covers TRLs 3 and 4; Simulation, which corresponds to TRLs 5 and 6; and Validation, with TRLs 7 and 8.

The 10.81% of the articles found to correspond to the earliest stage, where the prototype is developed and detailed. Experimentation has the highest number of articles with 70.27%; at this point, at least one proof of concept or validation in a laboratory has been performed. The next figure corresponds to simulation, where components are tested in a relevant environment usually simulated; this stage could be considered the most developed before commercialization and represents 16.22% of the articles. Finally, there is a validation stage where the product reaches the highest degree of feasibility, so that it can be marketable; this only accounts for 2.70%.

TABLE VIII. PERCENTAGES OF STUDIES BASED ON THE ACTUATOR EMPLOYED

State of development	Amount	
	Frequency	Percentage
Ideation	4	10,81%
Experimentation	26	70,27%
Simulation	6	16,22%
Validation	1	2,70%

G. Type of control

It is crucial to consider the type of control because the coupling between it and the mechanical design represents the efficiency of the mechanism, thus being able to develop an appropriate exoskeleton. As can be seen in Table IX, most of the studies focused on PID with 24.32% and the rest below 3%.

TABLE IX. PERCENTAGES OF STUDIES BASED ON TYPES OF CONTROL

Supported movements	Amount	
	Frequency	Percentage
PID	9	24,32%
PD	1	2,70%
PC	1	2,70%
ROS	1	2,70%
POILC	1	2,70%
MPC	1	2,70%
PID, CC, DC	1	2,70%
RNNs	1	2,70%
PSMC, PID, and FUZZY	1	2,70%
NE	12	32,43%

Note: PID: Proportional Integral Derivative, PD: Proportional-derivative, PC: Position controller, ROS: Robot Operating System, POILC: Parameter Optimal Iterative Learning Control, MPC: Model predictive controller, CC: Centralized Controller, DC: Distributed Controllers, RNNs: Neural Network, PSMC: Proxy-based sliding mode control, N.E.: Not specified.

H. Type of controller

This section considers the type of controller that allows the exoskeletons to perform actions on the actuators based on the data obtained from the sensors. Although most exoskeletons process signals and data in computers with higher capacities, some microcontrollers, and microprocessors act as intermediaries to manipulate the exoskeleton. Based on the articles reviewed and compiled in Table X, 10.81% of the exoskeletons are controlled by FPGAs; 8.11% use STM323F407 type ARM microcontrollers; another 8.11% use dSPACE processors; 5.41% of the exoskeletons are controlled by Raspberry pi 3 microprocessors. Exoskeletons that use controllers in 2.70% are the Arduino MEGA 2560, mbed LCP1864, and PC104; in the same percentage, some exoskeletons use two controllers which are the STM32F407 and Raspberry Pi 4B. Finally, 56.76% of reviewed articles do not mention the type of controller used by their exoskeletons.

TABLE X. PERCENTAGES OF STUDIES BASED ON TYPE CONTROLLER

Controller type	Amount	
	Frequency	Percentage
FPGA	4	10.81%
STM323F407	3	8.11%
dSPACE	3	8.11%
Raspberry Pi 3	2	5.41%
Arduino MEGA 2560	1	2.70%
mbed LCP1864	1	2.70%
PC104	1	2.70%
STM32F407- Raspberry Pi 4B	1	2.70%
NE	21	56.76%

Note: FPGA: Field-programmable gate array, N.E.: Unspecified.

IV. CONCLUSIONS

Since their creation, exoskeletons have served to improve the quality of life through movement assistance. Therefore, they present great potential in the field of rehabilitation. The present article reviewed multitype of research around rehabilitation in the knee and ankle joints. Multiple functions must be improved, as seen in the state of development of the projects, where most of the exoskeletons are still in the early stages of research.

Regarding the structural characteristics of the exoskeleton, it is observed that the most used sensor is the force sensor, the mechanism is primarily serial due to the relative ease for the development of mathematical models and the predominant type of material is aluminum. Regarding rehabilitation, it is observed that the treatment focuses mainly on both joints and that the most used movements are flexion-extension in the knee and plantarflexion in the ankle.

In summary, it is hoped that exoskeletons will continue to be developed in the identified areas of opportunity and that in the future, innovations in the form of rehabilitation with exoskeletons will be presented. It is also expected that the review will serve as a tool to see trends in the way exoskeleton projects have been developed in knee and ankle rehabilitation, and that new projects in this line can be developed.

REFERENCES

- [1] Mhairi K. MacLean and Daniel P. Ferris, "Energetics of Walking With a Robotic Knee Exoskeleton," Journal of Applied Biomechanics, pp. 320-326, 2019
- [2] Ignatius Deo Putranto, Eka Budiarto, Kidarsa, Lydia Anggraeni, "Mechanical Design of Knee and Ankle Exoskeleton to Help Patients with Lower Limb Disabilities," Mechanical Design of Knee and Ankle Exoskeleton to Help Patients with Lower Limb Disabilities, pp. 17-26, 2019.
- [3] Shuangyue Yu, Tzu-Hao Huang, Dianpeng Wang, Brian Lynn, Dina Sayd, Viktor Silivanov, Young Soo Park, Yingli Tian and Hao Su, "Design and Control of a High-Torque and Highly-Backdrivable Hybrid Soft Exoskeleton for Knee Injury Prevention during Squatting," IEEE ROBOTICS AND AUTOMATION LETTERS, 2019.
- [4] Raijmakers, Bart et al. "Use and Usability Of Custom-Made Knee-Ankle-Foot Orthoses In Polio Survivors with Knee Instability: A Cross-Sectional Survey." Journal of rehabilitation medicine vol. 54 jrm00261. 14 Feb. 2022, doi:10.2340/jrm. v53.1122

- [5] Hong Han, Wei Wang, Fengchao Zhang, Xin Li, Jianyu Chen, Jianda Han, and Juanjuan Zhang, "Selection of Muscle-Activity-Based Cost Function in Human-in-the-Loop Optimization of Multi-Gait Ankle Exoskeleton Assistance," *IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING*, vol. 29, pp. 944-952, 2021.
- [6] L. Lonini; N. Shawen; S. Hoppe-Ludwig; S. Deems-Dluhy; C. Mummidisetty, Y. Eisenberg, A. Jayaraman, "Combining Accelerometer and GPS Features to Evaluate Community Mobility in Knee Ankle Foot Orthoses (KAFO) Users," *IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING*, vol 29, pp. 1386-1393, 2021
- [7] G. Belforte, T. Raparelli, G. Eula, S. Sirolli, S. Appendino, G. Geminiani, E. Geda, M. Zettin, R. Virgilio and K. Sacco, "An Active Exoskeleton Called P.I.G.R.O. Designed for Unloaded Robotic Neurorehabilitation Training," *Medical Robotics - New Achievements*, 2019.
- [8] D. Gomez, F. Ballen, P. Barria, R. Aguilar, J. Azorin, M. Munera y C. Cifuentes, "The Actuation System of the Ankle Exoskeleton T-FLEX: First Use Experimental Validation in People with Stroke," *Multidisciplinary Digital Publishing Institute - brian sciences*, 2021.
- [9] S. Rezazadeh, D. Quintero, N. Divekar, E. Reznick, L. Gray and R. D. Gregg, "A Phase Variable Approach for Improved Rhythmic and Non-Rhythmic Control of a Powered Knee-Ankle Prosthesis," in *IEEE Access*, vol. 7, pp. 109840-109855, 2019.
- [10] I. Farkhatdinov, J. Ebert, Gijs van Oort, Mark Vlutters, Edwin van Asseldonk, et al. "Assisting Human Balance in Standing With a Robotic Exoskeleton," in *IEEE Robotics and Automation Letters*, IEEE 2019, 4 (2), pp.414-421.
- [11] D. Gomez-Vargas et al., "The Actuation System of the Ankle Exoskeleton T-FLEX: First Use Experimental Validation in People with Stroke," *Brain Sciences*, vol. 11, no. 4, p. 412, Mar. 2021.
- [12] D. Luviano Cruz, "Exoesqueleto Activo para Rehabilitación de Tobillo," *Academia Journals, Congreso Internacional De Investigación Academia Journals Celaya* 2021.
- [13] Lyu Mingxing, Chen Wei-Hai, Ding Xilun, Wang Jianhua, Pei Zhongcai, Zhang Baochang, "Development of an EMG-Controlled Knee Exoskeleton to Assist Home Rehabilitation in a Game Context," in *Frontiers in Neurorobotics*, vol. 13, 2019.
- [14] B. Freitas et al., "Design, Modelling and Control of an Active Weight-Bearing Knee Exoskeleton with a Series Elastic Actuator," 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG), 2019, pp. 1-4.
- [15] Chen, Ji et al. "A Pediatric Knee Exoskeleton With Real-Time Adaptive Control for Overground Walking in Ambulatory Individuals With Cerebral Palsy." *Frontiers in robotics and AI*, vol. 8, 2021.
- [16] Etenzi, E., Borzuola, R. & Grabowski, A.M. "Passive-elastic knee-ankle exoskeleton reduces the metabolic cost of walking," *J NeuroEngineering Rehabil* 17, n° 104, 2020.
- [17] Chen, Chunjie et al. "Iterative Learning Control for a Soft Exoskeleton with Hip and Knee Joint Assistance." *Sensors (Basel, Switzerland)* vol. 20,15 4333. 4 Aug. 2020.
- [18] T. Zhou, C. Xiong, J. Zhang, W. Chen, and X. Huang, "Regulating Metabolic Energy Among Joints During Human Walking Using a Multiarticular Unpowered Exoskeleton," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 662-672, 2021.
- [19] M. Grimmer, B. Quinlivan, S. Lee, P. Malcolm, D. Martineli, C. Siviyy, C. Walsh, "Comparison of the human-exosuit interaction using ankle moment and ankle positive power inspired walking assistance," *Journal of Biomechanics*, vol. 83, pp. 76-84, 2019.
- [20] S. Lee, D. Jin, S. Kang, D. Gaebler and L. Zhang, "Combined Ankle/Knee Stretching and Pivoting Stepping Training for Children With Cerebral Palsy," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 9, pp. 1743-1752, Sept. 2019.
- [21] Z. Wang, C. Yang, Z Ding, T. Yang, H. Guo, F. Jiang, B. Tian, "Study on the Control Method of Knee Joint Human-Exoskeleton Interactive System," *Sensors*, vol. 22, no. 3, p. 1040, Jan. 2022.
- [22] R. Hidayah, D. Sui, K. Wade, B. Chang, and S. Agrawal, "Passive knee exoskeletons in functional tasks: Biomechanical effects of a SpringExo coil-spring on squats," *Wearable Technologies*, vol. 2, p. e7, 2021.
- [23] H. Han et al., "Selection of Muscle-Activity-Based Cost Function in Human-in-the-Loop Optimization of Multi-Gait Ankle Exoskeleton Assistance," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 944-952, 2021.
- [24] I. Jammeli, A. Chemori, H. Moon, S. Elloumi and S. Mohammed, "An Assistive Explicit Model Predictive Control Framework for a Knee Rehabilitation Exoskeleton," in *IEEE/ASME Transactions on Mechatronics*, pp. 1-12, 2021.
- [25] Yacine Bougrinat, Sofiane Achiche, & Maxime Raison, "Design and development of a lightweight ankle exoskeleton for human walking augmentation," *Mechatronics*, vol. 64, 2019.
- [26] P. Franks, N. Bianco, G. Bryan, J. Hicks, S. Delp and S. Collins, "Testing Simulated Assistance Strategies on a Hip-Knee-Ankle Exoskeleton: a Case Study," 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob), pp. 700-707, 2020.
- [27] C. Sanz; M. Fantozzi; A. Parri; F. Giovacchini; A. Baldoni; M. Cempini; S. Crea; D. Lefeber; N. Vitiello, "A Knee-Ankle-Foot Orthosis to Assist the Sound Limb of Transfemoral Amputees," *IEEE TRANSACTIONS ON MEDICAL ROBOTICS AND BIONICS*, vol. 1, no. 1, pp. 38-48, 2019.
- [28] E.McCain, T. Dick, T. Giest, R. Nuckols, M. Lewek, K. Saul and G. Sawicki, "Mechanics and energetics of post-stroke walking aided by a powered ankle exoskeleton with speed-adaptive myoelectric control," *Journal of NeuroEngineering and Rehabilitation*, 2019.
- [29] L. Minchala, A. Velasco, J. Blandin, F. Astudillo, A. Vazquez, "Low-Cost Lower Limb Exoskeleton for Assisting Gait Rehabilitation: Design and Evaluation," *ACM Digital Library*, 2019.
- [30] J. Mu, H. Jiang, Y. Hua, J. Zhao, and Y. Zhu, "Design and Implementation of a Lightweight Lower Extremity Exoskeleton," *MATEC Web of Conferences*, 2019.
- [31] Z. Wang, C. Yang, F. Jiang, C. Yi, Z. Ding, B. Wei, J. Liu, "Design of a Lightweight and Flexible Knee Exoskeleton with Compensation Strategy," *ACM Digital Library*, 2021.
- [32] X. Liu, Z. Zhou, J. Mai, Q. Wang, "Real-time mode recognition based assistive torque control of bionic knee exoskeleton for sit-to-stand and stand-to-sit transitions," *Science Direct*, vol. 119, pp. 209-220, 2019
- [33] K. Seo; Y. Jin; J. Lee; S. Hyung; M. Lee, "RNN-Based On-Line Continuous Gait Phase Estimation from Shank-Mounted IMUs to Control Ankle Exoskeletons," *IEEE 16th International Conference on Rehabilitation Robotics*, pp. 24-28, 2019.
- [34] T. Lee, I. Kim, and S. Lee, "Estimation of the Continuous Walking Angle of Knee and Ankle (Talocrural Joint, Subtalar Joint) of a Lower-Limb Exoskeleton Robot Using a Neural Network," *Sensors*, 2021.
- [35] G. Bryan, P. Franks, S. Klein, R. Peuchen, S. Collins, "A hip-knee-ankle exoskeleton emulator for studying gait assistance," *The International Journal of Robotics Research*, vol. 40, pp. 722-746, 2020.
- [36] W. Zhao, A. Song, "Active Motion Control of a Knee Exoskeleton Driven by Antagonistic Pneumatic Muscle Actuators," *Multidisciplinary Digital Publishing Institute - Actuators*, 2020.
- [37] M. Rosenberg, B. Banjanin, S. Carga y K. Steele, "Predicting walking response to ankle exoskeletons using data-driven models," *The Royal Society*, 2020.
- [38] F. Mouzo, F. Michaud, U. Ligris, J. Cuadrado, "Leg-orthosis contact force estimation from gait analysis," *Mechanism and Machine Theory*, vol. 148, 2020.
- [1] J. R. Ortiz-Zacarias, Y. S. Valenzuela-Lino, J. Asto-Evangelista, and D. Huamanchahua, "Kinematic Position and Orientation Analysis of a 4 DoF Orthosis for Knee and Ankle Rehabilitation," 2021 6th International Conference on Robotics and Automation Engineering, ICRAE 2021, pp. 141-146, 2021, doi: 10.1109/ICRAE53653.2021.9657817.
- [2] J. Cornejo et al., "Mechatronic Exoskeleton Systems for Supporting the Biomechanics of Shoulder-Elbow-Wrist: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-9, doi: 10.1109/IEMTRONICS52119.2021.9422660.
- [3] D. Huamanchahua et al., "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [4] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo and J. Macuri-Vasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics

- & Mobile Communication Conference (UEMCON), 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [5] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [6] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.
- [7] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE), 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.

Hand Exoskeletons for Rehabilitation: A Systematic Review

Deyby Huamanchahua

*Department of Electrical and Mechatronic Engineering
Universidad de Ingeniería y Tecnología - UTEC
Lima, Peru
dhuamanchahua@utec.edu.pe*

Pedro Toledo-Garcia

*Department of Bioengineering
Universidad de Ingeniería y Tecnología - UTEC
Lima, Peru
pedro.toledo@utec.edu.pe*

Jack Aguirre

*Department of Mechatronic Engineering
Universidad Nacional de Trujillo - UNT
Trujillo, Peru
jaguirrev@unitru.edu.pe*

Sayda Huacre

*Professional School of Electrical Engineering
Universidad Nacional Mayor de San Marcos - UNMSM
Lima, Peru
sayda.huacre@unmsm.edu.pe*

Abstract—Medical conditions and accidents can cause immobility in certain parts of the body such as the upper extremities. To help people in the rehabilitation process, various robotic devices have been designed, such as exoskeletons; however, they have limited features that did not achieve optimal performance. This article presents a systematic review of the design of hand exoskeletons applied to rehabilitation. The goal is to provide the researcher with a structured matrix of features and components for the development of a much more efficient hand exoskeleton. Different databases and specialized search engines were used because they helped in the collection of information from the different investigations from the year 2019 to 2021. Filtering and selection of the investigations were carried out, so in the end, only 33 investigations were selected. Finally, it is concluded that there is the possibility of improvement in the development of hand exoskeletons that contribute to the rehabilitation process.

Index Terms—hand exoskeleton, degrees of freedom, review, rehabilitation

I. INTRODUCTION

Robotic rehabilitation devices have been perfected and are of great help to people with disabilities. These can replicate the manual work of the therapist, while improving motor recovery and functional independence for patients with physical or neurological disabilities. [38].

Currently, 15% of the world population has some type of motor disability in the lower or upper limbs due to cerebrovascular accidents, rheumatoid arthritis, carpal tunnel syndrome, or due to an injury [39]. When it comes to hand disabilities, exoskeletons help increase the effective range of motion of the joints for repeated activities of daily living. This is done through crucial rehabilitation movements in a planned and controlled way.

Furthermore, considering the complex anatomy and high mobility of the hand, exoskeletons are robotic devices capable of executing almost all the actions of a healthy hand, even being able to be configured according to the task to be performed. Also, they are comfortable and easy to handle by

the patient as they adapt to the size of the hand and the length of the fingers [40].

On the other hand, there are exoskeleton models such as HandMATE [4] and the one designed by Narvaez et. al. [5] that offer better finger mobility thanks to the 15 degrees of freedom they have. A key factor in exoskeleton design is to improve features such as simplicity, low cost, portability, and intuitive control [6]. This is the case of Butzer et. to the. who created a portable 3D prototype based on the Myo bracelet and electromyography (EMG) focused on child rehabilitation; Wang et. al. [37] designed and fabricated a hand exoskeleton controlled by voice commands via a mobile app and Li et. al. [6] created an efficient control system based on brain signals using Brainlink Lite. But, despite the advances that hand exoskeletons show day by day, there are still limited functions that do not achieve the real performance of the hand in its entirety.

Therefore, in this document, our objective is to present a systematic review of these devices and serve as a guide for future projects and research. The materials used in each study, the degrees of freedom, the control mechanisms and finally the state of development are detailed.

II. METHODOLOGY

For the development of this research, the "Emeral Insight", "Frontiers in Neuroscience", "IEEE Xplore", "MDPI", "SAGE Journals" and "Springer Link" databases were consulted. As well as articles from magazines such as "Politeknik Dergisi" and "ScienceDirect", and also university repositories such as the "Digital Repository of the Central University of Ecuador". Keywords such as "exoskeleton", "hand", "rehabilitation" and "robot" were important to identify the publications consulted in this article. From a total of 42 publications, that cover the years from 2019 to 2021, we had access to 33 of them. These articles were read, examined and interpreted within the present investigation.

III. REVIEW OF PUBLICATIONS

The information from the reviewed scientific articles on hand exoskeletons for rehabilitation was systematically arranged in a morphological table. This table was made up of the following fields: Reference, used sensors, number of degrees of

freedom (DoF), type of controller, actuators, type of treatment for which the exoskeleton was used, and the TRL or also called the maturity level of the technology.

The carried out review covers more information apart from that mentioned above, so information on these will be added later. Table I shows what was compiled in the 7 fields.

Table I: HAND EXOSKELETONS FOR REHABILITATION

References	Sensors	DoF	Controller	Actuator	Training	TRL
Ahmed, T. (2021) [8]	NE	14	Arduino UNO	14 M	Passive	6
Araujo, M. (2021) [9]	EEG electrodes	NE	Arduino UNO	DC M	Assisted	4
Castiblanco, J. (2021) [10]	EMG & Motion	4	Arduino Mega 2560	4 LS	Active	7
Esposito, D. (2021) [11]	FSR	11	Arduino UNO	S	Assisted	6
Li, G. (2021) [12]	Position and pair	3	NE	3 RA	Assisted	4
Li, M. (2021) [6]	Brainlink Lite	5	Arduino Mega 2560	5 LA	Passive	6
Meng, Q. (2021) [13]	FSR	NE	NE	1 MT	Assisted	6
Moreno, V. (2021) [14]	NE	5	Arduino Mega 2560	5 LA	Assisted	7
Serbest, K. (2021) [15]	NE	1	NE	1 LA	Passive	6
Shahid, T. (2021) [16]	EOG, current and flexion	NE	Arduino Mega 2560	2 S	Assisted	4
Sun, N. (2021) [17]	EMG and Force flexible sensors	3	NE	3 M with encoder	Assisted	4
Xiao, F. (2021) [18]	EMG, flexible and force	7	STM32F103ZET6 from computer	SGM	Assisted	6
Yang, L. (2021) [19]	Force and angle	3	FPGA chips	DC M	Active	6
Yang, S. (2021) [20]	Linear force sensing and Flex	5	Arduino Mega 2560	DC M	Assisted	4
Yumna, H. (2021) [21]	Sharp IR and touch sensors	NE	Arduino Mega 2560	DC M	Assisted	6
Birouas, F. (2020) [22]	Hall sensor	NE	NE	DC M with encoder	NE	6
Boser, Q. (2020) [23]	EMG and flexion	NE	NE	NE	Assisted	2
Erden, M. (2020) [24]	FSR, Pneumatic Pressure and Strain Gauge	3	Arduino Nano	S	Active	6
Haghshenas-Jaryani, M. (2020) [25]	Pressure, Vacuum and IMU sensors	6	Microcontroller and pneumatic controller	Soft and rigid, and tube-shaped actuators	Passive	7
Moya, R. (2020) [26]	Force and curvature sensors	3	NE	S	Assisted	4
Narvaez, V. (2020) [5]	NE	15	Atmega 328 and servo controller	MS and S	Passive and assisted	7
Sandison, M. (2020) [4]	FSR	15	Microcontroller and motor controller	LA with position feedback	Passive and assisted	6

Secciani, N. (2020) [27]	EMG	NE	Microcontroller	M and transmission cables	Assisted	2
Xu, D. (2020) [28]	FSR, Flexible Motion Angle and Strain	NE	NE	AC S	Assisted	6
Burns, M. (2019) [29]	EMG and flexion	10	Arduino Mega 2560	M, transmission cables and LA	Passive and active	7
Bützer, T. (2019) [30]	EMG	3	NE	NE	Active	7
Chowdhury, A. (2019) [31]	Force	3	NE	2 S	Passive and active	4
Jana, M. (2019) [32]	EMG and EEG	2	Arduino UNO	LA	Assisted	4
Jo, I. (2019) [33]	Linear force	4	NE	Linear Motor	Assisted	6
Li, M. (2019) [34]	Force	NE	Arduino UNO	Flexion and abduction actuators	NE	4
Marconi, D. (2019) [35]	Hall Encoder and MR Encoder Flexion, EMG, encoders and position	7	NI-sbRIO-9626 and FPGA	M and EA	NE	4
Rose, C. (2019) [36]	encoders and position	NE	Quanser Motor Controller and Acellus Panel	DC M	Passive	4
Wang, X. (2019) [37]	NE	5	Arduino UNO	MMGM	Assisted	6

Note: In training all of them are flexo-extension movements. Abbreviations: EMG: Electromyogram sensor, FPGA: Field-programmable gate array, FSR: Force-sensitive resistor, IMU: Inertial Measurement Units, NE: Non specified, M: Motor, DC M: DC Motor, LS: Linear servo, S: Servomotor, RA: Rotational actuator, LA: Linear Actuator, SGM: Steering gear motor, MT: Torque motor, MS: Micro servo, AC S: AC Servo motor, EA: Elastic Actuator, MMGM: Micro Metal Gearmotor, TRL 2: Formulated concept or technology, TRL 4: Laboratory test results, TRL 6: Results of the tests carried out at the prototype level in a relevant environment, TRL 7: Results of prototype-level tests carried out in an operational environment.

A. Sensors

In this section, a recount of the sensors used in the prototypes of the reviewed hand exoskeletons is made. Table II shows the proportions in percentages of the use of each sensor. It can be seen that the most used sensor is the force sensor, which represents 39.39 % and the least used are the curvature, current, vacuum sensors, etc., which each represent 3.03 % of the total. To calculate these percentages, we considered the 33 investigations as the total; however, the total number of sensors used is not always 33 since in some investigations it is mentioned that more than one sensor was used.

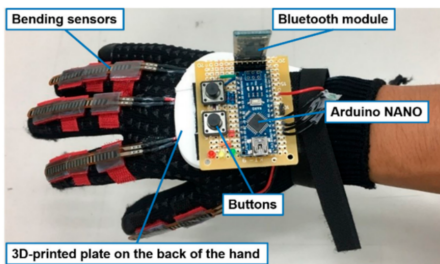


Fig. 1: Hand exoskeleton model from [20]

Table II: TYPES OF SENSORS

Sensors	Amount	
	Frequency	Percentage (%)
Force (includes FSR)	13	39.39
Curvature	1	3.03
EMG	9	27.27
Motion angle	2	6.06
Strain	2	6.06
Flexion	3	9.09
EEG	3	9.09
Encoders	2	6.06
Position	2	6.06
EOG	1	3.03
Current	1	3.03
Flexible	4	12.12
Hall	1	3.03
Touch	1	3.03
Pressure	2	6.06
Vacuum	1	3.03
IMU	1	3.03
NE	5	15.15

Note: Abbreviations: EEG: Electroencephalography electrodes, EMG: Elec-

tromyogram sensor/electrodes, EOG: Electrooculography electrodes, FSR: Force-sensitive resistor, IMU: Inertial Measurement Units, NE: Non specified.

B. Degrees of Freedom (DoF)

In this section, a count of the degrees of freedom (DoF) values of the hand exoskeleton prototypes of the reviewed articles is made. Table III shows the percentages corresponding to the proportion of the DoF. In this field, prototypes with 3 DoF were found more frequently, which represents 21.21 %; while prototypes based on 1, 2, 6, 10, 11, or 14 DoF are less frequent, accounting for 3.03 % individually.

Table III: DEGREES OF FREEDOM (DoF)

DoF	Amount	
	Frequency	Percentage (%)
1	1	3.03
2	1	3.03
3	7	21.21
4	2	6.06
5	4	12.12
6	1	3.03
7	2	6.06
10	1	3.03
11	1	3.03
14	1	3.03
15	2	6.06
NE	10	36.36

Note: Abbreviations: DoF: Degrees of Freedom, NE: Non specified.

C. Controller type

In this section, the proportion of types of controllers used in the prototypes of the hand exoskeletons is observed. Table IV shows the corresponding percentages for each type of controller. In this case, like the sensors, the total sum of percentages is not 100% because more than one controller was used in some projects. Regarding the most used type of controller, it is observed that this is the Arduino Mega 2560, which represents 21.21 % and the least used are the computer chip, the Arduino Nano, and the Atmega 328 controller, which individually represent 3.03 % taking into consideration the 33 items as the total.

Table IV: CONTROLLER TYPES

Controller type	Amount	
	Frequency	Percentage (%)
Arduino UNO	6	18.18
Arduino Mega 2560	7	21.21
Computer chip	1	3.03
FPGA chip	2	6.06
Arduino Nano	1	3.03
Microcontroller	3	9.09
Pneumatic controller	1	3.03
Atmega 328	1	3.03
Servo/motor controller	3	9.09
NE	11	33.33

Note: Abbreviations: FPGA: Field-programmable gate array, NE: Non specified.

D. Actuators

Table V shows the variety of actuators used in the prototypes of the hand exoskeletons. According to what was observed, these can be linear or rotational. In addition, it can be seen that the most used type of actuator is the motor (includes DC), which represents 39.39 % and the least used is, for example, the rotary actuator, which represents only 3.03 %.

Table V: ACTUATORS

Actuator type	Amount	
	Frequency	Percentage (%)
Motor (w/wo E)	7	21.21
DC motor (w/wo E)	6	18.18
Linear servo	1	3.03
Servo motor	6	18.18
Rotational actuator	1	3.03
Linear actuator	6	18.18
Steering gear motor	2	6.06
Soft and rigid actuator	1	3.03
Micro servo	1	3.03
Transmission cable	2	6.06
FA actuator	1	3.03
Elastic actuator	1	3.03
NE	2	6.06

Note: Abbreviations: E: Encoder, FA: Flexion and abduction, NE: Non specified, w/wo: With or without.

E. Training modes and movements

Information about the training modes, which can be passive, active, or assisted, and the types of rehabilitation movement is placed in this section. Regarding the movements, it was found that the most used and which is evidenced in all the articles is the movement of flexion and extension of the joints. Table VI shows the proportion through the percentages of the training modes corresponding to the flexo-extension movement (FE). In the same way, as in Table VI, the total sum of percentages exceeds 100 %, due to the use of more than one training mode in the consulted investigations. According to the table, it can be seen that the training mode based on assisted movements is the most frequent, which represents 57.58 % of the analyzed articles, while active movements are the least frequent, representing only 15.15 % . Likewise, it is important to mention that within Table VI the category "Mixed Movements" refers to training modes that combine the basic ones listed above; i.e. "Active and Passive Movements", "Passive Assisted Movements", etc.

Table VI: TRAINING MODES OF THE FE MOVEMENT

Training mode	Amount	
	Frequency	Percentage (%)
Active Movements	5	15.15
Assisted Movements	19	57.58
Passive Movements	8	24.24
Mixed Movements	4	12.12
NE	3	9.09

Note: Abbreviations: FE: Flexo-extension, NE: Non specified.

F. Technology Readiness Levels (TRL)

This section addresses the maturity levels of the hand exoskeleton prototypes of the reviewed articles. The TRL (Technology Readiness Levels) are numbers that go from 1 to 9, and these are assigned according to the level of development of the technology and its applicability to an environment. Table VII shows the proportion of TRLs using the corresponding percentages. According to the table, it can be seen that the most frequent TRL is 6 (results of the tests carried out at the prototype level in a relevant environment), which represents 42.42 % and the least frequent is 2 (concept or technology formulated), which represents 6.07 %.

Table VII: TECHNOLOGY READINESS LEVELS

TRL	Amount	
	Frequency	Percentage (%)
2	2	6.07
4	11	33.33
6	14	42.42
7	6	18.18

Note: Abbreviations: TRL: Technology Readiness Levels.

G. Control type

This section adds more information on the types of control used in the revised hand exoskeletons. Mostly, the type of control considered is the PID and in a large part of the articles, the type of control used is not specified. Table VIII shows the proportion of the types of control used in the projects. According to the table, it is evident that the most frequent type of control is PID, which represents 33.33 % and the least frequent is, for example, fuzzy control, which represents 3.03 %.

Table VIII: CONTROL TYPE

Control type	Amount	
	Frequency	Percentage (%)
ON - OFF	5	15.15
Fuzzy	1	3.03
PID	11	33.33
With potentiometer	1	3.03
Simple feedback	1	3.03
Proportional control	1	3.03
Closed loop control	1	3.03
Neural networks	1	3.03
NE	11	33.33

Note: Abbreviations: NE: Non specified.

H. Material

Finally, this section deals with more information about the used materials for the construction of the prototypes of the hand exoskeletons for rehabilitation. According to the reviewed articles, the prototypes are made through the use of different materials, highlighting among them the use of PLA through 3D printing. Table IX shows the proportion of the types of used materials. In this case, the total percentage exceeds 100 % due to the fact that some investigations used more than

one material in the construction process. Thus, according to Table IX, PLA represents 36.36 % of the total, while different materials such as VisiJet Glass, RTV Silicone Rubber, among others that were used only once, individually represent 3.03 % of the total.

Table IX: MATERIALS

Material	Amount	
	Frequency	Percentage (%)
ABS	3	9.09
Aluminium	2	6.06
Cristal visiJet	1	3.03
RTV Silicone Rubber	1	3.03
Markforged Onyx	1	3.03
PRR	1	3.03
PLA	12	36.36
Stainless Steel	1	3.03
TPU	2	6.06
NE	12	36.36

Note: Abbreviations: ABS: Acrylonitrile Butadiene Styrene, PLA: Polylactic acid, PRR: Photopolymer Rapid Resin, TPU: Thermoplastic Polyurethane, NE: Non specified.

IV. CONCLUSION

Thanks to technological progress and the improvement of current clinical studies, the development of hand exoskeletons has been evolving; therefore, future designers must adopt increasingly efficient hand exoskeletons designs. Such designs can only be possible after studying the design options of current devices.

In this article, we investigate a wide range of existing hand exoskeletons in the literature based on different design aspects, such as actuator technologies or control strategies.

However, the search for the most efficient device is not over yet. We hope that this review provides useful guidelines and practices for future designers as they create new and efficient hand exoskeletons.

For future works, based on the information obtained from this research about the materials and methods used in the consulted references, the design of a hand exoskeleton will be proposed as a new alternative to contribute to the rehabilitation of arthritis in the elderly.

REFERENCES

- [1] M. Sarac, M. Solazzi and A. Frisoli, "Design Requirements of Generic Hand Exoskeletons and Survey of Hand Exoskeletons for Rehabilitation, Assistive, or Haptic Use," in IEEE Transactions on Haptics, vol. 12, no. 4, pp. 400-413, 2019, doi: 10.1109/TOH.2019.2924881.
- [2] World Health Organization."Disability and health". 2021. <https://www.who.int/news-room/fact-sheets/detail/disability-and-health>.
- [3] A. Barrientos and J. del Cerro, "Robotics in medicine", Medicina Clínica (English Edition), vol. 152, no. 12, pp. 493-494, 2019, doi: 10.1016/j.medcle.2019.02.023.
- [4] M. Sandison et al., "HandMATE: Wearable Robotic Hand Exoskeleton and Integrated Android App for at Home Stroke Rehabilitation", Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2020-July, pp. 4867-4872, 2020, doi: 10.1109/EMBC44109.2020.9175332.

- [5] V. Narvaez, B. Bolaños, D. J. López, J. A. Guerrero, J. E. Mejía and S. E. Ruiz, "Diseño de un prototipo de exoesqueleto para rehabilitación postquirúrgica del síndrome del túnel del carpo", 2020 IX International Congress of Mechatronics Engineering and Automation (CIIMA), pp. 1–6, 2020.
- [6] M. Li et al., "An attention-controlled hand exoskeleton for the rehabilitation of finger extension and flexion using a rigid-soft combined mechanism", *Front. Neurobot.*, vol. 13, no May, pp. 1–13, 2019, doi: 10.3389/fnbot.2019.00034.
- [7] X. Wang, P. Tran, S. M. Callahan, S. L. Wolf and J. P. Desai, "Towards the development of a voice-controlled exoskeleton system for restoring hand function", 2019 Int. Symp. Med. Robot. ISMR 2019, vol. 1, pp. 1–7, 2019, doi: 10.1109/ISMR.2019.8710195.
- [8] T. Ahmed et al., "Flexohand: A hybrid exoskeleton-based novel hand rehabilitation device", *Micromachines*, vol. 12, no 11, 2021, doi: 10.3390/mi12111274.
- [9] R. S. Araujo, C. R. Silva, S. P. N. Netto, E. Morya and F. L. Brasil, "Development of a Low-Cost EEG-Controlled Hand Exoskeleton 3D Printed on Textiles", *Front. Neurosci.*, vol. 15, no June, pp. 1–13, 2021, doi: 10.3389/fnins.2021.661569.
- [10] J. C. Castiblanco, I. F. Mondragon, C. Alvarado-Rojas and J. D. Colorado, "Assist-as-needed exoskeleton for hand joint rehabilitation based on muscle effort detection", *Sensors*, vol. 21, no 13, pp. 1–16, 2021, doi: 10.3390/s21134372.
- [11] D. Esposito et al., "Design of a 3D-Printed Hand Exoskeleton Based on Force-Myography Control for Assistance and Rehabilitation", *Machines*, vol. 10, no 1, pp. 1–16, 2022, doi: 10.3390/machines10010057.
- [12] G. Li, L. Cheng and N. Sun, "Design, manipulability analysis and optimization of an index finger exoskeleton for stroke rehabilitation", *Mech. Mach. Theory*, vol. 167, no September 2021, p. 104526, 2022, doi: 10.1016/j.mechmachtheory.2021.104526.
- [13] Q. Meng, Z. Shen, Z. Nie, Q. Meng, Z. Wu and H. Yu, "Modeling and evaluation of a novel hybrid-driven compliant hand exoskeleton based on human-machine coupling model", *Appl. Sci.*, vol. 11, no 22, 2021, doi: 10.3390/app112210825.
- [14] V. Moreno-SanJuan, A. Císnal, J. C. Fraile, J. Pérez-Turiel and E. de-la-Fuente, "Design and characterization of a lightweight underactuated RACA hand exoskeleton for neurorehabilitation", *Rob. Auton. Syst.*, vol. 143, p. 103828, 2021, doi: 10.1016/j.robot.2021.103828.
- [15] K. Serbest and O. Eldogan, "Design, Development and Evaluation of a New Hand Exoskeleton for Stroke Rehabilitation at Home", *J. Polytech.*, vol. 0900, no 1, pp. 305–314, 2020, doi: 10.2339/politeknik.725310.
- [16] T. Shahid, D. Gouwanda, S. G. Nurzaman, A. A. Gopalai and T. K. Kheng, "Development of an Electrooculogram-activated Wearable Soft Hand Exoskeleton," 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 433-438, doi: 10.1109/IECBES48179.2021.9398797.
- [17] N. Sun, G. Li, y L. Cheng, "Design and Validation of a Self-Aligning Index Finger Exoskeleton for Post-Stroke Rehabilitation", *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1513–1523, 2021, doi: 10.1109/TNSRE.2021.3097888.
- [18] F. Xiao, L. Gu, W. Ma, Y. Zhu, Z. Zhang and Y. Wang, "Real time motion intention recognition method with limited number of surface electromyography sensors for a 7-DOF hand/wrist rehabilitation exoskeleton", *Mechatronics*, vol. 79, no November 2020, p. 102642, 2021, doi: 10.1016/j.mechatronics.2021.102642.
- [19] L. Yang, F. Zhang, J. Zhu and Y. Fu, "A Portable Device for Hand Rehabilitation with Force Cognition: Design, Interaction and Experiment", *IEEE Trans. Cogn. Dev. Syst.*, vol. 8920, no c, 2021, doi: 10.1109/TCDS.2021.3055626.
- [20] S. H. Yang et al., "An instrumented glove-controlled portable hand-exoskeleton for bilateral hand rehabilitation", *Biosensors*, vol. 11, no 12, pp. 1–12, 2021, doi: 10.3390/bios11120495.
- [21] H. Yumna, A. Arifin and A. F. Bagei, "Robotic Hand Exoskeleton with Tactile Force Feedback for Post-Stroke Spasticity Rehabilitation", *Proc. - 2021 Int. Semin. Intell. Technol. Its Appl. Intell. Syst. New Norm. Era, ISITIA 2021*, pp. 266–271, 2021, doi: 10.1109/ISITIA52817.2021.9502261.
- [22] F. I. Birouaş, R. C. Țarcă, S. Dzitac and I. Dzitac, "Preliminary results in testing of a novel asymmetric underactuated robotic hand exoskeleton for motor impairment rehabilitation", *Symmetry (Basel)*, vol. 12, no 9, 2020, doi: 10.3390/sym12091470.
- [23] Q. A. Boser, M. R. Dawson, J. S. Schofield, G. Y. Dziwenko and J. S. Hebert, "Defining the design requirements for an assistive powered hand exoskeleton: A pilot explorative interview study and case series", *Prosthet. Orthot. Int.*, 2020, doi: 10.1177/0309364620963943.
- [24] M. S. Erden, W. Mccoll, D. Abassebay and S. Haldane, "Hand Exoskeleton to Assess Hand Spasticity", 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob), pp. 1004–1009, 2020.
- [25] M. Haghshenas-Jaryani, R. M. Patterson, N. Bugnariu and M. B. J. Wijesundara, "A pilot study on the design and validation of a hybrid exoskeleton robotic device for hand rehabilitation", *J. Hand Ther.*, vol. 33, no 2, pp. 198–208, 2020, doi: 10.1016/j.jht.2020.03.024.
- [26] R. Moya-Jiménez, T. Magal-Royo, D. Ponce, M. Flores and M. Caiza, "Hand Exoskeleton Design for the Rehabilitation of Patients with Rheumatoid Arthritis", *Springer Nature Switzerland*, pp. 1307, 12–21, 2020.
- [27] N. Secciani, M. Pagliai, F. Bounamici, F. Vannetti, Y. Volpe and A. Ridolfi, "A Novel Architecture for a Fully Wearable Assistive Hand Exoskeleton System", *IFTToMM Italy 2020. Mech. Mach. Sci.*, vol. 91, 2020, doi: https://doi.org/10.1007/978-3-030-55807-9_14.
- [28] D. Xu, Q. Wu and Y. Zhu, "Development of a soft cable-driven hand exoskeleton for assisted rehabilitation training", *Ind. Rob.*, vol. 48, no 2, pp. 189–198, 2020, doi: 10.1108/IR-06-2020-0127.
- [29] M. K. Burns, D. Pei and R. Vinjamuri, "Myoelectric control of a soft hand exoskeleton using kinematic synergies", *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no 6, pp. 1351–1361, 2019, doi: 10.1109/TB-CAS.2019.2950145.
- [30] T. Butzer et al., "PEXO - A pediatric whole hand exoskeleton for grasping assistance in task-oriented training", *IEEE Int. Conf. Rehabil. Robot.*, vol. 2019-June, pp. 108–114, 2019, doi: 10.1109/ICORR.2019.8779489.
- [31] A. Chowdhury, S. S. Nishad, Y. K. Meena, A. Dutta and G. Prasad, "Hand-Exoskeleton Assisted Progressive Neurorehabilitation Using Impedance Adaptation Based Challenge Level Adjustment Method", *IEEE Trans. Haptics*, vol. 12, no 2, pp. 128–140, 2019, doi: 10.1109/TOH.2018.2878232.
- [32] M. Jana, B. G. Barua and S. M. Hazarika, "Design and Development of a Finger Exoskeleton for Motor Rehabilitation using Electromyography Signals", 2019 23rd Int. Conf. Mechatronics Technol. ICMT 2019, pp. 1–6, 2019, doi: 10.1109/ICMECT.2019.8932126.
- [33] I. Jo, Y. Park, J. Lee and J. Bae, "A portable and spring-guided hand exoskeleton for exercising flexion/extension of the fingers", *Mech. Mach. Theory*, vol. 135, pp. 176–191, 2019, doi: 10.1016/j.mechmachtheory.2019.02.004.
- [34] M. Li et al., "A 3D-printed soft hand exoskeleton with finger abduction assistance", 2019 16th Int. Conf. Ubiquitous Robot. UR 2019, pp. 319–322, 2019, doi: 10.1109/URAI.2019.8768611.
- [35] D. Marconi, A. Baldoni, Z. McKinney, M. Cempini, S. Crea and N. Vitiello, "A novel hand exoskeleton with series elastic actuation for modulated torque transfer", *Mechatronics*, vol. 61, no November 2018, pp. 69–82, 2019, doi: 10.1016/j.mechatronics.2019.06.001.
- [36] C. G. Rose y M. K. O'Malley, "Hybrid Rigid-Soft Hand Exoskeleton to Assist Functional Dexterity", *IEEE Robot. Autom. Lett.*, vol. 4, no 1, pp. 73–80, 2019, doi: 10.1109/LRA.2018.2878931.
- [37] D. Huamanchahua et al., "A Robotic Prosthesis as a Functional Upper-Limb Aid: An Innovative Review," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422648.
- [38] D. Huamanchahua, D. Yalli-Villa, A. Bello-Merlo and J. Macurivasquez, "Ground Robots for Inspection and Monitoring: A State-of-the-Art Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), 2021, pp. 0768-0774, doi: 10.1109/UEMCON53757.2021.9666648.
- [39] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [40] D. Huamanchahua, A. Tadeo-Gabriel, R. Chávez-Raraz and K. Serrano-Guzmán, "Parallel Robots in Rehabilitation and Assistance: A Systematic Review," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), 2021, pp. 0692-0698, doi: 10.1109/UEMCON53757.2021.9666501.

Artificial Intelligence Applied in Human Medicine with the Implementation of Prostheses

Ismael Alvarado-Landeo
Department of Mechatronics
Engineering
Universidad Continental
Huancayo, Perú
74174245@continental.edu.pe

Erick Surichaqui-Montalvo
Department of Mechatronics
Engineering
Universidad Continental
Huancayo, Perú
74479295@continental.edu.pe

Kener Velasquez-Colorado
Department of Mechatronics
Engineering
Universidad Continental
Huancayo, Perú
76636453@continental.edu.pe

Deyby Huamanchahua
Department of Mechatronics
Engineering
Universidad Continental
Huancayo, Peru
dhuamanchahua@continental.edu.pe

Abstract— *The use of artificial intelligence (AI) in medicine is already a reality. Everywhere there is talk of the advantages that AI can mean for the future in our daily lives, as well as its possible applications. The future of "standard" medical practice could appear here ahead of schedule, where a patient could go to a computer before seeing a doctor. Through advances in AI, it becomes more possible for the days of misdiagnosis and treatment of the symptoms of the disease, rather than its root cause, to be left behind. Think about how many years of blood pressure measurements you have or how much storage you would need to remove so that you can fit a complete 3D image of an organ on your laptop. The idea of artificial intelligence in medicine may make you think of robots roaming the halls of a hospital in the distant future, but AI is already here.*

Keywords— *Artificial intelligence, prosthesis implantation, prosthetic rehabilitation, nanomaterials, modeling.*

I. INTRODUCTION

Artificial intelligence (AI) is a branch of computer science that includes very transversal concepts related to logic and learning [1]. The process of implanting the human prosthesis is simplified and can be performed in a less complex way aesthetically, being possible to restore a member of the body without so much complexity in its assembly and operation. For this article, he has focused on seeing how artificial intelligence takes a big step by establishing this type of implementation and that it makes it more efficient in the process of it. Health professionals must know this technology, its advantages, and its disadvantages because it will be an integral part of their work [2].

It is key to highlight that there are different types of human prosthesis implantations, which in many cases can be excessively expensive due to the necessary inputs to perform the implantation adding the time it takes the specialist to perform this process that tries to respond to our problem.

Concerning AI, his contribution was in the integration of nanomaterials and biomaterials to be able to reconstruct part of the affected skin. As you already know, this has given a very important takeoff in the development of prostheses, although it has not exploited all the resources it could provide and not only to prostheses but to medicine in general. This leads to a review of articles related to health topics, applications of AI, prostheses, and new technologies in implants. For example, maxillofacial prosthetic rehabilitation replaces missing structures to regain function and aesthetics related to facial defects or injuries [3].

However, it is known that the nerve cuff electrodes have remained stable during the four months since implantation. These results suggest that 16-channel neuroprostheses will provide stronger knee extension moments for longer before fatigue during standing and transfers [4]. That is why AI can be used to address many challenges facing the world's healthcare system, from disease detection to building predictive models for treatment, thereby improving quality, and reducing the cost of patient care. For example, in recent decades, medicine, mechatronics, mathematics, and materials science have progressed together in the search for the ideal active prosthesis for the upper limb [5].

II. METHODOLOGY

To indicate the intensity of the articles reviewed on the artificial intelligence applied in medicine, a systematic review of the simple literature was carried out. It inquired about the topics of artificial intelligence in the implementation of prostheses in different databases such as Science Direct, Scopus Preview, and Springer, each in the "article" segment where the search for articles related to Artificial Intelligence and Medicine was carried out.

The following syntax is used for the search: Prosthesis implantation, bone, direct skeletal union, limb amputation, neurocentral, osseointegration, prosthetic rehabilitation, Nanomaterials for orthopedic implants and applications, Rehabilitation and prosthetics post-amputation, Integration of robotics and neuroscience, Neural Interface Implanted. It is appreciable that "AI" and "prostheses" are the keywords that were entered for search engines on the respective information bank pages.

In addition, it is key to note that these databases were used for the reliability of their studies. The search for information was initially carried out with 130 articles, where due to the Mendeley program, it has been possible to summarize and catalog the respective data and cite them precisely, providing the 83 articles investigated properly to the central study topic. Then, all the information was organized and systematized in tables, in an organized way, indicating the country, material (characteristics), and citation of the different articles collected.

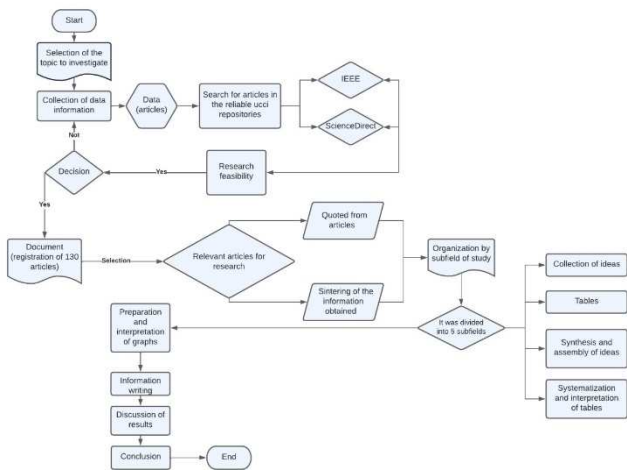


Fig. 1. Flowchart on the elaboration of the article.

III. RESULT AND ANALYSIS

The results that were obtained reflect that the largest number of articles from AI go from the year 2016-2021 collected from repositories such as IEEE and Science Direct having, in the same way, a considerable sum of people involved in recent years to the preparation of these articles and who are in constant exploration on the advancement of AI

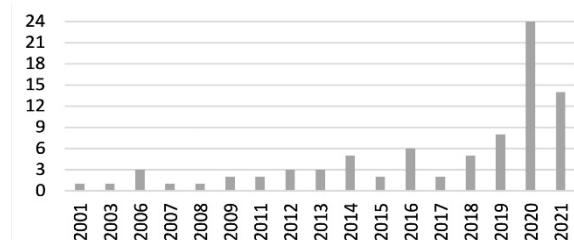


Fig. 2. Graph of bars of scientific articles by years of publication.

Fig. 2 shows that the publications about Artificial Intelligence applied in Human Medicine with the Implementation of Prostheses, the years from 2001 to 2013, do not exceed 3 articles. However, as of 2014, it shows an increase to a maximum of 6 articles. This trend continues until 2020 when it manifests an amount greater than 20 articles.

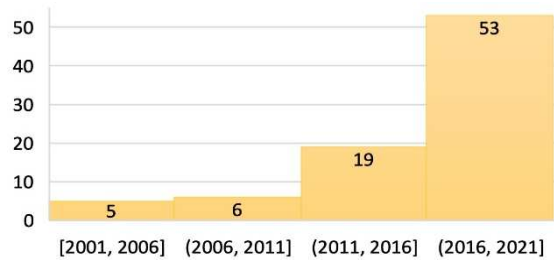


Fig. 3. Histogram of the number of articles collected.

Fig. 3 shows the number of publications we collect for our topic Artificial Intelligence applied in Human Medicine with the Implementation of Prostheses. The compilation is 83 publications in total, according to the year, we divide them into intervals of 5 years starting in 2001. As you can see, most are from recent years, as they range from 2016 to 2021.

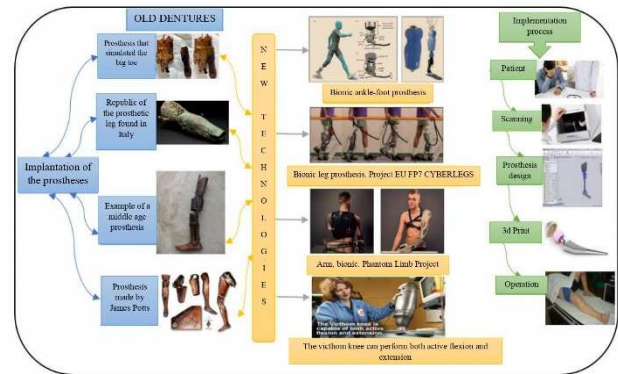


Fig. 4. Process diagram of the evolution of the prosthesis and the implantation process [6][7].

Fig. 4 shows, the evolution of the prosthesis that had over the years which was both the change the evolution of the prosthesis since these date from the twentieth century, until today where modern prostheses are developed and produced thanks to AI and technological evolution within the medicine of computer and robotic advances, it also shows the process involved in the implementation of a prosthesis in humans.

IV. FUNDAMENTAL PRINCIPLES OF AI IN MEDICINE

Table I. Systematization Table Fundamental principles within medicine

Comparative Studies	AI Factors	Medical Factors	Response variable
[8]	Data collection	Prosthesis design	Novel bionic foot when walking.
[5]	Data collection	Prosthesis design	Ideal prosthesis for the upper limb

[1]	Data collection	Medical Decision Making	Personalized medicine
[2]	Data collection	Medical Decision Making	Technology adapted to doctors
[9]	Detection	Treatment prediction	Improving hospital care
[10]	Feedback	Sensory information	Neural prosthesis technologies

According to Table I, 21.4% of the authors focus on the use of AI for the design of prostheses, and the same percentage of authors focus on the use of AI for medical decision making, on the other hand, 14.3% of articles study the application of AI for the prediction of treatments, The same number of articles focus on healthcare, 28.6% of the authors studied the application of AI in four different topics in which sensory information, neurocomputing, systems description and solving previous medical problems are found. It is concluded that in this table studies on the application of AI in favor of prosthesis design and medical decision making predominate.

Therefore, technology and medicine follow a parallel path during the last decades. On the other hand, AI is made up of a series of logical algorithms [11]. That is why, in this sub-theme, we will review the significant advances in sensory neural prosthesis technologies made in recent decades [12]. For example, CIBERESUCICOVID aims to determine through artificial intelligence analysis what are the risk and prognostic factors of patients infected with COVID [13]. In addition, practical intelligent applications (PRINTEPS) are being developed, which is a user-centric platform for developing integrated intelligent applications only by combining four types of modules [14]. For example, the prosthetic knee is one of the key elements in the design of modular prostheses and the appearance of new mechanisms together with the proliferation of knees has complicated its classification and prescription.

Telepresence surgery, on the other hand, is an interactive computerized system, so fast and intuitive that the computer disappears from the surgeon's mind, transforming the environment generated by the system into something real [15]. For example, artificial neural networks (ANNs) are well established in BCI research and have numerous successful applications [16]. On the other hand, thanks to AI, transtibial amputees can better approximate typical movement patterns at slow, normal walking speeds using the new bionic prosthesis [17]. In the past, natural heart valves functioned primarily as live check valves, which are purely fluid dynamics devices in nature [18].

That is why AI can be used to address many challenges facing the world's healthcare system, thereby improving the quality, and reducing the cost of patient care [9]. For example, in recent decades, medicine, mechatronics, mathematics, and materials science have progressed together

in the search for the ideal active prosthesis for the upper limb [19]. However, there is a reflection on the evolution of the field of Neurocomputing that has witnessed the sequence of editions of the International Working Conference on Artificial Neural Networks [20]. Similarly, this review demonstrates that the application of advanced ai methods in healthcare has the potential to improve the quality of care by uncovering non-obvious and clinically relevant relationships and enabling timely care intervention [21].

V. ARTIFICIAL INTELLIGENCE FOR MEDICAL DIAGNOSIS

Table II. Table of Systematization of artificial intelligence for medical diagnosis

Comparative Studies	AI Factors	Medical Factors	Response variable
[22]	Data collection	Prosthesis design	Transfemoral Prosthesis Model
[23]	Data collection	Prosthesis design	A prosthesis system is an artificial arm
[24]	Data collection	Prosthesis design	Implementing AI in Radiology
[25]	Data collection	Prosthesis design	Portable devices for upper limb amputees
[26]	Estimation algorithm	Prosthesis implantation	The precision of implantation of the prosthesis
[27]	Data collection	Implant durability	Neo-intimate incorporation of the Stentrod
[26]	Learning algorithm	Control of electromyogram patterns	Smart prostheses controlled by EMG-PR
[28]	Biomechanics	Postoperative monitoring	Implants with load detection

According to Table II, 23.5% of the authors focus on the use of AI for the design of prostheses, and the same percentage of authors focus on the use of AI for medical decision making, on the other hand, 17.6% of articles study the application of AI for the prediction of treatments, 35.3% of the authors studied the application of AI in six different topics in which medical care, implantation of prostheses, the durability of implants, control of myogram patterns, integration of the nervous system and postoperative monitoring are found.

In such a way in the era of Industry 4.0, sustainable chemistry and processes are undergoing a drastic transformation of the continuous flow system towards the next level of operations [29]. Likewise, technological capabilities are creating an opportunity for machine learning and artificial intelligence (AI) to enable "intelligent" brain-machine interfaces (BMI) designed by nanoengineering [30]. Thus, complementing the study of the active knee joint and the active knee-ankle transfemoral prosthesis [21]. At the same time, the prosthesis system is an artificial arm produced by INAIL en Vigorso [22]. As well as, on the report of electrodes mounted on a stent (Stentrod) capable of

chronically recording neural signals from inside a blood vessel with high fidelity [26]. AI is known to be used to identify new drug therapies and improve a physician's efficiency [31].

However, advances in AI are contributing to a growing reliance on algorithms to make decisions important to humans [32]. Another important fact, breast cancer is the leading cause of cancer death. Survival rates in developing countries range from 50% to 60% due to late detection [33]. However, the use of an intelligent learning algorithm for electromyogram pattern recognition in upper limb prostheses is considered an important clinical option [25]. In contrast, orthopedics has not yet adapted to innovative trends in health control, from an obvious entry point during orthopedic surgeries, physicians remain unable to objectively examine the structural integrity and biomechanics in the region operated through implantable sensors [27].

The use of AI models applied to diagnostics offers multiple benefits such as its large data storage capacity [34]. Thus, we propose an auxiliary decision support system that combines joint learning with case-based reasoning to help clinicians improve the accuracy of predicting breast cancer recurrence [10]. That's why AI has the potential to improve all levels of radiology workflow and practice [23]. Likewise, portable devices (WD) have evolved from purely mechanical devices to intelligent mechatronic systems thanks to the continuous advancement of technology integrating sensors, actuators, novel materials, and above all sensory feedback [24].

VI. ADAPTATION OF THE HUMAN PROSTHESIS

Table III. Table of Systematization of Adaptation of the human prosthesis

Comparative Studies	AI Factors	Medical Factors	Response variable
[35]	Modeling	Prosthesis implantation	Electronic skins
[36]	Information Collection	Prosthesis implantation	Skeletal fixation technology and neuromuscular control technology.
[37]	Modeling	Prosthesis design	Sophisticated portable gripper
[38]	Modeling	Prosthesis design	Spring-shock absorber
[39]	Information Collection	Prosthesis design	Physical activity tracking devices
[40]	Modeling	Prosthesis design	Bioelectric prosthetic hand
[4]	Modeling	Integration of the nervous system	Implanted neuroprosthesis

According to Table III, 28.5% of the authors focus on the use of AI for medical decision making, the same percentage of authors focus on the use of AI for the design of prostheses, on the other hand, 21.4% of authors focus on the use of AI for the implantation of prostheses, the same percentage of authors focus on the use of AI for nervous

system integration. It can be said that in this table the studies on the different approaches are homogeneous.

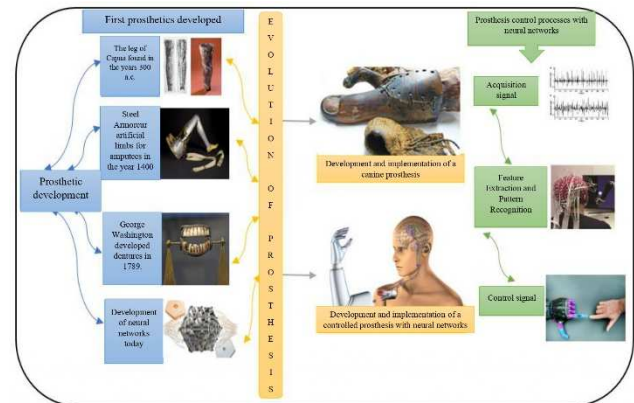


Fig. 5. Evolution of the prosthesis [41][42].

To some extent, advanced AI technology also creates new opportunities to explore the scientific basis of traditional Chinese medicine (TCM) and develop the standardization and digitization of TCM pulse diagnostic methodology. That is why we review and discuss the possible application of AI technology in the diagnosis of TCM pulses [43]. For example, the ability to learn accurate, temporarily abstracted predictions is shown through two case studies: the myoelectric control of a non-disabled robotic arm and the interactions of an amputee with a myoelectric training robot [44]. On the other hand, this system is divided into four separate categories: force detection, temperature sensation, motion detection, and final feedback allows the user to monitor and control the hand through an Android app [39].

After all, robotic technology can help build bipedal robots that allow human gait to be emulated, aimed at building lower prostheses or exoskeletons as can be seen in Fig. 6, which help to walk and perform other muscular activities that would otherwise be impossible [35]. That's why wearable sensors have evolved from fitness trackers that are worn on the body to multifunctional, highly integrated, compact, and versatile sensors [38].

VII. SIMULATION FOR THE DEVELOPMENT OF PROSTHESES

Table IV. Simulation Systematization Table for the development of prostheses.

Comparative Studies	AI Factors	Medical Factors	Response variable
[45]	Modeling	Prosthesis implantation	Control technology
[46]	Information Collection	Prosthesis implantation	Amputee Rehabilitation
[47]	Information Collection	Prosthesis implantation	Joint stiffness
[48]	Modeling	Prosthesis implantation	Application of nanomaterials
[49]	Modeling	Prosthesis design	Osseointegrated prosthesis
[50]	Information Collection	Integration of the nervous system	Artificial neural networks

According to Table IV, 31.5% of the authors focus on the use of AI for medical decision-making, on the other hand, 26.3% of the authors focus on the use of AI for the design of prostheses, as do the authors who focus on the use of AI for the implantation of prostheses. 15.7% of the authors focus on the use of AI for nervous system integration. That's why Gen Z uses their smartphones daily more than any other generation. This aspect of your life should be considered when talking about education, since all your free time is dedicated to these smart devices, educational apps could implement some aspect of learning when using these devices [51]. This leads to a new wave of technologies making their way into clinical practice, including mHealth, which enables constant monitoring of biological parameters, anytime, anywhere, of hundreds of patients at the same time [37]. Thus, AI is viable for patients to rehabilitate, and improve the quality of surgical equipment so that it provides the final prosthesis and meets the objectives set [45].

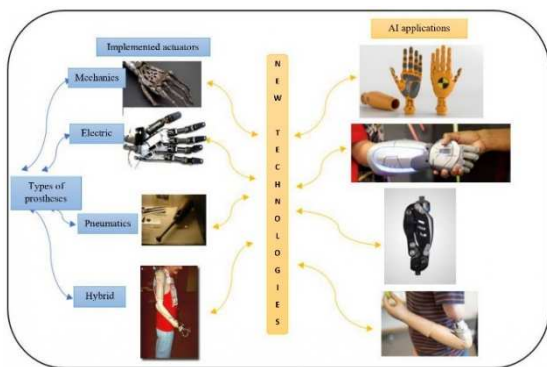


Fig. 6. Types of prosthetic limbs [52].

There have been applications of artificial intelligence (AI) in all aspects of CT imaging (acquisition, reconstruction, analysis, and measurement of new imaging features) that can further enhance value and reduce cost [53]. Likewise, the implementation of hybrid neuroprosthetic systems (HBS) relates artificial intelligence to the human nervous system, so being very important, many researchers investigate the subject more to recover sensorimotor functions in people suffering from different disabilities [54]. In addition, as shown in Fig. 6, an in-line control system for upper limb prostheses based on EEG motor imaging is performed, consisting of the brain-computer interface unit (BCI), motion controller, and target orientation module [44].

Nanomaterials hold promise for orthopedic applications due to their excellent tribological properties, wear resistance, osseointegration, and tissue regeneration capacity. Nanostructured materials play a pivotal role in orthopedic research due to their size range and ability to mimic the characteristics and hierarchical structure of native bones [47]. On the other hand, the WiseTOP real-time communication platform incorporated in an osseointegrated prosthesis is used for the recovery of hand function after amputation and presents the initial results of the performance evaluation [48].

It also features an integrated, real-time framework for finger strength control in upper extremity prostheses [55].

VIII. NEW TECHNOLOGIES FOR THE MEDICAL INDUSTRY IN THE APPLICATION OF PROSTHESES

Table V. Table of Systematization of New Technologies for the Medical Industry in the Application of Prostheses

Comparative Studies	AI Factors	Medical Factors	Response variable
[56]	Detection	Treatment prediction	Diagnosis and therapy and the relationship with AI
[57]	Modeling	Prosthesis design	Smart textiles
[36]	Data collection	Prosthesis design	Upper limb prosthesis
[58]	Modeling	Prosthesis implantation	Restore, retain, or modify damaged tissues
[59]	Information Collection	Integration of the nervous system	The artificial neural feedback network

According to Table 5, 33.3% of the authors focus on the use of AI for healthcare, and the same percentage of the authors focus on the use of AI for the prediction of treatments, in addition, 16.7% of the authors focus on the use of AI for the design of prostheses, 11.1% of the authors analyze the application of AI in the implantation of prostheses, finally, 5.6% of the articles are focused on the study of the application of AI in the integration of the nervous system. Robotics uses information about neuroscience for the implementation of hardware and control in biomedical engineering, implantation of myoelectric hands, hands with portable detection, and haptics [60]. On the other hand, to detect urinary bladder cancer, the multilayer perceptron method was found combined with other common methods, such as convolution neural networks focused on learning [61].

From the basic step of taking a patient's history to processing data and then extracting the information from the data for diagnosis, artificial intelligence has many applications in medical science [62]. AI along with virtual reality help analyze complex anatomy in three dimensions[63]. Likewise, AI improves performance in cancer diagnosis through immunotherapy, also optimizes treatment, predicts outcomes, and reduces personnel costs [34]. On the other hand, it can modify the design of clinical trials which begins with preparation until execution to reduce the R&D burden in terms of pharmacy [64]. For example, the purpose of this latest study was to implement an artificial neural feedback network (FFANN) to predict the KCF of the medial condyle corresponding to two different gait modifications known as medial thrust and trunk swing [58].

The diffusion of AI will depend on the forces that favor the reproduction of robots, systems, codes, and intelligent algorithms, as well as R&D groups and budgets. This is part of a special issue on the future of AI [64]. In conclusion,

artificial intelligence is a very productive factor in medical devices and many technologies today[65].

IX. CONCLUSION

First of all, the combination between medicine and technology was seen to generate various efficient medical methods that benefit medicine, AI helps to progress exponentially and allows the medicine to evolve in favor of humanity, however, it has a great capacity to learn and analyze possible diseases or deficiencies at a very high speed, allowing healthcare professionals to gain time in medical diagnosis to generate medical treatment in the shortest possible time, thus saving many more lives. Traditional prostheses only tried to replace the amputated limb for the displacement or simulation of human movement, however, thanks to the implementation of artificial intelligence, it is now possible to simulate or replicate the different senses that the lost limb possessed, thus giving greater perception, efficiency, and comfort to the user of the prosthesis. It can be observed that to develop a functional and effective prosthesis, an AI-assisted simulation must be carried out to have feedback that allows the prosthesis to be developed much faster and thus verify its correct operation, prevent various accidents or incompatibilities that may affect the user and avoid unnecessary expense in the construction of various prototypes still incomplete.

On the other hand, AI has not yet reached its peak, there is still much to explore and develop in the field of medicine and its advancement will be reflected in the fact of how many resources are used to deepen, develop and replicate the various investigations that promise to implement in a more effective and complex way the use of AI that encompasses the study of medicine and its branches. In short, both technology and medicine have been present for a long period. These technological advances are changing the vision we have about health and how they influence the knowledge of new technologies. Artificial intelligence consists of a series of logical instructions that machines use to be able to decide specific cases based on general principles. This technology is mainly used to make a diagnosis and health analysis of patients since in this way an individual prognosis can be appreciated. However, if we combine this technology and robotics, we can create machines and intelligent devices that can provide diagnostic suggestions or be more efficient in their work. Artificial intelligence will be a technology present in our day-to-day through machines or computer algorithms, in a true way for users, it will help in the daily life of the assistance processes. Professionals must know about this technology, its advantages, and its disadvantages because they will be integrated into the present work.

REFERENCES

[1] J. F. Avila-Tomás, M. A. Mayer-Pujadas, and V. J. Quesada-Varela, "Artificial intelligence and its applications in medicine I: antecedent

introduction to AI and robotics", *Aten. Primary*, vol. 52, no. 10, pp. 778–784, 2020.

[2] J. F. Ávila-Tomás, M. A. Mayer-Pujadas, and V. J. Quesada-Varela, "Artificial intelligence and its applications in medicine II: current importance and practical applications", *Aten. Primary*, vol. 53, no. 1, pp. 81–88, 2021.

[3] A. J. Rahyussalim, A. F. Marsetio, I. Saleh, T. Kurniawati, y Y. Whulanza, "The needs of current implant technology in orthopedic prosthesis biomaterials application to reduce prosthesis failure rate", *J. Nanomater.*, vol. 2016, pp. 1–9, 2016.

[4] Y. Mine, S. Suzuki, T. Eguchi, y T. Murayama, "Applying deep artificial neural network approach to maxillofacial prostheses coloration", *J. Prosthodont. Res.*, vol. 64, núm. 3, pp. 296–300, 2020.

[5] L. E. Fisher *et al.*, "Preliminary evaluation of a neural prosthesis for standing after spinal cord injury with four contact nerve-cuff electrodes for quadriceps stimulation", *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2006, pp. 3592–3595, 2006.

[6] elavancedelasprotesis, "Historia de las prótesis – El Avance de las Prótesis", El Avance de las Prótesis. [En línea]. Disponible en: <https://elavancedelasprotesis.wordpress.com/category/historia-de-las-protesis/>. [Consultado: 20-may-2022].

[7] "ACTUALIDAD EN PRODUCTOS SANITARIOS", Sefh.es. [En línea]. Disponible en: https://gruposdetrabajo.sefh.es/gps/images/stories/publicaciones/pam_2018_42%20411_256-259.pdf. [Consultado: 20-may-2022].

[8] S. Kumar, M. Nehra, D. Kedia, N. Dilbaghi, K. Tankeshwar, y K.-H. Kim, "Nanotechnology-based biomaterials for orthopedic applications: Recent advances and prospects", *Mater. Sci. Eng. C Mater. Biol. Appl.*, vol. 106, núm. 110154, p. 110154, 2020.

[9] A. Torres *et al.*, "CIBERESUCICOVID: a strategic project for a better understanding and clinical management of COVID-19 in critical patients", *Arch. Bronchopeumol.*, vol. 57 Suppl 2, pp. 1–2, 2021.

[10] J. Wojnarowski y K. Mirota, "Generalized criterion for in vitro testing of artificial heart valves", *J. Biomech.*, vol. 39, p. S619, 2006.

[11] J. A. Gegúndez Fernández, "Tecnificación versus humanización. Artificial intelligence at the service of medical diagnosis", *Arch. Soc. Esp. Oftalmol.*, vol. 93, no. 3, pp. e17–e19, 2018.

[12] C. Castellini, "Upper limb active prosthetic systems—overview", en *Wearable Robotics*, Elsevier, 2020, pp. 365–376.

[13] J. F. Avila-Tomás, M. A. Mayer-Pujadas, and V. J. Quesada-Varela, "Artificial intelligence and its applications in medicine I: antecedent introduction to AI and robotics", *Aten. Primary*, vol. 52, no. 10, pp. 778–784, 2020.

[14] D. Gu, K. Su, y H. Zhao, "A case-based ensemble learning system for explainable breast cancer recurrence prediction", *Artif. Intell. Med.*, vol. 107, núm. 101858, p. 101858, 2020.

[15] K. Nakamura, T. Morita, y T. Yamaguchi, "A user-centric platform PRINTEPS to develop integrated intelligent applications and application to robot teahouse", *Procedia Comput. Sci.*, vol. 112, pp. 2309–2318, 2017.

[16] R. Bravo and A.M. Lacy, "Medicine and Robotics," *Med. Clin. (Barc.)*, vol. 145, no. 11, pp. 493–495, 2015.

[17] T. Berger *et al.*, "Role of the hippocampus in memory formation: restorative encoding memory integration neural device as a cognitive neural prosthesis", *IEEE Pulse*, vol. 3, núm. 5, pp. 17–22, 2012.

[18] K. De Pauw *et al.*, "Prosthetic gait of unilateral lower-limb amputees with current and novel prostheses: A pilot study", *Clin. Biomech. (Bristol, Avon)*, vol. 71, pp. 59–67, 2020.

[19] D. B. Neill, "Using artificial intelligence to improve hospital inpatient care", *IEEE Intell. Syst.*, vol. 28, núm. 2, pp. 92–95, 2013.

[20] C. Castellini, "Upper limb active prosthetic systems—overview", en *Wearable Robotics*, Elsevier, 2020, pp. 365–376.

[21] A. Prieto, M. Atencia, y F. Sandoval, "Advances in artificial neural networks and machine learning", *Neurocomputing*, vol. 121, pp. 1–4, 2013.

[22] G. Kalnoor, "The brain-machine interface, nanosensor technology, and artificial intelligence: Their convergence with a novel frontier", in *Handbook of Nanomaterials for Sensing Applications*, Elsevier, 2021, pp. 575–587.

[23] Y. Chen, B. Xuan, Y. Geng, S. Ding, y L. Chen, "Modeling and control of knee-ankle-toe active transfemoral prosthesis", *IEEE Access*, vol. 8, pp. 133451–133462, 2020.

[24] D. J. Weber, M. Hao, M. A. Urbin, C. Schoenewald, y N. Lan, "Sensory information feedback for neural prostheses", en *Biomedical Information Technology*, Elsevier, 2020, pp. 687–715.

- [25] L. Letourneau-Guillon, D. Camirand, F. Guilbert, y R. Forghani, "Artificial intelligence applications for workflow, process optimization and predictive analytics", *Neuroimaging Clin. N. Am.*, vol. 30, núm. 4, pp. e1–e15, 2020.
- [26] S. T. Kakileti, H. J. Madhu, G. Manjunath, L. Wee, A. Dekker, y S. Sampangi, "Personalized risk prediction for breast cancer pre-screening using artificial intelligence and thermal radiomics", *Artif. Intell. Med.*, vol. 105, núm. 101854, p. 101854, 2020.
- [27] C. Bonivento, A. Davalli, y C. Fantuzzi, "Tuning of myoelectric prostheses using fuzzy logic", *Artif. Intell. Med.*, vol. 21, núm. 1–3, pp. 221–225, 2001.
- [28] O. W. Samuel *et al.*, "Intelligent EMG pattern recognition control method for upper-limb multifunctional prostheses: Advances, current challenges, and future prospects", *IEEE Access*, vol. 7, pp. 1–1, 2019.
- [29] V. Vemulapalli *et al.*, "Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data", *Artif. Intell. Med.*, vol. 74, pp. 1–8, 2016.
- [30] X. Y. Tai, H. Zhang, Z. Niu, S. D. R. Christie, y J. Xuan, "The future of sustainable chemistry and process: Convergence of artificial intelligence, data and hardware", *Energy and AI*, vol. 2, núm. 100036, p. 100036, 2020.
- [31] N. L. Opie *et al.*, "Micro-CT and histological evaluation of a neural interface implanted within a blood vessel", *IEEE Trans. Biomed. Eng.*, vol. 64, núm. 4, pp. 928–934, 2017.
- [32] A. Haleem, M. Javaid, R. P. Singh, y R. Suman, "Applications of Artificial Intelligence (AI) for cardiology during COVID-19 pandemic", *Sustainable Operations and Computers*, vol. 2, pp. 71–78, 2021.
- [33] B. Lepri, N. Oliver, y A. Pentland, "Ethical machines: The human-centric use of artificial intelligence", *iScience*, vol. 24, núm. 3, p. 102249, 2021.
- [34] V. A. S. Ramakrishna, U. Chamoli, G. Rajan, S. C. Mukhopadhyay, B. G. Prusty, y A. D. Diwan, "Smart orthopaedic implants: A targeted approach for continuous postoperative evaluation in the spine", *J. Biomech.*, vol. 104, núm. 109690, p. 109690, 2020.
- [35] A. H. Sadeghi *et al.*, "Virtual reality and artificial intelligence for 3-dimensional planning of lung segmentectomies", *JTCVS Tech*, vol. 7, pp. 309–321, 2021.
- [36] A. Badawy y R. Alfred, "Myoelectric prosthetic hand with a proprioceptive feedback system", *J. King Saud Univ. - Eng. Sci.*, vol. 32, núm. 6, pp. 388–395, 2020.
- [37] L. Leclercq, "Smart medical textiles based on cyclodextrins for curative or preventive patient care", en *Active Coatings for Smart Textiles*, Elsevier, 2016, pp. 391–427.
- [38] M. Pikhart, "Intelligent information processing for language education: The use of artificial intelligence in language learning apps", *Procedia Comput. Sci.*, vol. 176, pp. 1412–1419, 2020.
- [39] M. Pitkin, C. Cassidy, R. Muppavarapu, y D. Edell, "Recording of electric signal passing through a pylon in direct skeletal attachment of leg prostheses with neuromuscular control", *IEEE Trans. Biomed. Eng.*, vol. 59, núm. 5, pp. 1349–1353, 2012.
- [40] V. M. Petrovic, "Artificial intelligence and virtual worlds – toward human-level AI agents", *IEEE Access*, vol. 6, pp. 39976–39988, 2018.
- [41] "Historia de la Ingeniería Biomédica timeline", Timetoast timelines. [En línea]. Disponible en: <https://www.timetoast.com/timelines/evolucion-de-protesis>. [Consultado: 20-may-2022].
- [42] C. Y. Tecnología ElSancarlstaU, "Prótesis controladas por el cerebro", ElsancarlstaU.com, 25-sep-2017. [En línea]. Disponible en: <https://elsancarlstaU.com/2017/09/24/protesis-controladas-por-el-cerebro/>. [Consultado: 20-may-2022].
- [43] J. J. Huaroto, E. Suárez, y E. A. Vela, "Wearable mechatronic devices for upper-limb amputees", en *Control Theory in Biomedical Engineering*, Elsevier, 2020, pp. 205–234.
- [44] X. Wu *et al.*, "Artificial multisensory integration nervous system with haptic and iconic perception behaviors", *Nano Energy*, vol. 85, núm. 106000, p. 106000, 2021.
- [45] S. Micera *et al.*, "On the control of a robot hand by extracting neural signals from the PNS: preliminary results from a human implantation", *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2009, pp. 4586–4589, 2009.
- [46] C. Baladrón, J. J. Gómez de Diego, y I. J. Amat-Santos, "Big data and new information technology: what cardiologists need to know", *Rev. Esp. Cardiol. (Engl. Ed.)*, vol. 74, núm. 1, pp. 81–89, 2021.
- [47] J. F. Ávila-Tomás, M. A. Mayer-Pujadas, and V. J. Quesada-Varela, "Artificial intelligence and its applications in medicine II: current importance and practical applications", *Aten. Primaria*, vol. 53, no. 1, pp. 81–88, 2021.
- [48] A. Sun, B. Fan, y C. Jia, "Motor imagery EEG-based online control system for upper artificial limb", en *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, 2011.
- [49] A. Pokkalath, D. Nadar, P. Ravikumar, y S. P. Sawarkar, "Nanomaterials for orthopaedic implants and applications", en *Handbook on Nanobiomaterials for Therapeutics and Diagnostic Applications*, Elsevier, 2021, pp. 229–270.
- [50] H. A. Bayram, C.-H. Chien, y B. L. Davis, "Active functional stiffness of the knee joint during activities of daily living: a parameter for improved design of prosthetic limbs", *Clin. Biomech. (Bristol, Avon)*, vol. 29, núm. 10, pp. 1193–1199, 2014.
- [51] F. Khoshmanesh, P. Thurgood, E. Pirogova, S. Nahavandi, y S. Baratchi, "Wearable sensors: At the frontier of personalised health monitoring, smart prosthetics and assistive technologies", *Biosens. Bioelectron.*, vol. 176, núm. 112946, p. 112946, 2021.
- [52] Arcesw.com. [En línea]. Disponible en: <http://www.arcesw.com/pms1.htm>. [Consultado: 20-may-2022].
- [53] K. Devinuwara, A. Dworak-Kula, y R. J. O'Connor, "Rehabilitation and prosthetics post-amputation", *Orthop. Trauma*, vol. 32, núm. 4, pp. 234–240, 2018.
- [54] D. Dey, A. Lin, D. Han, y P. J. Slomka, "Computed tomography and artificial intelligence", en *Machine Learning in Cardiovascular Medicine*, Elsevier, 2021, pp. 211–239.
- [55] L. Bergamini, M. P. Sole, J.-D. Decotignie, y P. Dallemagne, "WiseTOP: a quality of service-aware low power acquisition and wireless communication platform for prosthesis control", en *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2019.
- [56] W. Batayneh, E. Abdulhay, y M. Alothman, "Prediction of the performance of artificial neural networks in mapping sEMG to finger joint angles via signal pre-investigation techniques", *Heliyon*, vol. 6, núm. 4, p. e03669, 2020.
- [57] K. Akrawinhawong, K. Majkut, S. Ferreira, y A. Mehdirad, "Voltage-dependent inappropriate right ventricular capture by right atrial lead pacing as a cause of cardiac resynchronization therapy non-responder", *J. Am. Coll. Cardiol.*, vol. 69, núm. 11, p. 2138, 2017.
- [58] F. Lamandé, J.-C. Dupré, P. Talbot, M. Gillet, T. Januscevic, and M. Dréjas-Zielinska, "Upper Limb Amputation", *EMC - Kinesitherapy - Med. Fis.*, vol. 35, no. 2, pp. 1–20, 2014.
- [59] S. Harrer, P. Shah, B. Antony, y J. Hu, "Artificial intelligence for clinical trial design", *Trends Pharmacol. Sci.*, vol. 40, núm. 8, pp. 577–591, 2019.
- [60] C. Potluri, M. Anugolu, D. S. Naidu, M. P. Schoen, y S. C. Chiu, "Real-time embedded frame work for sEMG skeletal muscle force estimation and LQG control algorithms for smart upper extremity prostheses", *Eng. Appl. Artif. Intell.*, vol. 46, pp. 67–81, 2015.
- [61] M. Santello *et al.*, "Hand synergies: Integration of robotics and neuroscience for understanding the control of biological and artificial hands", *Phys. Life Rev.*, vol. 17, pp. 1–23, 2016.
- [62] L. B. Yang, "Application of artificial intelligence in electrical automation control", *Procedia Comput. Sci.*, vol. 166, pp. 292–295, 2020.
- [63] D. Tandon y J. Rajawat, "Present and future of artificial intelligence in dentistry", *J. Oral Biol. Craniofac. Res.*, vol. 10, núm. 4, pp. 391–396, 2020.
- [64] Z. Xu, X. Wang, S. Zeng, X. Ren, Y. Yan, y Z. Gong, "Applying artificial intelligence for cancer immunotherapy", *Acta Pharm. Sin. B.*, vol. 11, núm. 11, pp. 3393–3405, 2021.
- [65] M. M. Ardestani *et al.*, "Feed forward artificial neural network to predict contact force at medial knee joint: Application to gait modification", *Neurocomputing*, vol. 139, pp. 114–129, 2014.
- [66] D. L. Waltz, "Evolution, sociobiology, and the future of artificial intelligence", *IEEE Intell. Syst.*, vol. 21, núm. 3, pp. 66–69, 2006.

Land-Mobile Robots for Rescue and Search: A Technological and Systematic Review

Deyby Huamanchahua
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 dhuamanchahua@utec.edu.pe

Kevin Aubert
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 kevin.aubert@utec.edu.pe

Mirella Rivas
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 mirella.rivas@utec.edu.pe

Eduardo Guerrero
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 eduardo.guerrero@utec.edu.pe

Laura Kodaka
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 laura.kodaka@utec.edu.pe

Diego Guevara
*Department of Electrical and
 Mechatronics Engineering
 Universidad de Ingeniería y
 Tecnología - UTEC*
 Lima, Perú
 diego.guevara@utec.edu.pe

Abstract—Due to the large number of emergencies and disasters happening around the globe, the technology development to face these difficulties has risen in different areas, one of them being the robotics field: rescue robots. A great amount of incentives have been presented over the years to increase the numbers of studies in the creation of rescue robots, like competitions and funding opportunities. The objective of writing this paper is to document and review different rescue robot works to provide, to interested individuals in the research area, a basis to start in this field. Twenty-six articles on rescue robotics development, ranging from 2017 to 2021, were found and selected using academic search engines.

Index Terms—rescue robots, TRL, sensors, systematic review

I. INTRODUCTION

In recent years, rescue robot development has reached an unprecedented speed due to them being one of the most effective tools in post-disaster search and extraction efforts. Around the world, emergencies pose a frequent and serious threat to the lives of people. These include natural disasters (e.g. earthquakes, landslides, floods), industrial accidents (spills, fires, explosions and structural failures) and terrorist attacks [1]. After a disaster, one of the main interests is to timely rescue and medically attend affected people, and to evacuate affected areas [2].

Due to the highly variable and risk-prone nature of after-accident environments [3], coupled with the low time constraints required for a successful rescue operation, human rescue workers and explorers have been phased out by the

use of modern rescue robots [4]. Recent examples of major disasters from around the world (such as earthquakes in Kobe, Turkey, Iran-Iraq and Fukushima [5] [6] [7]) exemplify the need for robust and dynamically-adaptable robots for S&R operations [8].

Because these robots have a wide range of potential markets and use cases, companies have been involved in stimulating the research and development of rescue robots. Currently, there are some examples of state-of-the-art commercially available robots such as HURCULES [9], a military-employed rescue and casualty extraction robot equipped with high-power actuator and motor systems, a fuzzy logic-based control system, multiple high-definition cameras and dedicated GPUs; and MSRBOTs [10], a coal-mine rescue robot equipped with hyperspectral binocular vision sensors (appropriate for the specialized rescue setting), gas sensors, and an electrohydraulic-powered locomotion system.

Notably, competitions such as the Robocup Rescue branch of the Robot World Cup [11] and the DARPA Robotics Challenge [12] provide a platform for simulation, construction and testing of S&R robots, and further the study of theory in rescue robotics. Five key metrics for rescue robots identified in common within these events include survivability, mobility, sensing, communicability and operability [13]. In order to achieve maximum performance in each of these metrics, the fusion of multiple design, construction and control techniques is considered to be required [14]. Several key insights can be extracted from the multiple published academic works derived from these series of competitions, some of which can be used to improve upon the aforementioned state-of-the-art robots.

Therefore, this article is a paper review summarizing these

TABLE I
RESCUE AND SEARCH ROBOTS

Reference	Sensors	CM	Actuators	DoF	Computing and Control Units	TRL
Ziegler, J. [15]	CO2 TC CAM LRF MC IMU	H	BML SDL SASM SAMP SDM	7	GC TC AVR	7
Liu, D. [16]	CO2 CAM LRF IMU	H	BML SDM	7	GC TC	2
Watanabe, A. [17]	CO2 CAM LRF MC IMU TS	H	DML SDL	5	OTH	6
Kitani, M. [18]	CO2 TC CAM LRF MC IMU	M	BML DML SDM	6	GC AVR	3
Habibian, S. [19]	CO2 TC CAM LRF MC IMU	H	DML SDM	7	GC ARM	6
Arbanas, B. [20]	CAM LRF IMU GPS	H	BML DML SDM	6	GC OTH	6
Yu Seop, S [21]	CO2 CAM MC	M	BML DML SDM	6	GC ARM OTH	4
Minami, Y. [22]	CO2 TC CAM LRF IMU	M	NS	6	GC ARM	5
Kohlbrecher, S. [23]	TC CAM LRF IMU GPS	H	NS	0	NS	6
Sharma, B. [24]	CAM IMU IS	M	NS	5	NS	7
Lin, X. [25]	CAM LRF IMU	H	BML BMM DMM	9	GC ARM	6
Xu, Q. [26]	CO2 TC CAM MC IMU	A	DML	0	GC	5
Francoeur, A. [27]	CO2 TC CAM LRF MC IMU	M	DML	0	TC	4
Najafi, F. [28]	CO2 TC CAM LRF IMU	H	DML SDM	6	GC ARM	3
Alizadeh, I. [29]	CO2 CAM LRF MC IMU TS US CS AS	H	DML SDM	8	OTH	6
Najafi, F. [30]	CO2 TC CAM LRF IMU TS	H	DML SDM	6	GC ARM	7
Phunopas, A. [31]	CO2 TC CAM LRF MC IMU TS	H	DML SDM	6	GC AVR PIC	6
Beller, D. [32]	CO2 TC CAM LRF	H	DML SDM	6	GC ARM	3
Yavuz, S. [33]	CO2 TC CAM LRF MC IMU TS GPS US	H	DML	0	GC TC AVR OTH	5
Jenabzadeh, M. [34]	CO2 TC CAM LRF MC IMU TS US CS AS	H	DML SDM	5	GC PIC AVR	6
Aguilar, L. [35]	CO2 TC MC TS HS	H	SDL SDM	3	AVR	5
Pescador, D. [36]	CO2 TC CAM MC TS	H	CML CEM DMM	6	GC TC AVR PIC	4
Shiotani, M. [37]	CAM LRF IMU MC	H	BML DML SDM CEM	6	TC ARM OTH	6
Kopiás, P. [38]	CO2 TC CAM LRF IMU CS GNSS	H	DML SDM	5	TC ARM	6
Szrek, J. [39]	TC CAM	H	DML	0	GC AVR	7
Zhang, X. [40]	CAM LRF IMU	H	DML	0	IPC ARM	6

Note: CM: Control Method, DoF: Degrees of Freedom, CAM: camera (include RGB camera, RGB-D camera, IP camera, stereo camera and other types of cameras except for the thermal ones), IMU: inertial measurement unit, LRF: laser rangefinder sensor, CO2: CO2 sensor, TC: thermal camera, MC: microphone, TS: temperature sensor, US: ultrasonic sensor, GPS: GPS sensor, CS: compass sensor, AS: accelerometer sensor, IS: inclinometer sensor, GNSS: GNSS sensor, HS: humidity sensor, H: Hybrid, M: Manual, A: Autonomous, BML: BLDC Motor for locomotion, SDL: Servo DC for locomotion, DML: DC Motor for locomotion, SASM: AC Servo Actuator Synchronous for manipulator, SAMP: Servo Actuator mini for manipulator, SDM: Servo DC Motor for manipulator, CEM: electric cylinder for manipulator, DMM: DC Motor manipulator, GC: General Computer (include Mini Computers, Laptops, Motherboards, PCs, and other types of computing units like these), TC: Tiny Computer (include NVIDIA Jetson TX2, Raspberry Pi, Single Board Computers and other types of computers like these), ARM: ARM Based microcontrollers and controller boards, AVR: AVR based microcontrollers and boards, PIC: PIC microcontrollers, IPC: Industrial PC, OTH: other computing or control units, NS: Not specified.

findings within rescue robot development, including research on sensors, control methods, actuators, technology readiness levels (TRLs), computing components and control units. The objective is to systematically review the trends in each of these design considerations measuring the frequency. Thereafter, conclusions will be drawn on the observed patterns.

II. METHODOLOGY

In order to conduct this article, we looked for different papers submitted in “Google Scholar”, “SpringerOpen”, RoboCup Rescue Team Description pages and papers, “ResearchGate”, “IEEE Xplore” and other academic search engines. For this search, we use a combination of the words “rescue”, “search”, “ground robots”, “sensors”, “actuators” and “RoboCup Rescue”. From this exploration, we found a total of 53 articles related to the topic of Rescue Robots in the range from the year 2017 until 2021.

However, in the end, we select 26 papers for the present review based on three reasons. The first one is the lack of access to some papers, especially the ones for the years 2020 and 2021. The second one is the absence of paper information

on the characteristics reviewed in this article. The third one is that many papers explain about the same robot, especially the ones from the RoboCup Rescue Team Description for the characteristic that this competition is taken abroad every year.

III. REVIEW OF PUBLICATIONS

The reviewed articles have been gathered and classified to understand the diversity of areas to consider before developing a rescue robot. This robots have been sorted by their control method, actuators, degrees of freedom (DoF), computing and control units, sensors and Technology Readiness Level (TRL). Table I shows a summary of these reviewed areas.

A. Sensors

In this area, a review is carried out about the sensors used to obtain data on the robot’s surroundings. Table II shows 14 different types of sensors found in the 26 revised papers. Which, cameras are the most common device with 96.15% for remote tele operator’s view of the environment, 3D mapping and image processing for victim identification [23] [22].

Furthermore, 80.77% include IMU sensors to obtain orientation, inclination and acceleration data [33] [31]. Besides, 76.92% include laser rangefinder sensors for mapping the surroundings [31] [30] and 30.77% have temperature sensors for two purposes. The first one is to measure the victim’s body temperature [36] [34] [29], while the second one is to acquire data on the ambient conditions of the operating environment [35].

In addition, 73.08% of the robots possess CO2 sensors, 65.38% include thermal cameras and 53.85% include microphones principally for human detection through the identification of respiratory activity, body heat or emitted sounds of the victims respectively [17] [23] [29]. Finally, other types of sensors are included each in less than 12% of the reviewed investigations.

TABLE II
PERCENTAGE OF SENSORS USAGE IN STUDIES

Sensor	Amount	
	Frequency	Percentage
CAM	25	96.15%
IMU	21	80.77%
LRF	20	76.92%
CO2	19	73.08%
TC	17	65.38%
MC	14	53.85%
TS	8	30.77%
US	3	11.54%
GPS	3	11.54%
CS	3	11.54%
AS	2	7.69%
IS	1	3.85%
GNSS	1	3.85%
HS	1	3.85%

Note: CAM: camera (include RGB camera, RGB-D camera, IP camera, stereo camera and other types of cameras except for the thermal ones), IMU: inertial measurement unit, LRF: laser rangefinder sensor, CO2: CO2 sensor, TC: thermal camera, MC: microphone, TS: temperature sensor, US: ultrasonic sensor, GPS: GPS sensor, CS: compass sensor, AS: accelerometer sensor, IS: inclinometer sensor, GNSS: GNSS sensor, HS: humidity sensor.

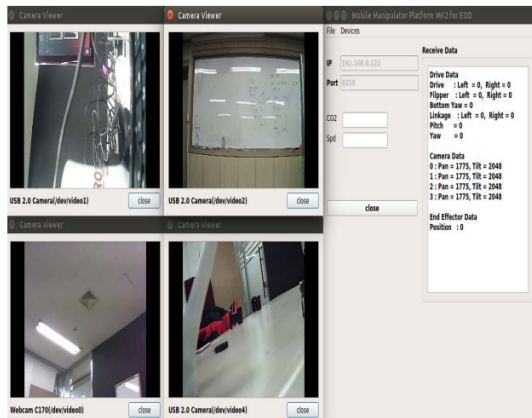


Fig. 1. Cameras for teleoperation by MIT Robotics Team [32].

B. Control Method

In this area, a review of the control method used to drive the rescue robot is performed, according to the research reviewed 3 methods were identified. The autonomous method, the hybrid method, and the manual method. The first one drives the robot using spatial identification algorithms and images to move in a place fulfilling certain functions. The second one uses teleoperation where the robot through artificial vision detects some tags that define its actions and its movement is given with the use of a joystick controller. The last one only navigates and actuates elements using the joystick controller.

As is shown in Fig. 1, all these methods are supported by a command station to analyze the robot’s cameras and sensors.

In Table III, the method that stands out the most is the hybrid method with 76.92% of the research conducted, while the manual method is used by 19.23% and only 3.85% could have an autonomous method for the most part.

TABLE III
PERCENTAGE OF CONTROL METHOD USAGE IN STUDIES

Control Method	Amount	
	Frequency	Percentage
H	20	79.92%
M	5	19.23%
A	1	3.85%

Note: H: Hybrid, M: Manual, A: Autonomous.

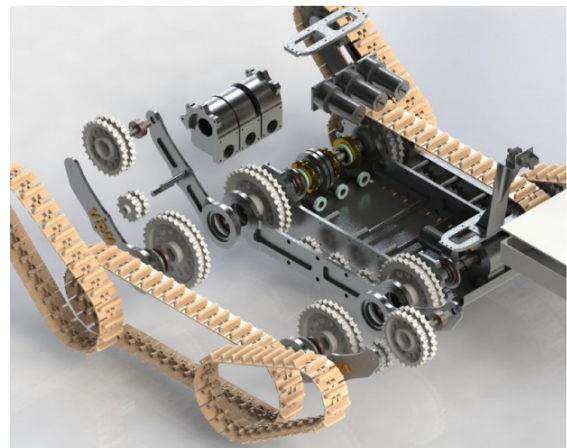


Fig. 2. Schematic of the actuators used by the VRU team [29].

C. Actuators

In this area, a review of the actuators used to generate the robot’s movements over the environment is performed. These actions can be classified into two types: locomotion and manipulative movement.

Table IV shows 8 different types of actuators found in the 26 papers reviewed. Which, DC motors are the most used for the locomotion of the robot with 69.23% since their operation is optimal for the rotation of gears, pulleys, or wheels.

In addition, 57.69% include DC Servomotors, which allow having a more specific angular movement for the manipulator movement. Also, 23.08% include BLDC motors for the robot movement since it has a fast response, high efficiency, and speed and 11.54% have DC Servomotors in the locomotion, which are more used for other purposes in robots, as seen above. While for robot arm joints 7.69% of robots have electric cylinders and a DC motor. And only 3.85% of robot arms have an AC servo motor and mini AC servo motor. The actuators are located on the robots as shown in the schematic diagram in Fig. 2

TABLE IV
PERCENTAGE OF ACTUATORS USAGE IN STUDIES

Actuators	Amount	
	Frequency	Percentage
DML	18	69.23%
SDM	15	57.69%
BML	6	26.92%
SDL	3	11.54%
NS	3	11.54%
CEM	2	7.69%
DMM	2	7.69%
SASM	1	3.85%
SAMM	1	3.85%

Note: BML: BLDC Motor for locomotion, SDL: Servo DC for locomotion, DML: DC Motor for locomotion, SASM: AC Servo Actuator Synchronous for manipulator, SAMM: Servo Actuator mini for manipulator, SDM: Servo DC Motor for manipulator, CEM: electric cylinder for manipulator, DMM: DC Motor manipulator.



Fig. 3. 6 DoF robotic arm of the MRL team [28].

D. Degrees of freedom (DoF)

In this area, a review of the degrees of freedom of exoskeleton prototypes is freedom that the robot arms of the rescue robots have. The robot arm is a device used as a manipulator that helps to mobilize elements and also allows holding objects to open doors or move levers. As shown in Table V, the largest number of robots according to the studies reviewed, with a percentage of 55.56%, have 6 DoF, as shown

in Fig. 3. In second place, robots with 5 DoF have a 22.22%, followed by a 16.67% with 7 DoF. Furthermore, robot arms that have 3 DoF have a percentage of 5.56%. While 23.08% have no robot arm. Besides, no robot arms with 4, 2, or 1 DoF were found.

TABLE V
PERCENTAGE OF DOF OF ROBOT ARM IN STUDIES

Degrees of freedom	Amount	
	Frequency	Percentage
9	1	3.85%
8	1	3.85%
7	3	11.54%
6	10	38.46%
5	4	15.38%
3	1	3.85%
0	6	23.08%

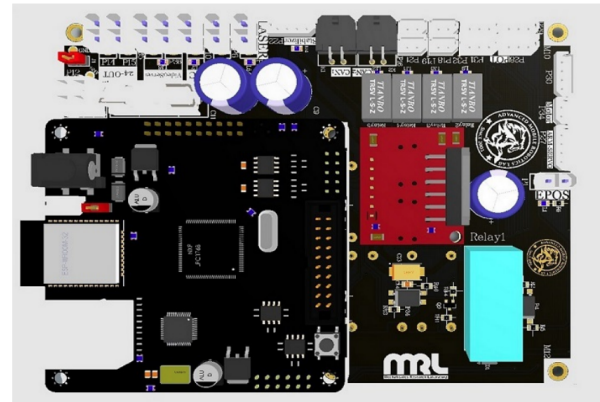


Fig. 4. Main Control Unit of ARKA robot, from the MRL Team [28].

E. Computing and Control Units

This area summarizes the main computing and control units of each robot. As shown in Table VI, the General computer units such as PCs, motherboards, mini computers or laptops, represent 65.38% of the computing and control units applied to robots. ARM based microcontrollers and controller boards have been used 38.46% of the time, while AVR based microcontrollers and boards followed closely behind with a percentage of 30.77%. On the other hand, 26.92% of the robots have used tiny computers such as NVIDIA Jetson TX2, Raspberry Pi or Single Board Computers, whereas other units not listed in Table VI represent 23.08% . PIC microcontrollers have been used less frequently, thus having 11.54% of appearances. Finally, Industrial PCs were the least applied units in the selected articles, with one appearance that represents the 3.85% of the times used. As presented in Fig. 4, the main control unit of the MRL Team robot, ARKA, is ARM based [28].

TABLE VI
PERCENTAGE OF COMPUTING AND CONTROL UNITS USAGE IN STUDIES

Computing and Control Units	Amount	
	Frequency	Percentage
GC	17	65.38%
ARM	10	38.46%
AVR	8	30.77%
TC	7	26.92%
OTH	6	23.08%
PIC	3	11.54%
IPC	1	3.85%
NS	2	7.69%

Note: GC: General Computer (include Mini Computers, Laptops, Motherboards, PCs, and other types of computing units like these), TC: Tiny Computer (include NVIDIA Jetson TX2, Raspberry Pi, Single Board Computers and other types of computers like these), ARM: ARM Based microcontrollers and controller boards, AVR: AVR based microcontrollers and boards, PIC: PIC microcontrollers, IPC: Industrial PC, OTH: other computing or control units, NS: Not specified.

F. Technology Readiness Level

In this area, a review is made of the level of development that the studies have reached. For a purpose of standardization, the present paper uses the Technology Readiness Levels to classify the advance of each investigation into 9 categories described in [41]. TRL 1 level corresponds to the observation of the principles in the study. TRL 2 and 3 refer to the formulation and the experimental proof of the technology concept respectively. TRL 4 is related to the validation of the components in a first basic prototype. TRL 5 and 6 correspond to the validation of an advanced prototype in a relevant environment like the RoboCup Rescue arenas. TRL 7 is related to the demonstration of the robot in a live environment like fire departments disaster trials or mines. Finally, TRL 8 and 9 refer to a successful proven system in operational environments

As is shown in Table VII, TRL 6 is the most common level with 42.31 %. This is due to the large amount of RoboCup Rescue teams included in this review. A consideration that in general represents that the prototypes had been validated in environments similar to the Robocup Rescue arenas like the one shown in Fig. 5. These are grounds that have different obstacles like stairs, ramps, gravel terrains or readiness tests. In addition, it can be seen that 73.07% of the review studies have an advanced prototype with a TRL of 5 or more.

TABLE VII
PERCENTAGE OF STUDIES BASED ON TRL

Technology Readiness Level	Amount	
	Frequency	Percentage
TRL 2	1	3.85%
TRL 3	3	11.54%
TRL 4	3	11.54%
TRL 5	4	15.38%
TRL 6	11	42.31%
TRL 7	4	15.38%

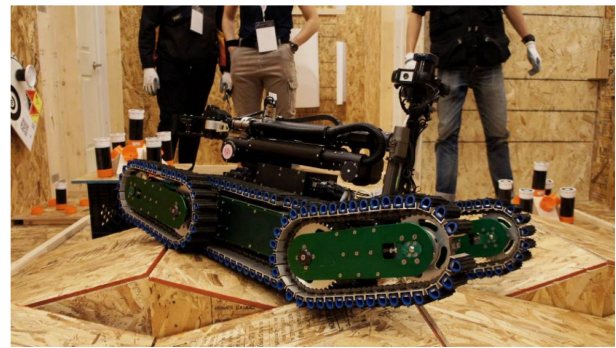


Fig. 5. Robot of iRAP SECHZIG team in a RoboCup Rescue arena [31].

IV. CONCLUSION

The development technology is having these days is making possible for a lot of difficulties in the world to be in the way of being solved. In a relatively new field like robotics, competition with funding opportunities as well as support to research departments made a great motivation to a lot of the teams who submitted the studied works in this article to create a path for new people to continue creating new designs for rescue robots for emergencies like natural disasters. This paper was intended to give a basic knowledge for interested individuals on how to begin in the creation of rescue robots, relaying in analysis of the trending components like sensors, computing and control units as well as Technology Readiness Level of each work.

It can be observed that most of the realized robots are still manipulated by humans using teleoperation, however it is getting closer and closer to being able to realize autonomous systems. For the operation, the cameras, the IMU sensors and the LRF sensors are the most common measurement devices. In addition, the locomotion of the robot, DC motors are used in most of the studies analyzed because the access to them and their use is more optimal for this task. While the configuration with 6 degrees of freedom is the most used in the manipulator of the studied robots. Also, the majority of investigations are in the stage of prototype validation in relevant and live environments.

Finally, the next step of our team will be the usage of these founded patterns in the investigation and construction of a land-mobile robot for search and rescue to participate in the 2023 edition of the RoboCup Rescue Robot League.

REFERENCES

- [1] Z. Xuexi, A. Yuming, F. Genping, L. Guokun, and L. Shiliu, "Survey on key technology of a robocup rescue robot," in *2019 Chinese control conference (CCC)*. IEEE, 2019, pp. 4746–4750.
- [2] L. Battistuzzi, C. T. Recchiuto, and A. Sgorbissa, "Ethical concerns in rescue robotics: a scoping review," *Ethics and Information Technology*, vol. 23, no. 4, pp. 863–875, 2021.
- [3] Z. Wang and H. Gu, "A review of locomotion mechanisms of urban search and rescue robot," *Industrial Robot: An International Journal*, 2007.
- [4] M. A. A. Hassan, "A review of wireless technology usage for mobile robot controller," in *Proceeding of the International Conference on System Engineering and Modeling (ICSEM 2012)*, 2012, pp. 7–12.
- [5] S. Tadokoro, "Earthquake disaster and expectation for robotics," in *Rescue Robotics*. Springer, 2009, pp. 1–16.

- [6] A. Shahri, M. Norouzi, A. Karambakhsh, A. Mashat, J. Chegini, H. Montazerzohour, M. Rahmani, M. Namazifar, B. Asadi, M. Mashat *et al.*, “Robocuprescue 2011-robot league team mrl rescue robot (iran),” *Vahid Khorani*, 2011.
- [7] N. Michael, S. Shen, K. Mohta, V. Kumar, K. Nagatani, Y. Okada, S. Kiribayashi, K. Otake, K. Yoshida, K. Ohno *et al.*, “Collaborative mapping of an earthquake damaged building via ground and aerial robots,” in *Field and service robotics*. Springer, 2014, pp. 33–47.
- [8] T. XU GH, “The development status and trend of mobile robots,” in *RTA*, vol. 3, 2001, pp. 7–14.
- [9] B. Choi, W. Lee, G. Park, Y. Lee, J. Min, and S. Hong, “Development and control of a military rescue robot for casualty extraction task,” *Journal of Field Robotics*, vol. 36, no. 4, pp. 656–676, 2019.
- [10] G. Zhai, W. Zhang, W. Hu, and Z. Ji, “Coal mine rescue robots based on binocular vision: A review of the state of the art,” *IEEE Access*, vol. 8, pp. 130561–130575, 2020.
- [11] H. Kitano, “Robocup rescue: A grand challenge for multi-agent systems,” in *Proceedings fourth international conference on MultiAgent systems*. IEEE, 2000, pp. 5–12.
- [12] G. Ishigami, K. Iagnemma, J. Overholt, and G. Hudas, “Design, development, and mobility evaluation of an omnidirectional mobile robot for rough terrain,” *Journal of Field Robotics*, vol. 32, no. 6, pp. 880–896, 2015.
- [13] G. Pratt and J. Manzo, “The darpa robotics challenge [competitions],” *IEEE Robotics & Automation Magazine*, vol. 20, no. 2, pp. 10–12, 2013.
- [14] P. Zhang, F. Gao, and F. Shuang, “Design and implementation of omni-directional mobile robot control system based on ros,” *Modular Machine Tool and Automatic Manufacturing Technique*, vol. 7, pp. 89–92, 2018.
- [15] J. Ziegler, J. Gleichauf, and C. Pfitzner, “Robocup rescue 2018 team description paper autonohm,” *Department of Electrical Engineering, TH Nuremberg Georg Simon Ohm., Nuremberg, Bavaria, Germany*, p. 4, 2018.
- [16] D. Liu, Y. Xu, T. Luo, C. Hailin, W. Jingbin, Z. Jinyan, Z. Xinye, and Z. Lingjin, “Robocup rescue team description paper sgbot,” 2018.
- [17] A. Watanabe, H. Miura, A. Wakayama, T. Kawaguchi, H. Mizugaki, R. Kamimura, and T. Teramoto, “Robocup rescue 2018 team description paper ait pickers,” 2018.
- [18] M. Kitani, T. Yokotani, K. Hoshino, K. Ushimaru, M. Totani, and N. Sato, “Robocup rescue 2018 team description paper nitro,” 2018.
- [19] S. Habibian, M. Dadvar, B. Peykari, A. Hosseini, M. H. Salehzadeh, A. H. Hosseini, and F. Najafi, “Design and implementation of a maxi-sized mobile robot (karo) for rescue missions,” *Robomech Journal*, vol. 8, no. 1, pp. 1–33, 2021.
- [20] B. Arbanas, F. Petric, A. Batinović, M. Polić, I. Vatavek, L. Marković, M. Car, I. Hrabar, A. Ivanović, and S. Bogdan, “From erl to mbzirc: Development of an aerial-ground robotic team for search and rescue,” 2021.
- [21] S. Yu Seop, T. Y. Lee, K. Yeong Jun, C. Yeong Hoon, J. Young Joon, L. Min Ho, and K. Jun Ho, “Robocup rescue 2017 team description paper robit,” 2017.
- [22] Y. Minami, G. Ramos, S. Hernandez, M. Rivero, C. Pineda, J. Azuara, and C. Garcia, “Robocup rescue 2017 team description paper finder-unam,”
- [23] S. Kohlbrecher, C. Rose, D. Koert, P. Manns, K. Daun, A. Stumpf, and O. von Stryk, “Robocup rescue 2017 team description paper hector darmstadt,” in *Proc. of the Intern. RoboCup Symposium*, 2017.
- [24] B. Sharma, B. M. Pillai, and J. Suthakorn, “Live displacement estimation for rough terrain mobile robot: Bart lab rescue robot,” in *2021 International Siberian Conference on Control and Communications (SIBCON)*. IEEE, 2021, pp. 1–6.
- [25] X. Lin, I. Cardenas, N. Kanyok, A. Shaker, P. Poudel, H. Jeong, N. Camacho, J. Butcher, J.-H. Kim, and G. P. Sharma, “Robocup rescue 2019 team description paper atr team,” 2019.
- [26] Q. Xu, Z. Shan, R. Li, X. Long, and S. Schwertfeger, “Robocup rescue 2019 team description paper mars-rescue,” 2019.
- [27] A. Francoeur, L. Vanasse, and M.-O. Bélisle, “Robocup rescue 2019 team description paper club capra,”
- [28] F. Najafi, H. Bagheri, N. Bonakdar Hashemi, A. Pouryayvali, M. Ahadi, S. Mohammad moshgforoush, M. Mojahedpour, M. Soltani, and A. Sharifi, “Robocup rescue 2019 team description paper mrl,”
- [29] I. Alizadeh, M. Alizadeh, R. Alizadeh, and M. Khalooei, “Robocup rescue 2017 team description paper vru.”
- [30] F. Najafi, M. Dadvar, S. Habibian, A. Hosseini, H. Haeri, M. Arvan, B. Peykari, and H. Bagheri, “Robocup rescue 2018 team description paper mrl.”
- [31] A. Phunopas, N. Pudchuen, and A. Blattler, “Robocup rescue 2019 team description paper irap sechzig.”
- [32] D. Beller, D. Mayo, J. Muller, M. Tan, R. Baird, and L. Beyer, “Robocup rescue 2017 team description paper mit robotics team.”
- [33] S. Yavuz, M. F. Amasyali, E. Uslu, F. Cakmak, N. Altuntas, S. Marangoz, M. B. Dilaver, and A. E. Kırılı, “Robocup rescue 2017 team description paper yildiz,” 2017.
- [34] M. Jenabzadeh, “Robocup rescue 2017 team description paper yra.”
- [35] L. Aguilar, A. Fernandez, C. Hernandez, J. Macias, E. Olivares, A. Perez, and I. Rangel, “Robocup rescue 2019 team description paper tecnobot,” 2019.
- [36] D. Pescador, A. Avalos, F. Perez, L. Bolanos, J. Hernandez, D. Cruz, D. Barboza, L. Gomez-Sanchez, J.-A. Rodriguez, M. Ramirez-Sosa, and D. Martinez-Peon, “Robocup rescue 2019 team description paper x-kau itnl.”
- [37] M. Shiotani, R. Kitamura, S. Iwamoto, T. Ono, T. Noake, and H. Wakiyama, “Robocup rescue 2018 team description paper nexis-r.”
- [38] P. Kopiás and M. Krauter, “Robocup rescue 2017 team description paper rescube.”
- [39] J. Szrek, R. Zimroz, J. Wodecki, A. Michalak, M. Góralczyk, and M. Worsa-Kozak, “Application of the infrared thermography and unmanned ground vehicle for rescue action support in underground mine—the amicos project,” *Remote Sensing*, vol. 13, no. 1, p. 69, 2020.
- [40] X. Zhang, J. Lai, D. Xu, H. Li, and M. Fu, “2d lidar-based slam and path planning for indoor rescue using mobile robots,” *Journal of Advanced Transportation*, vol. 2020, 2020.
- [41] Horizon-2020, “Work programme 2014- 2015 general annexes,” p. 29, 2015.

Efficient Simulation of Variable-Speed Diesel-Engine Generators Using Constant-Parameter Voltage-Behind-Reactance Formulation

Erfan Mostajeran

Department of Electrical and
Computer of Engineering
Univeristy of British Columbia
Vancouver, Canada
mostajeran@ece.ubc.ca

Arash Safavizadeh

Department of Electrical and
Computer of Engineering
Univeristy of British Columbia
Vancouver, Canada
arash.safavizadeh@ece.ubc.ca

Seyyedmilad Ebrahimi

Department of Electrical and
Computer of Engineering
Univeristy of British Columbia
Vancouver, Canada
ebrahimi@ece.ubc.ca

Juri Jatskevich

Department of Electrical and
Computer of Engineering
Univeristy of British Columbia
Vancouver, Canada
jurij@ece.ubc.ca

Abstract—Diesel-engine generators can operate with variable speed to enhance their efficiency over a wide range of loading conditions. Constant-parameter voltage-behind-reactance (CPVBR) models have been developed for efficient simulation of three-phase wound-field synchronous generators (WFSGs). The CPVBR models use fixed branches for interfacing (without any incompatibility, e.g., the need for snubbers, etc.) and are not required to update their interfacing circuit in every time step of the simulation. This paper presents the application of the CPVBR model of a WFSG for efficient simulation of a variable-frequency/speed diesel-engine generator-rectifier system. Simulation studies are carried out in MATLAB using PLECS toolbox to verify the computational performance of the CPVBR model. It is demonstrated that the CPVBR model achieves superior computational performance (without compromising the accuracy) compared to the built-in models of the toolbox, i.e., $qd0$, stator VBR, and full VBR models.

Keywords—Constant-parameter voltage-behind-reactance (CPVBR) model, diesel-engine generator (DEG), numerical efficiency, variable-speed, wound-field synchronous generator (WFSG).

I. INTRODUCTION

Diesel-engine generators (DEGs), aka gensets, are reliable sources for supplying energy for example in microgrids and remote residential areas. They are also considered as emergency back-up power supplies in hospitals [1], data centers [1], etc. DEGs usually operate with constant speed to have fixed nominal frequency at their terminal [2]. It has been demonstrated in the literature that the constant-speed regime hinders the efficiency of DEGs, and instead, adjusting the speed of the machine according to the system load level can increase their efficiency by reducing the fuel consumption [2], [3]. Hence, multiple control methods have been addressed in [1]–[3] to realize the variable-speed operation for DEGs.

In most studies involving DEGs [4], [5], the model for the generator is either in the form of the conventional $qd0$ model [6], or chosen from the available built-in models in simulation programs [7]–[9]. In the $qd0$ model, the generator is implemented as voltage-controlled current sources wherein the stator voltages and currents are considered as the inputs and the

outputs of the model, respectively [10]. This specific input-output relationship requires the usage of fictitious snubbers across the machine terminals to provide the necessary voltages for the model when the machine is connected to an inductive network or a power-electronic converter [10]. These artificial snubbers cause inevitable numerical stiffness and errors and may degrade the computational performance [10].

To address the interface incompatibility of the $qd0$ model, the coupled-circuit phase domain (CCPD) models have been developed which achieve direct interfacing capability (i.e., no need for snubbers, etc.) [11]. However, due to having rotor-position-dependent inductance matrices, the CCPD models are computationally expensive and slow [10], [12]. The voltage-behind-reactance (VBR) models have been proposed [12]–[14] to improve the computational performance by reducing the size of the rotor-position-dependent inductance matrices (compared to the CCPD model) into only one matrix, with the size of $n \times n$ with n being the number of machine phases. Although the CCPD and VBR models lead to a compatible interface with the rest of the network, they result in variable interfacing inductances which cause several simulation challenges. First, modeling of variable inductances is not readily available in all commercial simulation packages (e.g., PLECS [9] allows variable inductance but Simscape Electrical [8] does not). Secondly, the feature of variable-parameter interfacing forces the simulation solver to update the system matrix at each time-step which is computationally costly. To overcome these issues, constant-parameter VBR (CPVBR) models have been proposed in [15], [16] to not only achieve a direct interface similar to the VBR and CCPD models, but also to make the interfacing circuit constant for faster simulations [16].

This paper investigates the numerical performance of the CPVBR model presented in [16] for modeling the DEGs with variable speed/frequency operation, compared to the $qd0$ model and two VBR models [13], [14] available in the PLECS toolbox. It is demonstrated that the CPVBR model benefits from simplicity of implementation and low computational cost with acceptable accuracy compared to the existing counter-part models built in the PLECS library.

II. VBR FORMULATION FOR A THREE-PHASE WOUND-FIELD SYNCHRONOUS MACHINE

For the purpose of this paper, a conventional three-phase wye-connected synchronous machine with sinusoidal distribution of stator windings is assumed here. The machine has N damper windings on the q -axis and M damper windings in addition to one field winding on the d -axis. For having a consistent formulation as [6], [16], the motor convention is used, i.e., the positive direction for the stator currents is into the windings. All the variables are considered to be referred to the stator side. It should be noted that all the formulations are derived from the machine model initially expressed in the $qd0$ rotor reference frame [6], [16] as illustrated in Fig. 1. Also, for the purpose of notational consistency with [16], the vector of three-phase variables is denoted by $\mathbf{f}_{abc} = [f_a \ f_b \ f_c]$, and the vector of $qd0$ variables is denoted by $\mathbf{f}_{qd} = [f_q \ f_d]$. It should be noted that the zero-sequence variables are not considered here.

A. Variable-Parameter VBR

To set the stage for the CPVBR formulation, the derivation of the variable-parameter VBR model is presented first. The VBR formulation was first introduced in [12] by considering the rotor flux linkages as state variables leading to a current-input-voltage-output model. The full-order VBR model has the following stator interface [12], [16]

$$\mathbf{v}_{abc} = r_s \mathbf{i}_{abc} + p[\mathbf{L}'_{abc}(\theta_r) \mathbf{i}_{abc}] + \mathbf{e}''_{abc}, \quad (1)$$

where $\mathbf{L}'_{abc}(\theta_r)$ is the rotor-position-dependent subtransient inductance matrix as defined in [12], and \mathbf{v}_{abc} and \mathbf{i}_{abc} are stator voltages and currents, respectively. The subtransient voltages \mathbf{e}''_{abc} are as [12], [16]

$$\mathbf{e}''_{abc} = \mathbf{K}_s^{-1}(\theta_r) \mathbf{e}''_{qd}, \quad (2)$$

$$e''_q = \omega_r \lambda''_d + \sum_{j=1}^N \left(\frac{L''_{mq} r_{kqj}}{L_{lkqj}^2} (\lambda_{mq} - \lambda_{kqj}) \right), \quad (3)$$

$$e''_d = -\omega_r \lambda''_q + \sum_{j=1}^M \left(\frac{L''_{md} r_{kdj}}{L_{lkdj}^2} (\lambda_{md} - \lambda_{kdj}) \right) + \frac{L''_{md}}{L_{lfd}} v_{fd} + \frac{L''_{md} r_{fd}}{L_{lfd}^2} (\lambda_{md} - \lambda_{fd}), \quad (4)$$

where $\mathbf{K}_s^{-1}(\theta_r)$ is the inverse of Park's transform [6]; The resistances and leakage inductances of damper windings on the q - and d -axis are denoted by (r_{kqj}, L_{lkqj}) and (r_{kdj}, L_{lkdj}) , respectively. Similarly, the resistance and the leakage inductance of the field winding is specified by (r_{fd}, L_{lfd}) and its voltage is shown by v_{fd} . The subtransient magnetizing inductances L''_{mq} and L''_{md} in (3)–(4) are defined as [12], [16]

$$L''_{mq} = \left(\frac{1}{L_{mq}} + \sum_{j=1}^N \frac{1}{L_{lkqj}} \right)^{-1}, \quad (5)$$

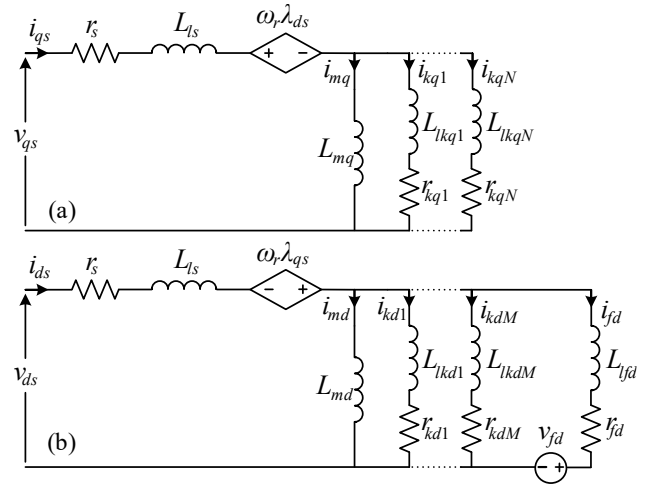


Fig. 1. The equivalent $qd0$ model of a three-phase synchronous machine expressed in rotor reference frame shown for: (a) q -axis, and (b) d -axis.

$$L''_{md} = \left(\frac{1}{L_{md}} + \frac{1}{L_{lfd}} + \sum_{j=1}^M \frac{1}{L_{lkdj}} \right)^{-1}, \quad (6)$$

where L_{mq} and L_{md} are the q - and d -axis magnetizing inductances, respectively. The subtransient flux linkages are also defined as

$$\lambda''_q = L''_{mq} \left(\sum_{j=1}^N \frac{\lambda_{kqj}}{L_{lkqj}} \right), \quad (7)$$

$$\lambda''_d = L''_{md} \left(\frac{\lambda_{fd}}{L_{lfd}} + \sum_{j=1}^M \frac{\lambda_{kdj}}{L_{lkdj}} \right), \quad (8)$$

and the magnetizing fluxes are obtained as

$$\lambda_{mq} = L''_{mq} \left(i_{qs} + \sum_{j=1}^N \frac{\lambda_{kqj}}{L_{lkqj}} \right), \quad (9)$$

$$\lambda_{md} = L''_{md} \left(i_{ds} + \frac{\lambda_{fd}}{L_{lfd}} + \sum_{j=1}^M \frac{\lambda_{kdj}}{L_{lkdj}} \right). \quad (10)$$

Subsequently, the state derivatives of the VBR model are represented as

$$p\lambda_{kqj} = -\frac{r_{kqj}}{L_{lkqj}} (\lambda_{kqj} - \lambda_{mq}); \quad j = 1, \dots, N, \quad (11)$$

$$p\lambda_{kdj} = -\frac{r_{kdj}}{L_{lkdj}} (\lambda_{kdj} - \lambda_{md}); \quad j = 1, \dots, M, \quad (12)$$

$$p\lambda_{fd} = -\frac{r_{fd}}{L_{lfd}} (\lambda_{fd} - \lambda_{md}) + v_{fd}, \quad (13)$$

where p is the Heaviside's operator for the derivative with respect to time. The mechanical subsystem is formulated similar to the $qd0$ model and is expressed as

$$p\theta_r = \omega_r, \quad (14)$$

$$p\omega_r = T_e - T_m, \quad (15)$$

where θ_r and ω_r are rotor position and angular electrical speed; T_m is the mechanical torque and the electromagnetic torque T_e is obtained by

$$T_e = \frac{3P}{4} (\lambda_{md} i_{qs} - \lambda_{mq} i_{ds}), \quad (16)$$

where P denotes the number of machine poles. Given the variable nature of the interfacing inductances in (1), the network solver has to re-calculate and re-construct the differential equations of the system at each time-step which significantly increases the model computational burden [15]. Additionally, modeling variable-parameter inductances is not available in many simulation packages [8].

B. Constant-Parameter VBR

To achieve a compatible and constant-parameter interface, multiple versions of CPVBR model were proposed in [16]. The main idea of all such models is to transfer the varying variables in the interfacing circuit to the subtransient voltages (i.e., \mathbf{e}''_{abc}) and represent the interfacing branches with constant RL elements. As discussed in [16], after moving the varying parameters to the subtransient voltages, the decoupled constant-interface circuit (assuming a wye (Y) connection in the stator windings and having access to the neutral) can be written as

$$\mathbf{v}_{abcs} = r_D \mathbf{i}_{abcs} + L_D p \mathbf{i}_{abcs} + L_0 p i_{ng} + \mathbf{e}''_{abc}, \quad (17)$$

where

$$r_D = r_s, \quad (18)$$

$$L_D = L_{ls} + L''_{md}, \quad (19)$$

$$L_0 = -\frac{1}{3} L''_{md}, \quad (20)$$

and i_{ng} is the current of the neutral path of the machine. The subtransient voltages in (17) in the qd frame are defined as

$$e''_q = \omega_r \lambda''_d + p(L''_q - L''_d) i_{qs} + \sum_{j=1}^N \left(\frac{L''_{mq} r_{kqj}}{L''_{lkqj}} (\lambda_{mq} - \lambda_{kqj}) \right), \quad (21)$$

$$e''_d = -\omega_r (L''_q - L''_d) i_{qs} + \sum_{j=1}^M \left(\frac{L''_{md} r_{kdj}}{L''_{lkdj}} (\lambda_{md} - \lambda_{kdj}) \right) - \omega_r \lambda''_q + \frac{L''_{md}}{L''_{lfd}} v_{fd} + \frac{L''_{md} r_{fd}}{L''_{lfd}} (\lambda_{md} - \lambda_{fd}), \quad (22)$$

where

$$L''_{qq} = L_{ls} + L''_{mq}, \quad (23)$$

$$L''_{dd} = L_{ls} + L''_{md}. \quad (24)$$

As noted in (21), this formulation needs the value of the derivative of stator current at the present time which has to be approximated by either a high-pass filter or other numerical methods to prevent the formulation of a nonproper state model [16]. As demonstrated in [16], this approximation leads to a

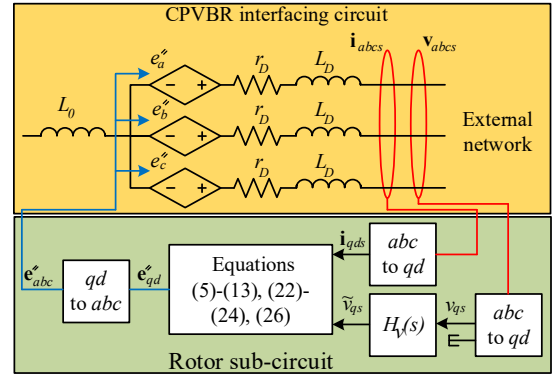


Fig. 2. Implementation of the CPVBR model using a low-pass filter $H_v(s)$ for relaxation of the algebraic loop.

larger error compared to the other following implementation where alternatively, the current derivative term is firstly obtained based on system inputs and states as [16]

$$p i_{qs} = \frac{1}{L''_q} (v_{qs} - r_s i_{qs} - \omega_r L''_d i_{ds} - \omega_r \lambda''_d) - \frac{1}{L''_q} \sum_{j=1}^N \left(\frac{L''_{mq} r_{kqj}}{L''_{lkqj}} (\lambda_{mq} - \lambda_{kqj}) \right). \quad (25)$$

Then, the new q -axis subtransient voltage can be written by substituting (25) in (21) as [16]

$$e''_q = \frac{L''_d}{L''_q} \left\{ \omega_r (\lambda''_d - (L''_q - L''_d) i_{ds}) + \sum_{j=1}^N \left(\frac{L''_{mq} r_{kqj}}{L''_{lkqj}} (\lambda_{mq} - \lambda_{kqj}) \right) \right\} + \frac{L''_q - L''_d}{L''_q} (v_{qs} - r_s i_{qs}). \quad (26)$$

Using (26), the need for approximating the derivative of stator current is eliminated. However, the presence of the stator voltage term v_{qs} in (26) forms an algebraic feedthrough in the model. Solving this algebraic loop requires iterative solutions of nonlinear equations, and hence increases the computational cost of the model. Relaxation of this algebraic loop can be achieved by using a low-pass filter as [16]

$$H_v(s) = \frac{p_0}{s + p_0}, \quad (27)$$

where $-p_0$ is the pole of the filter. The selection of the filter pole is a trade-off between the accuracy and stiffness of the system. As the magnitude of the pole of the low-pass filter increases, the accuracy improves at the cost of making the system stiffer. The systematic approach presented in [17] can be adopted to tune the pole of the filter for the desired accuracy.

It should be noted that the rest of the formulations of the CPVBR model are similar to that of variable-parameter VBR including the mechanical subsystem (14)–(16) and rotor flux linkages dynamics (11)–(13). A block-diagram of the implementation of the CPVBR model is depicted in Fig. 2. As seen, controlled voltage sources in series with decoupled and

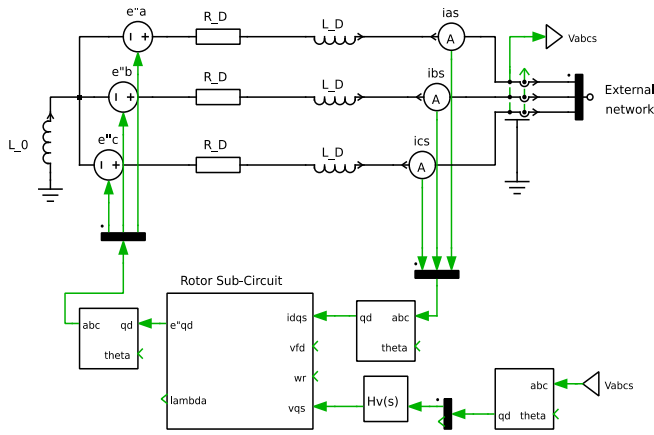


Fig. 3. Implementation of the CPVBR model in PLECS using conventional library components.

constant series resistances and inductances are used for interfacing with the external network. The implementation of the CPVBR model in PLECS environment is also illustrated in Fig. 3 using its conventional library components. It is noted that for a clear presentation, the mechanical subsystem is not shown in Fig. 3.

III. COMPUTER STUDIES

Here, the numerical performance of the CPVBR model is compared with the models available in the PLECS library [9]. Presently, PLECS has three built-in models for synchronous machines, i.e., *qd0*, stator VBR (SVBR) [13], and full VBR (FVBR) models [14]. As stated in [14], the two SVBR and FVBR models are algebraically equivalent and both result in variable-parameter interface which hinders their computational performance. The only difference between these two models is that the field sub-circuit in SVBR needs artificial snubbers to interface with an inductive network, whereas the FVBR model has a direct interface capability both in stator and field windings.

For the purpose of this paper, the DEG-rectifier system shown in Fig. 4 is implemented in MATLAB/Simulink using PLECS toolbox. Here, a three-phase salient wound-field synchronous generator (WFSG) is considered which supplies a resistive load through a six-pulse diode rectifier. All the parameters of the system are summarized in the Appendix. The voltage control loop consists of a PI controller and an actuator (modeled as a transfer function) which regulates the rectifier dc voltage at 1.0 pu by adjusting the voltage applied to the field winding. Also, to limit the voltage ripple introduced by the diode-bridge, a large LC filter is used at the output of the rectifier as shown in Fig. 4. To realize the variable speed operation of the generator and obtain the efficient power-speed trajectory, a considerable amount of data of the engine and the generator including saturation is needed [2]. Here, without loss of generality, it is assumed that the reference speed is changed from two arbitrary values at the moment of load change. In this regard, the regulation of the rotor speed around the set-point is achieved by the speed control loop consisting of another PI controller and an actuator which is modeled as a transfer function. A simplified diesel engine is also used similar to [4], which is modeled as a time-delay unit as shown in Fig. 4. Three fictitious snubbers are also required across the machine terminal in the *qd0* model. As the machine terminal voltages are regulated at the nominal value, the snubbers draw the same amount of

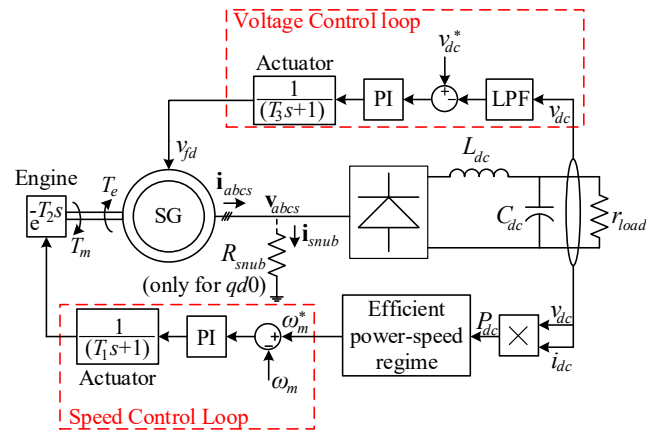


Fig. 4. Variable-speed DEG-rectifier system consisting of voltage and speed control loops supplying a dc load.

current at all operating points. Therefore, as the loading of the machine decreases from the nominal operating point, a larger portion of stator currents i_{abc} will be formed by snubbers currents i_{snub} . Hence, in different operating points, different values of snubbers are required to maintain the stator current error in an acceptable range (i.e., $\approx 1\%$). In this study, the snubbers are chosen to be 350 pu to ensure that the stator current error in the *qd0* model stays in the proximity of 1% for the considered loading conditions. Also, the pole of the low-pass filter required for CPVBR model is chosen to be -4300 , corresponding to the switching frequency of 360 Hz because of the six-pulse rectification [17]. This way, the solutions error of the CPVBR model remains in the acceptable range of 1%.

For consistency, all the subject models are run using MATLAB's stiff solver *ode23tb* with maximum and minimum time steps of 1ms and $0.01\mu s$, respectively. The absolute and relative errors are chosen to be 10^{-4} . The simulations studies are performed on a personal computer with Intel® Core™ i7-10750H CPU @ 2.60 GHz processor. Since the FVBR formulation includes no approximations/snubbers in its implementation, it is chosen as the reference model. The reference solution is obtained with the maximum time-step of $1\mu s$ using the same *ode23tb* solver.

In similar studies for comparing different models of ac machines [18], [19] in the presence of power-electronic converters, two different approaches are commonly adopted. In one approach, the speed of the machine is considered as an input to the system, e.g., in wind turbine studies. This way, the same speed is considered for all models which eliminates the need for solving the mechanical subsystem and excludes the impact of its error on simulation results [18]. The other alternative is the consideration of a grid-connected application where the grid provides a point of reference for the system [19]. These considerations cause all the results from different models to be synchronized, and a fair comparison can be drawn between all implementations by calculating the 2-norm error [16]. A similar approach is adopted here by inserting the speed profile obtained by the reference model into all subject models, and consequently, eliminating the impact of the error introduced by the mechanical subsystem in the simulation results. This approach is justifiable since the major differences between all the subject models originate from the properties of the electrical subsystem; and neglecting the impact of the error of the

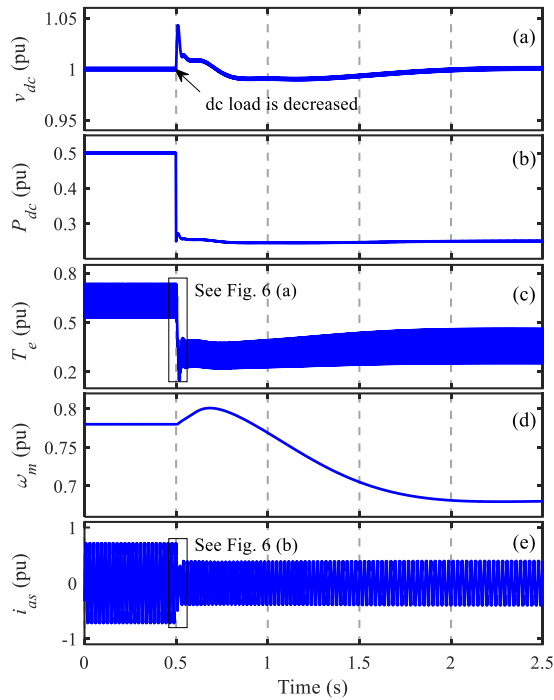


Fig. 5. Transient response of several variables when a load rejection of 0.25 pu and step down in reference speed occur at $t = 0.5$ s: (a) rectifier dc voltage, (b) dc load power, (c) electromagnetic torque, (d) rotor speed, and (e) stator phase- a current.

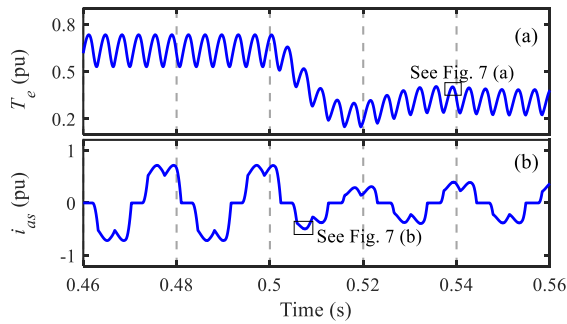


Fig. 6. Magnified view during the moment of transient: (a) electromagnetic torque in Fig. 5(c), and (b) stator phase- a current in Fig. 5(e).

mechanical subsystem does not affect the final conclusions about the numerical performance of the subject models.

For the considered transient study, the system in Fig. 4 is simulated for 2.5 seconds. The simulation results for the selected variables are provided in Fig. 5. As shown in Fig. 5(a)–(d), the system initially operates in the steady-state where the generator is supplying a load of 0.5 pu at the speed of 0.78 pu while the dc voltage is maintained constant at 1.0 pu. Then at $t = 0.5$ s, a 0.25 pu load rejection occurs (as shown in Fig. 5(b)) which leads to an immediate 0.08 pu over-voltage across the load, and it is regulated back to 1.0 pu by the control loop as shown in Fig. 5(a). At the moment of load change, the reference speed is also stepped down to 0.68 pu which is assumed to be the fuel-efficient speed according to the new loading condition. The governor adjusts the rotor speed to the new set point according to Fig. 5(d). As can be seen in Fig. 5(c), the electromagnetic torque profile follows the same transient as the dc power profile and decreases suddenly at the moment of load rejection. The stator phase- a current is also affected by the decrease in the load

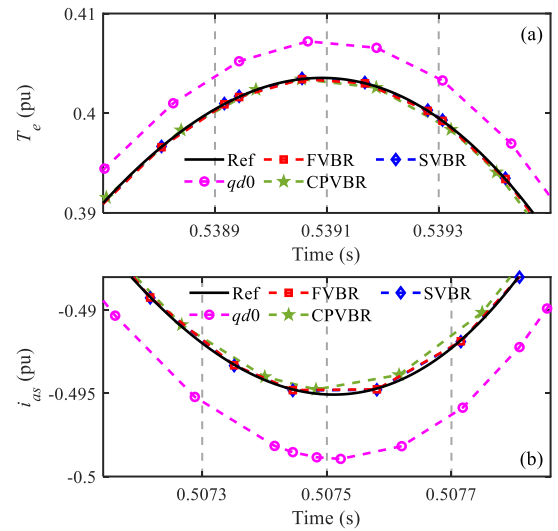


Fig. 7. Magnified view of variables in Fig. 6: (a) electromagnetic torque, and (b) stator phase- a current.

TABLE I. COMPARISON BETWEEN THE COMPUTATIONAL PERFORMANCE OF SUBJECT MODELS

Model	2-norm error		No. of steps (<i>ode23tb</i>) (<i>ode45</i>)	CPU time (<i>ode23tb</i>) (<i>ode45</i>)
	Torque (T_e)	Stator current (i_{as})		
<i>qd0</i>	1.05%	1.04%	53,918	5.71 s
			1,484,733	103.78 s
FVBR	0.03%	0.03%	41,928	10.92 s
			31,947	8.09 s
SVBR	0.02%	0.03%	42,808	6.48 s
			28,717	6.22 s
CPVBR	0.13%	0.15%	43,102	3.56 s
			35,162	3.25 s

as depicted in Fig. 5(e). To provide a more detailed snapshot of the considered transient about the moment of load rejection, the magnified view of electromagnetic torque and stator phase- a current are demonstrated in Fig. 6.

To qualitatively compare the results of the subject models with respect to the reference model, a further magnified view of the electromagnetic torque and the stator current are shown in Fig. 7. It is observed in Fig. 7 that the VBR models and the CPVBR model predicted the current and torque with great accuracy contrary to the *qd0* model which has a visibly large error due to the interfacing snubbers. For a quantitative comparison between the accuracy of the subject models, the cumulative 2-norm errors of the stator phase- a current and the electromagnetic torque for the considered transient study are calculated and provided in Table I. As observed in Table I, both FVBR and SVBR obtained the torque and current with very small error ($\sim 0.03\%$) which is the result of having a direct interface with the network. The *qd0* model has around 1% error both in torque and current due to having fictitious snubbers. Meanwhile, the CPVBR model obtained an acceptable (i.e., less than 1%) error in torque (0.13%) and in stator current (0.15%).

The comparison of the computational performance between the subject models using *ode23tb* (stiff) and *ode45* (non-stiff) solvers are also included in Table I. As seen in Table I, the CPVBR model achieves the fastest simulations compared to

other models. Specifically, when using the stiff solver *ode23tb*, the CPVBR model allows 38%, 67%, and 45% faster simulation compared to the *qd0* (3.56s vs. 5.71s), FVBR (3.56s vs. 10.92s), and SVBR (3.56s vs. 6.48s) models, respectively. It is worth mentioning that, as a result of having constant inductances, the *qd0* model was faster than the FVBR and SVBR models which have variable-inductances in their interface, although the *qd0* model needs more time steps (i.e., 53,918 vs. 41,928 and 42,808, respectively). It is noted that the number of steps used by the CPVBR model is on the same order as the VBR models (i.e., 43,102). Using the non-stiff solver *ode45*, the *qd0* model takes significantly more time (103.78s) and computational resources (1,484,733 time steps) due to having large snubbers across the machine terminals to maintain the acceptable accuracy. Meanwhile, the CPVBR outperforms all other models by achieving 97%, 60%, and 48% efficiency in CPU time compared to the *qd0* (3.25s vs. 103.78s), FVBR (3.25s vs. 8.09s), and SVBR models (3.25s vs. 6.22s), respectively. It should be noted that the number of time steps of the two VBR models (31,947 for FVBR and 28,717 for SVBR) and the CPVBR model (35,162) models are significantly fewer than the time steps used by the *qd0* model.

IV. CONCLUSION

In this paper, the application of the previously-introduced CPVBR model has been investigated for simulation of variable-speed diesel-engine generators (DEGs) consisting of wound-field synchronous generators (WFSGs). Due to having a current-input-voltage-output model, CPVBR model can be connected to power-electronics-based networks without any snubbers contrary to the conventional *qd0* model. Also, as a result of having a constant-parameter interfacing circuit, it is computationally more efficient than other variable-parameter VBR models. The computational advantages of the CPVBR model were validated through offline simulation studies. It was verified that the CPVBR model could achieve the fastest CPU time compared to the available built-in models of synchronous machines in PLECS for simulation of DEGs while having acceptable error of about 0.1~0.2%.

APPENDIX

Three-phase WFSG parameters [8]:

Rated ac voltage: 2.4 kV, rated dc voltage: 3.24 kV, rated power 3.125 MVA, poles: 4, rated speed: 1800 rpm, inertia time-constant: 1.07s.

$r_s = 0.0036$ pu, $X_{ls} = 0.052$ pu, $X_{md} = 1.508$ pu, $X_{mq} = 1.008$ pu, $r_{kd} = 0.0616$ pu, $X_{lkq} = 0.1532$ pu, $r_{fd} = 0.00245$ pu, $X_{lfd} = 0.2911$ pu, $r_{kd} = 0.0159$ pu, $X_{lkd} = 0.2563$ pu.

DC filter parameters:

$L_f = 1$ mH, $C_f = 10$ mF.

Voltage controller parameters [20]:

$K_p = 500$, $K_I = 10$, $T_3 = 0.02$ s.

Speed controller parameters [4]:

$K_p = 6.09$, $K_I = 5.9$, $T_1 = 0.1$ s, $T_2 = 0.1$ s.

REFERENCES

- [1] I. Choi, Y. Jeung, and D. Lee, "Variable speed control of diesel engine-generator using sliding mode control," in *Proc. IEEE Transport. Elec. Conf. and Expo. Asia-Pacific (ITEC Asia-Pacific)*, pp. 1–5, Oct. 2017.
- [2] M. Mobarra, M. Rezkallah, and A. Ilinca, "Variable speed diesel generators: Performance and characteristic comparison," *Energies*, vol. 15, no. 2, p. 592, Jan. 2022.
- [3] J. H. Lee, S. H. Lee, and S. K. Sul, "Variable-speed engine generator with supercapacitor: Isolated power generation system and fuel efficiency," *IEEE Trans. Ind. Appl.*, vol. 45, no. 6, pp. 2130–2135, Nov. 2009.
- [4] T. Theubou, R. Wamkeue, and I. Kamwa, "Dynamic model of diesel generator set for hybrid wind-diesel small grids applications," in *Proc. IEEE Canadian Conf. on Elec. and Comp. Eng. (CCECE)*, pp. 1–4, Oct. 2012.
- [5] B. Singh and J. Solanki, "Load Compensation for diesel generator-based isolated generation system employing DSTATCOM," *IEEE Trans. Ind. Appl.*, vol. 47, no. 1, pp. 238–244, Apr. 2007.
- [6] P. C. Krause, O. Wasynczuk, and S. D. Sudhoff, "Analysis of electric machinery and drive systems," 3rd ed., NJ, USA: IEEE Press, 2013.
- [7] (2022). Simulink Dynamic System Simulation Software Users Manual (MathWorks Inc.). [Online]. Available: www.mathworks.com
- [8] (2022). Simscape Electrical User's Guide (Electronics, Mechatronics, and Power Systems) (The MathWorks Inc.). [Online]. Available: www.mathworks.com
- [9] Piecewise Linear Electrical Circuit Simulation (PLECS) User Manual, Version. 4.6.5, Plexim GmbH, 2022. [Online]. Available: www.plexim.com
- [10] L. Wang *et al.*, "Methods of interfacing rotating machine models in transient simulation programs," *IEEE Trans. Power Deliv.*, vol. 25, no. 2, pp. 891–903, Apr. 2010.
- [11] X. Cao, A. Kurita, H. Mitsuma, Y. Tada, and H. Okamoto, "Improvements of numerical stability of electromagnetic transient simulation by use of phase-domain synchronous machine models," *Elec. Eng. Jpn.*, vol. 128, no. 3, pp. 53–62, Apr. 1999.
- [12] S. D. Pekarek, O. Wasynczuk, and H. J. Hegner, "An efficient and accurate model for the simulation and analysis of synchronous machine/converter systems," *IEEE Trans. Energy Convers.*, vol. 13, no. 1, pp. 42–48, Mar. 1998.
- [13] D. C. Aliprantis, O. Wasynczuk and C. D. Rodríguez Valdez, "A voltage-behind-reactance synchronous machine model with saturation and arbitrary rotor network representation," *IEEE Trans. Energy Convers.*, vol. 23, no. 2, pp. 499–508, Jun. 2008.
- [14] A. M. Cramer, B. P. Loop, and D. C. Aliprantis, "Synchronous machine model with voltage-behind-reactance formulation of stator and field windings," *IEEE Trans. Energy Convers.*, vol. 27, no. 2, pp. 391–402, Jun. 2012.
- [15] M. Chapariha, L. Wang, J. Jatskevich, H. Dommel and S. D. Pekarek, "Constant-parameter RL-branch equivalent circuit for interfacing AC machine models in state-variable-based simulation packages," *IEEE Trans. Energy Convers.*, vol. 27, no. 3, pp. 634–645, Sep. 2012.
- [16] M. Chapariha, F. Therrien, J. Jatskevich and H. W. Dommel, "Explicit formulations for constant-parameter voltage-behind-reactance interfacing of synchronous machine models," *IEEE Trans. Energy Convers.*, vol. 28, no. 4, pp. 1053–1063, Dec. 2013.
- [17] F. Therrien, M. Chapariha and J. Jatskevich, "Pole selection procedure for explicit constant-parameter synchronous machine models," *IEEE Trans. Energy Convers.*, vol. 29, no. 3, pp. 790–792, Sep. 2014.
- [18] S. Ebrahimi, N. Amiri, L. Wang and J. Jatskevich, "Efficient modeling of six-phase PM synchronous machine-rectifier systems in state-variable-based simulation programs," *IEEE Trans. Energy Convers.*, vol. 33, no. 3, pp. 1557–1570, Sep. 2018.
- [19] N. Amiri, S. Ebrahimi, J. Jatskevich and H. W. Dommel, "Saturable and decoupled constant-parameter VBR model for six-phase synchronous machines in state-variable simulation programs," *IEEE Trans. Energy Convers.*, vol. 34, no. 4, pp. 1868–1880, Dec. 2019.
- [20] I. D. Hassan, R. Weronick, R. M. Bucci and W. Busch, "Evaluating the transient performance of standby diesel-generator units by simulation," *IEEE Trans. Energy Convers.*, vol. 7, no. 3, pp. 470–477, Sep. 1992.

Detection and analysis types of DDoS attack

1st Erkin Navruzov

*Faculty of Applied Mathematics and
Intelligent Technologies
National University of Uzbekistan
Tashkent, Uzbekistan
erkinbek0989@gmail.com*

2nd Anvar Kabulov

*Faculty of Applied Mathematics and
Intelligent Technologies
National University of Uzbekistan
Tashkent, Uzbekistan
anvarkabulov@gmail.com*

Abstract—The problem of detecting types of DDOS attacks in large-scale networks is considered. The complexity of detection is explained by the presence of a large number of connected and diverse devices, the high volume of incoming traffic, the need to introduce special restrictions when searching for anomalies. The technology of developing information security models using data mining (DM) methods is proposed. The features of machine learning of DM algorithms are related to the choice of methods for preprocessing big data (Big Data). A technique for analyzing the structure of relations between types of DDOS attacks has been developed. Within the framework of this technique, a procedure for pairwise comparison of data by types of attacks with normal traffic is implemented. The result of the comparison is the stability of features, the values of which are invariant to the measurement scales. The analysis of the structure of relations by grouping algorithms was carried out according to the stability values on the determined sets of features. When forming the sets, the stability ranking was used. For classification, various existing methods of machine learning are analyzed.

Index Terms—types of DDOS attacks, Machine Learning, classification, informative features

I. INTRODUCTION

A DDoS attack is the actions of intruders aimed at disrupting the performance of a company's infrastructure and client services. To carry out a DDoS attack, hackers use the so-called botnet networks, which create an avalanche-like increase in requests to an online resource in order to increase the load on it and disable it. Protecting a system from DDOS attacks is about detecting and preventing an attack before it affects the end user. At the same time, detection should be performed with a high attack detection rate and a low false positive rate [1]. Recently, data mining has become an important component to prevent DDOS attacks.

DDoS attack detection methods play a very important role in protecting the security of computer networks. However, the existing methods of detecting DDoS attacks based on the flow face a significant time delay and are not common for different types of DDOS attacks with different speeds [2], [30]–[32]. To fill this gap in research, a quick approach to detecting DDOS attacks based on informative signs is proposed.

Machine learning can be defined as the ability of a program (system) to learn in order to improve its performance

relative to a given task for a certain time [13], [14]. Machine learning methods focus on building a system that improves its performance based on previous results, that is, machine learning methods have the ability to change their execution strategy based on newly acquired knowledge [3]. This feature is the main advantage of this approach, the main disadvantage is resource intensity. In many cases, machine learning methods coincide with statistical methods and data mining methods.

The main prerequisite for creating machine learning classifiers is the effective and efficient detection of DDOS attacks. However, their achievement of machine learning models to distinguish between DDOS attacks depends on how to choose appropriate and minimum attributes in network flows [4], [33]. A method for the formation of a new feature space from heterogeneous (quantitative and nominal) features is proposed [8], [34]–[37]. The formation process is based on a nonlinear transformation of features using the functions of belonging objects to classes [6]–[8].

The purpose of the work is to review existing algorithms for classifying data based on machine learning, evaluate their effectiveness and develop a method for detecting network DDOS attacks using machine learning algorithms [5]. Its achievement involves solving the following main tasks: selection of a training data set, data preprocessing, formation of a feature space, justification of the choice of a machine learning model, quality assessment and testing of the model in real conditions.

DDOS attack detection can be viewed as a classification problem [6], [7]. It is required to classify objects as representing a certain type of attacks and normal traffic. Classification is one of the 5 standard types of patterns used in data mining [11], [12]. Its main purpose is to predict whether new objects belong to classes [9], [10].

In order to achieve maximum effect in forecasting both accuracy and resource costs, the tasks of detecting hidden patterns through an informative feature are considered. In this paper, an experimental comparison of various machine learning models used to detect computer attacks is carried out. The training and testing of models was performed on one of the most relevant data sets, CICDDOS2019, was selected among the available public datasets. Previously, an analysis of the significance of the features was carried out and a reduction in the dimension of the feature space was performed.

II. CICDDOS2019 DATASET. PROBLEM STATEMENT

The data set contains 12 different types of DDoS attacks, which were described using the TCP/UDP protocols. Classification of attacks in the data set is made in terms of attacks based on exploitation and reflection. The main purpose of forming the set was to use it for training and testing. Attacks identified as UDP, SNMP, NetBIOS, LDAP, TFTP, NTP, SYN, WebDDoS, MSSQL, UDP-Lag, DNS, and SSDP are included as attack types in the training. As a separate set, testing data was generated, which contains 7 types of DDoS attacks: SYN, MSSQL, UDP-Lag, LDAP, UDP, PortScan and NetBIOS. PortScan attack data is not included in the test and training set for internal evaluation of the detection system. The data description contains 88 quantitative and qualitative characteristics of the flow, which were obtained using CICFlowMeter tools. The CICDDoS2019 data is available on the website of the Canadian Cyber Security Institute. Brief information about the CICDDoS2019 data [15] is contained in Table 1 [39]–[43].

TABLE I. Values of compactness measure by types of DDOS attacks.

	Type of DDOS attacks	Number of objects
1	Benign	56 863
2	DNS	5071011
3	LDAP	2179930
4	MSSQL	4522492
5	NetBIOS	4093279
6	SNMP	5159870
7	UDP-Lag	366461
8	WebDDoS	439
9	SYN	1582289
10	NTP	1202642
11	SSDP	2610611
12	UDP	3134645
13	TFTP	20082580
	Total	50006249

A set of $E_0 = \{S_1, \dots, S_m\}$ objects is considered, divided into l disjoint subsets (classes) $K_1, K_2, \dots, K_l, E_0 = \bigcup_{i=1}^l K_i$. The sample objects are described by a set of n different types of features $X(n) = \{x_1, \dots, x_n\}$, ξ of which are measured in interval scales, $(n - \xi)$ – in nominal.

Required:

- according to the description of objects E_0 in $X(n)$, form a set of features $Y(n)$ in the interval scale of measurements;
- the description of the class $K_i, i = 1, \dots, l$ by $Y(n)$ is represented as a row in the table “object-property” $T(l, n)$;
- to investigate the structure of the connections of objects from $T(l, n)$ by dividing them into a given number of disjoint groups according to the specified sets of features from $Y(n)$.

A. Data preprocessing

When developing a model based on machine learning algorithms, it is important to decide which features should be used

as input data [38]–[45].

Data preprocessing consists of the following steps:

- removing socket features such as source and destination IP address, source and destination port, timestamp and stream ID;
- saving one representative from repeating objects;
- removing objects with omissions (unmeasured values) in the data.

According to the results of data preprocessing, more than 10 million objects described by a set of 80 features were obtained. The number of objects for each type of DDOS attacks is shown in Table 2.

TABLE II. Values of compactness measure by types of DDOS attacks.

	Type of DDOS attacks	Number of objects
1	Benign	50635
2	DNS	108211
3	LDAP	28871
4	MSSQL	193656
5	NetBIOS	18017
6	SNMP	120332
7	UDP-Lag	89085
8	WebDDoS	414
9	SYN	155519
10	NTP	1131109
11	SSDP	958766
12	UDP	1156894
13	TFTP	6109716
	Total	10070590

Due to large amounts of data, attempts to solve the problem of selecting informative features [1,2,9] by classical methods for 12 classes of DDOS attacks are unpromising. Based on the specifics of the selection task, its solution is proposed to be sought as follows. Consider data samples $\Omega_i(K_1, K_2), i=1,2,\dots,12$ in which class K_1 is represented by normal traffic, K_2 is one of the types of DDOS attacks. For each sample $\Omega_i(K_1, K_2)$ arrange the features from $X(n), n=80$ by their stability values (stability calculation is given below). To form 12 sequences of features based on the results of their ordering on $\Omega_i(K_1, K_2)$.

The goals that are pursued in the selection of informative features.

- Ordering the types of DDOS attacks according to the complexity of their detection;
- Selection and justification of rules for classification of types of DDOS attacks.

In the presence of large volumes of initial data, it is relevant to solve the problem of choosing and justifying algorithms for calculating the values of the stability of features.

B. Formation of a feature space

When developing a machine learning model, it is important to decide which features should be used as input for the learning algorithm [11]. In some problems, the dimension of the feature space can be very large [12].

A method of forming a new feature space from different types (quantitative and nominal) features is proposed [11]. The formation process is based on the nonlinear transformation of features using the class membership functions of objects.

The inclusion of features in the informative set is based on their ranking with respect to stability. The implementation of the proposals is associated with the modification of the algorithms of the interval method described in [19]. The essence of the modification is to reduce the combinatorial complexity of the algorithms of the method due to the large volume of source data.

The formation of a feature space is an important stage in solving classification problems [18]. The specificity of the considered problem of space formation is that the raw features are presented in different scales and scales of measurements [19]. To achieve invariance to the measurement scales, it is proposed to use the transformation of quantitative features using interval methods.

The choice of both the number and the boundaries of the division of quantitative features into intervals is widely practiced in solving problems of discriminant data analysis [16], [17]. Within the scope of this study, it is not possible to make a complete review and qualitative analysis of existing methods for such a choice.

There is a need to apply interval methods to calculate generalized estimates of objects. Generalized estimates are considered as latent features, the analysis of which is aimed at finding hidden patterns in the data. For decision making, the composition of the set of raw features used in calculating the values of generalized estimates is of interest [17], [18].

A hypothesis test is required: For each sample of $\Omega_i(K_1, K_2)$, there is a unique set of raw features for making a decision on a specific type of DDOS attacks. When proving the truth of the hypothesis, it is proposed to use the method of calculating generalized estimates of objects.

In the classical version, the calculation of generalized estimates by a set of different types of features is performed by a stochastic algorithm. The results of the stochastic algorithm analysis depend on the choice of initial approximations. The question remained open which of the sets should be considered informative for decision-making. To answer this question it is suggested:

- To convert the values of quantitative characteristics into nominal gradations;
- Implement the calculation of generalized estimates by a deterministic algorithm;
- Use a hierarchical agglomerative feature grouping algorithm to calculate generalized object ratings.

The implementation of the proposals is associated with the modification of the algorithms of the interval method described in [19]–[21]. The essence of the modification is to reduce the combinatorial complexity of the algorithms of the method due to the large amount of initial data.

In this paper, we consider the division into intervals for each quantitative feature according to the $\Omega_i(K_1, K_2)$, $i = 1, \dots, 12$ sample, within the boundaries of which the values

of objects of class K_1 or K_2 dominate. To do this, the values of the $x_c \in X(n)$ attribute are ordered in ascending order

$$r_{c_1}, r_{c_2}, \dots, r_{c_m}. \tag{1}$$

According to the criterion defined below, the sequence (1) is divided into τ_c , ($\tau_c \geq 2$) disjoint intervals $[r_{c_u}, r_{c_v}]^i$, $1 \leq u, u \leq v \leq m$, $i = \overline{1, \tau_c}$. The values of the features lying in the interval $[r_{c_u}, r_{c_v}]^i$ can then be considered as a gradation of the nominal feature.

Let $d_1^i(u, v)$, $d_2^i(u, v)$ be the number of representatives of classes K_1, K_2 respectively, in the interval $[r_{c_u}, r_{c_v}]^i$. For the recursive procedure for selecting r_{c_u}, r_{c_v} values, the

$$\left| \frac{d_1^i(u, v)}{|K_t|} - \frac{d_2^i(u, v)}{|K_{3-t}|} \right| \rightarrow \max. \tag{2}$$

criterion is used. The boundaries of the first interval $[r_{c_u}, r_{c_v}]^1$ on sequence (1) are calculated according to the maximum of criterion (2). Similarly, the boundaries for $[r_{c_u}, r_{c_v}]^p$, $p > 1$ are determined on values (1) that are not included in $[r_{c_u}, r_{c_v}]^1, \dots, [r_{c_u}, r_{c_v}]^{p-1}$. The condition for stopping the procedure is to cover all values (1) with disjoint intervals.

One of the options for using the division of feature values into intervals according to (2) is the grouping of objects. The transformation of features into nominal gradations makes it easier to search for similar objects in the training set $\Omega_i(K_1, K_2)$. Otherwise, in the raw feature space, it was necessary to look for a method of normalization, a measure of proximity between objects, set grouping criteria and prove the stability of the solution for them.

Let $\eta_{1i} = \frac{d_1^i(u, v)}{|K_1|}$, $\eta_{2i} = \frac{d_2^i(u, v)}{|K_2|}$ denote the results of optimal partitioning by (2) for each interval $[r_{c_u}, r_{c_v}]^i$, $i = \overline{1, \tau_c}$. The value of the membership function of feature $x_c \in X(n)$ to K_1 over the interval $[r_{c_u}, r_{c_v}]^i$ is defined as

$$f_{ci} = \frac{\eta_{1i}}{\eta_{1i} + \eta_{2i}}. \tag{3}$$

If the feature is $c \in J$, then η_{1i}, η_{2i} in (3) can be considered as the number of values of the i -th gradation, respectively, in classes K_1, K_2 . The generalized estimate is calculated using stochastic algorithms

$$g_c = \frac{1}{m} \sum_{\{[r_{c_u}, r_{c_v}]^i\}} \begin{cases} f_{ci}(v - u + 1), & f_{ci} > 0.5, \\ (1 - f_{ci})(v - u + 1), & f_{ci} < 0.5 \end{cases} \tag{4}$$

Stability values (4) are necessary for the conversion of the original raw features into binary values. Experts can use the division into intervals according to (2) when forming linguistic rules for knowledge bases. The number of class dominance intervals indirectly indicates the status of patterns [22], [23]. The smaller the intervals of dominance, the stronger the manifestation of regularity on a specific feature in the class. This property can be used when ranking quantitative indicators in applied tasks. The highest ranks are obtained by those indicators, the number of intervals of dominance

of classes K_1, K_2 of which is minimal [24]. An additional alternative for ranking by K_2 with an equal number of intervals is an indicator that expresses the degree of uniformity (non-displacement) of the values of the c -th feature of objects within the boundaries of the dominance intervals determined by (2).

C. Machine learning algorithms for classification

SVM is widely used to transfer different types of data by including a nuclear function for mapping into the data space [25]. Such functions are most often used as a linear kernel, a polynomial kernel, a Gaussian kernel with a radial base function, and a sigmoid kernel.

The purpose of SVM is to classify data points of an X_n -dimensional space using an $(n - 1)$ - dimensional hyperplane [26]. Any hyperplane can be written as a set of points X satisfying $w^T x + b = 0$, where the vector w is a normal vector perpendicular to the hyperplane and b is the displacement of the hyperplane $w^T x + b = 0$ from the original point along the direction w .

The distance from the data point to the separating hyperplane $w^T x + b = 0$ can be calculated as $r = (w^T x + b)/|w|$, and the data points closest to the hyperplane are called reference vectors.

Linear SVM is solved by formulating the quadratic optimization problem as follows:

$$\begin{aligned} \arg \min_{w,b} (\frac{1}{2}|w|^2), \\ y(w^T x + b) \geq 1 \end{aligned} \tag{5}$$

Random Forest (RF) is a machine learning algorithm that combines two ideas: a decision tree and ensemble learning [25], [26]. The forest contains many decision trees that use randomly selected data attributes as input. The forest has a collection of controlled dispersion trees. Finally, the result of the classification can be determined by a majority vote or a weighted vote. One of the advantages of random forest is that the variance of the model decreases as the number of trees in the forest increases, while the bias remains the same. In addition, random forests have many other advantages, such as a small number of parameters and resilience to overfitting.

The decision tree (DT) is a tree-like structural model that has leaves that represent classes or solutions, and branches that represent conjunctions of features that lead to these classifications [26]. Tree classification of the input vector is performed by traversing the tree, starting from the root node and ending with a leaf [25]. Each tree node computes an inequality based on one of the input variables. Each sheet is assigned to a specific class. Each inequality that is used to partition the input space is based on only one of the input variables. Linear DT is similar to binary DT, except that the inequality calculated at each node has an arbitrary linear form, which may depend on several variables. The DT depends on the "if-then" rule, but does not require any parameters and metrics. This simple and interpretable structure allows decision trees to consider attribute problems of various types [25], [27].

The Decision Tree can also manage missing values or noisy data. However, they cannot guarantee optimal accuracy unlike other machine learning methods. Although Decision Trees are easy to learn and implement, they don't seem to be popular intrusion detection methods. A possible reason for the lack of popularity is that finding the smallest Decision Tree is an NP-hard task.

Algorithm of k-nearest neighbors. The construction of a classification model based on the k-nearest neighbor algorithm consists in memorizing a training sample of data [25], [28]. To make a prediction for a new data instance, the algorithm searches for the nearest point of the training sample to it, thereby finding the nearest neighbors. The new instance is assigned a label belonging to the nearest point of the training set. The algorithm allows you to take into account not only one nearest neighbor, but also an arbitrary number of them (k). Let's give a training sample with pairs of the form "object-response":

$$x^m - \{(x_1, y_1), \dots, (x_m, y_m)\} \tag{6}$$

Let the distance function $\rho(x, x')$ be given on the set of objects, which should be a fairly adequate model of similarity of objects, that is, the greater the value of this function, the less similar are the objects x, x' . For an arbitrary u , arrange the training sample objects x_i in order of increasing distance to u :

$$p(u, x_{1,u}) \leq p(u, x_{2,u}) \leq \dots \leq p(u, x_{m,u}) \tag{7}$$

where $x_{i,m}$ is the training sample object that is the i th neighbor of object u . Similarly, we introduce the notation for $y_{i,u}$ to answer the i th neighbor. The nearest neighbor algorithm in its most general form looks like this:

$$a(u) = \arg \max_{i=1}^m [x_{i,m} = y]w(i, u) \tag{8}$$

where $w(i, u)$ is a given weight function that estimates the degree of importance of the i -th neighbor in classifying the object u [122]. This function must be non-negative and must not increase in i -th.

The k-nearest neighbor algorithm has several advantages, including ease of interpretation of the model, satisfactory prediction quality that can be obtained without using a large number of settings [28]. Moreover, usually this algorithm allows you to build a classification model very quickly. However, if there is a large training set, the model needs additional time to learn. It is also worth noting that the k -nearest neighbors algorithm is inefficient for working with sparse datasets that have zero values, as well as with datasets where there are many features for model estimation. The k -nearest neighbors algorithm is not so often used in practice due to the fact that it has a relatively low computational speed and cannot process a large number of features.

Naive Bayes is a simple Bayesian network model that assumes that all variables are independent [27], [28]. Using

the Bayes rule for NB classification, it is necessary to find the maximum likelihood hypothesis that determines the class label for the test data x . Given observed data x and a group of labels of class $C = c_j$, the naive Bayes classifier can be solved by the maximum a posteriori probability (MAP) hypothesis as follows:

$$\arg \max_{c_j \in C} P(x|c_j)P(c_j) \tag{9}$$

NB is effective for inference problems and is based on the assumption of independent variables.

D. Metrics for evaluating classification methods

To evaluate the performance of classifiers, the following metrics are used: accuracy of classification, completeness, accuracy and F-measure. Four classification cases are possible for any classification algorithm, and this helps to understand the difference between the metrics under consideration: true-positive results (True Positives, TP), false-positive results (False Positives, FP), true-negative results and false-negative results (False Negatives, FN).

Accuracy of classification can be defined as the proportion of correct results that is achieved by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

The accuracy shows what proportion of objects identified by the classifier as positive is really positive.

$$\text{precision} = \frac{TP}{TP + FP} \tag{11}$$

Completeness shows which part of the positive objects was allocated by the

$$\text{recall} = \frac{TP}{TP + FN} \tag{12}$$

classifier.

The F-score is a metric that combines measures of precision and recall:

$$F1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{13}$$

III. COMPUTATIONAL EXPERIMENT

For the computational experiment, the CICDDoS 2019 data set was used [29], the presentation form of which is given in Table.2. For each type of DDOS, a selection is made of features whose stability value by (4) is greater than 0.90. The results of such selection are given in Table 3.

TABLE III. Values of compactness measure by types of DDOS attacks.

	Type of DDOS attacks	Number of objects
1	DNS	11
2	LDAP	26
3	MSSQL	23
4	NetBIOS	20
5	SNMP	22
<i>continued on next page</i>		

<i>continued from previous page</i>		
	Type of DDOS attacks	Number of
6	UDP-Lag	4
7	WebDDoS	9
8	SYN	13
9	NTP	19
10	SSDP	26
11	UDP	26
12	TFTP	26

Based on the results of feature selection (see Table 3), the following conclusions can be drawn. The worst in terms of stability (4) are the 3 types of DDOS attacks UDP-Lag, WebDDoS and DNS. The first 20 signs from the descending ordered (average by types of attacks) stability values of the sequence are given in Table 4.

TABLE IV. Values of compactness measure by types of DDOS attacks.

	Feature name	Average value
1	Fwd Packet Length Mean	0,946463889
2	Max Packet Length	0,946463889
3	Avg Fwd Segment Size	0,9427
4	Average Packet Size	0,941544389
5	Packet Length Mean	0,941252833
6	Packet Length Mean	0,931808306
7	Fwd Packet Length Min	0,927713917
8	Fwd Packet Length Max	0,925263917
9	Bwd Packets/s	0,924822222
10	Flow IAT Std	0,922172167
11	Flow IAT Max	0,921088861
12	Flow Packets/s	0,920752833
13	Flow Duration	0,920708333
14	Flow IAT Mean	0,917563861
15	Total Length of Fwd Packets	0,915383361
16	Fwd Packets/s	0,910708278
17	Subflow Fwd Bytes	0,910708278
18	Inbound	0,899558333
19	Flow Bytes/s	0,897586167
20	Total Backward Packets	0,895905583

The stability values (4) are proposed to be used for ordering (ranking) features and forming a new data sample. The purpose of the formation is to study the structure of links of 12 types of DDOS attacks. An important role in the experiment is played by the number of options from combinations of features. Table 5 shows a comparative analysis of the considered approaches to detecting DDOS attacks.

TABLE V. Values of compactness measure by types of DDOS attacks.

	Type of DDOS attacks	Number of objects
1	DNS	11
2	LDAP	26
3	MSSQL	23
<i>continued on next page</i>		

<i>continued from previous page</i>		
	Type of DDOS attacks	Number of
4	NetBIOS	20
5	SNMP	22
6	UDP-Lag	4
7	WebDDoS	9
8	SYN	13
9	NTP	19
10	SSDP	26
11	UDP	26
12	TFTP	26

Random Forest and KNN proved to be the best machine learning model for detecting DDOS attacks. When using feature selection methods, the Random Forest classifier showed the greatest efficiency. We managed to reduce the attribute from 87 to 20 without compromising the accuracy of the classifier. Feature selection improves the performance and accuracy of the classifier.

IV. CONCLUSION

A methodology for analyzing the structure of the relationship of DDOS attacks types using the values of the stability of the signs for pairs of classes normal traffic and type of DDOS attacks” has been developed. Sets of features for calculation were formed based on the results of their ranking by stability. The analysis technique is used to justify and construct classifiers of DDOS attacks types, taking into account their connectivity.

In the course of the work, the existing algorithms for classifying data based on machine learning were considered, and the possibility of their application for detecting network distributed DDOS attacks was evaluated. A comparative analysis of the considered approaches to the detection of DDOS attacks is given.

REFERENCES

[1] S. Rezaei and X. Liu, Deep learning for encrypted traf classification: An overview, IEEE Communications Magazine, vol. 57, pp. 7681, 2019. [Online]. Available: <https://doi.org/10.1109/MCOM.2019.1800819>

[2] A. Finamore, M. Mellia, M. Meo, and D. Rossi, Kiss: Stochastic packet inspection classier for udp trafic, IEEE/ACM Transactions on Networking, vol. 18, pp. 15051515, 2010. [Online]. Available: <https://doi.org/10.1109/TNET.2010.2044046>

[3] L. Vu, C. Bui, Q. Nguyen, and D. Rossi, A deep learning based method for handling imbalanced problem in network trafic classification. December 2017, pp. 333339. [Online]. Available: <https://doi.org/10.1145/3155133.3155175>

[4] G. Aceto, D. Ciunzo, A. Montieri, and P. A, Multi-classification approaches for classifying mobile app trafic, Journal of Network and Computer Applications, vol. 57, pp. 131145, 2018. [Online]. Available: <https://doi.org/10.1016/j.jnca.2017.11.007>

[5] P. Wang, C. Xuejiao, Y. Feng, and S. Zhixin, A survey of techniques for mobile service encrypted traf classification using deep learning, IEEE Access, vol. 7, pp. 5402454033, 2019. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2912896>

[6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traf characterization, 2018, pp. 108116. [Online]. Available: <https://doi.org/10.5220/0006639801080116>

[7] P. N. Matheus, F. C. Luiz, L. Jaime, and L. P. Mario, Long shortterm memory and fuzzy logic for anomaly detection and mitigation in software-dened network environment, 2020, pp. 8376583781. [Online]. Available: <https://doi.org/10.1109/ACCESS.2020.2992044>

[8] B. Naveen and S. Manu, Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting ddos attacks, Romanian journal of information science and technology, vol. 23, no. 3, p. 250 261, 2020.

[9] S.E.Mahmoud,L.Nhien-An,D.Soumyabrata,andD.J.Anca,Ddosnet: A deep-learning model for detecting network attacks, in 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). IEEE, 31 Aug.-3 Sept. 2020, pp. 18. [Online]. Available: <https://doi.org/10.1109/WoWMoM49955.2020.00072>

[10] M. S. Yin, P. A. Pye, and S. H. Aye, A slow ddos attack detection mechanism using feature weighing and ranking, Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore, pp. 45004509, March. 7-11, 2021.

[11] A. H. Lashkari, D. G. Gerard, M. M. Mamun, and A. A. Ghorbani, Characterization of tor traf using time based features, 2017, pp. 253262. [Online]. Available: <https://doi.org/10.5220/0006105602530262>

[12] N. Miloslavskaya, A. Tolstoy, and S. Zapechnikov, Taxonomy for unsecure big data processing in security operations centers, Aug.2224 2016, pp. 154159. [Online]. Available: <https://doi.org/10.1109/W-FiCloud.2016.42>

[13] N. Miloslavskaya and A. Makhmudova, Survey of big data information security, vol. 8, Aug.22-24 2016, pp. 133138. [Online]. Available: <https://doi.org/10.1109/W-FiCloud.2016.38>

[14] N. A. Ignatiev, On nonlinear transformations of features based on the functions of objects belonging to classes, Pattern Recognition and Image Analysis, vol. 2, no. 31, pp. 197204, June 30 2021. [Online]. Available: <http://dx.doi.org/10.1134/S1054661821020085>

[15] N. A. Ignatyev and M. A. Rakhimova, Formation and analysis of sets of informative features of objects by pairs of classes, Artificial intelligence and decision making, no. 4, pp. 18 26, 2021. [Online]. Available: <http://dx.doi.org/10.14357/20718594210402>

[16] N. Miloslavskaya, A. Tolstoy, V. Budzko, and D. Maniklal, Blockchain application for iot cybersecurity management, pp. 141168, 2019. [Online]. Available: <https://doi.org/10.1201/9780429674457-7>

[17] N. G. Zagoruiko, I. A. Borisova, and O. A. Kutnenko, Constructing a concise description of data using the competitive similarity function, Siberian Journal of Industrial Mathematics, vol.1, no.16, pp.2941,2013.

[18] S. F. Madrakhimov, K. T. Makharov, and M. Y. Lolayev, Data preprocessing on input, AIP Conference Proceedings, vol. 1, no. 16, pp. 2941, 2021. [Online]. Available: <https://doi.org/10.1063/5.0058132>

[19] N. A. Ignatiev, Structure choice for relations between objects in metric classification algorithms, Pattern Recognition and Image Analysis, vol. 28, no. 4, pp. 695702, 2018. [Online]. Available: <https://doi.org/10.1134/S1054661818040132>

[20] Ddos evaluation dataset (cic-ddos2019), 2018. [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>

[21] I. Sharafaldin, A. H. Lashkari, H. Saqib, and A. Ghorban, Developing realistic distributed denial of service (ddos) attack dataset and taxonomy, in In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCST). IEEE, Oct. 1-3, pp. 18. [Online]. Available: <https://doi.org/10.1109/CCST.2019.8888419>

[22] A. Kabulov, I. Normatov, E. Urunbaev, F. Muhammadiev, "Invariant continuation of discrete multi-valued functions and their implementation," 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.

[23] A. Kabulov, I. Saymanov, "Application of IoT technology in ecology (on the example of the Aral Sea region)," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.

[24] A. Kabulov, I. Saymanov, I. Yarashov, F. Muxammadiev, "Algorithmic method of security of the Internet of Things based on steganographic coding," 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.

[25] A. Kabulov, M. Berdimurodov, "Parametric Algorithm for Searching the Minimum Lower Unity of Monotone Boolean Functions in the Process Synthesis of Control Automates," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-5.

- [26] A. Kabulov, M. Berdimurodov, "Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal)," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [27] A. Kabulov, I. Saymanov, M. Berdimurodov, "Minimum logical representation of microcommands of cryptographic algorithms (AES)," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [28] A. Kabulov, I. Normatov, I. Kalandarov, I. Yarashov, "Development of An Algorithmic Model and Methods for Managing Production Systems Based on Algebra over Functioning Tables," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [29] A. Kabulov, I. Kalandarov, I. Yarashov, "Problems of Algorithmization of Control of Complex Systems Based on Functioning Tables in Dynamic Control Systems," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-4.
- [30] A. Kabulov, E. Urunboev, I. Saymanov, "Object recognition method based on logical correcting functions," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2020, pp. 1-4.
- [31] A. Kabulov, A. Babadzhanov, I. Saymanov, "Completeness of the linear closure of the voting model," AIP Conference Proceedings, 2022 (accepted).
- [32] A. Kabulov, A. Babadzhanov, I. Saymanov, "Correct models of families of algorithms for calculating estimates," AIP Conference Proceedings, 2022 (accepted).
- [33] A. Kabulov, I. Normatov, S. Boltaev, I. Saymanov, "Logic method of classification of objects with non-joining classes," Advances in Mathematics: Scientific Journal, 2020, 9(10), p. 8635–8646.
- [34] A. Kabulov, I. Kalandarov, I. Saymanov, "Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules," Smart transport conference 2022 Conference, pp. 1-11.
- [35] A. Kabulov, "Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions," Cybernetics, 1979, 15(3), p. 305–311.
- [36] A. Kabulov, G. Losef, "Local algorithms simplifying the disjunctive normal forms of Boolean functions," USSR Computational Mathematics and Mathematical Physics, 1978, 18(3), p. 201–207.
- [37] A. Kabulov, "Local algorithms on yablonskii schemes," USSR Computational Mathematics and Mathematical Physics, 1977, 17(1), p. 210–220.
- [38] H. Khujamatov, I. Siddikov, E. Reypnazarov, D. Khasanov. Research of Probability-Time Characteristics of the Wireless Sensor Networks for Remote Monitoring Systems // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [39] I. Siddikov, D. Khasanov, H. Khujamatov, E. Reypnazarov. Communication Architecture of Solar Energy Monitoring Systems for Telecommunication Objects // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [40] I. Siddikov, K. Khujamatov, E. Reypnazarov, D. Khasanov. CRN and 5G based IoT: Applications, Challenges and Opportunities // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [41] K. Khujamatov, A. Lazarev, N. Akhmedov, E. Reypnazarov, A. Bekturdiyev. Methods for Automatic Identification of Vehicles in the its System // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [42] S. Tanwar, H. Khujamatov, B. Turumbetov, E. Reypnazarov, Z. Allamuratova. Designing and Calculating Bandwidth of the LTE Network for Rural Areas // International Journal on Advanced Science, Engineering and Information Technology, 2022, 12(2), pp. 437-445.
- [43] H. Zaynidinov, D. Singh, S. Makhmudjanov, I. Yusupov. Methods for Determining the Optimal Sampling Step of Signals in the Process of Device and Computer Integration // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 471–482.
- [44] H. Zaynidinov, D. Singh, I. Yusupov, S. Makhmudjanov. Algorithms and Service for Digital Processing of Two-Dimensional Geophysical Fields Using Octave Method // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 460–470.
- [45] H. Zaynidinov, S. Anarova, J. Jabbarov. Determination of Dimensions of Complex Geometric Objects with Fractal Structure // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 437–448.

Suicide Rate and Factors Analysis: Pre and Post COVID Pandemic

Helen Antonova
hantonov.gmu.edu

Spencer Liao
slia6@gmu.edu

Eswara Chandra Sai Pamidimukkala
epanidim@gmu.edu

Ebrima Ceesay
eceedsay@gmu.edu

Abstract—According to the 2021 Report from the World Health Organization (WHO), more than 700,000 people have taken their life. Suicide can be prevented but so far most of the efforts to do so have fallen short. However, the use of machine learning and artificial intelligence offers new opportunities to increase the accuracy level of prediction and aid the goal of suicide prevention. This paper reviews literature concerning the machine learning methods used to help identify various risk factors and help prevent suicide. This paper also presents our research and analysis findings which were used to identify various suicide risk factors and additional analysis of whether there are any correlations or variations in the risk factors from pre- and post-pandemic datasets regarding suicide rates. This is especially important during times of high stress, such as a worldwide pandemic and quarantine. The dataset(s) obtained from WHO suggest that high levels of risk factor identification are possible and This paper and the analysis serve as supporting research and guide to aid in the continued ambitious goal of suicide prevention worldwide

Index Terms—Suicide Prevention, Python, R, Post-COVID, Quarantine, Mental Health, Volume Analysis, Algorithm, Suicide, Machine Learning, Risk Factors, AI

I. INTRODUCTION

Suicide is a complex global public health problem. According to the World Health Organization (WHO), every year, more than 700,000 people take their own life and there are many more who attempt to do so. Only 38 countries report having a national strategy for suicide prevention. Suicide and suicide attempts have a devastating ripple effect that impacts the community, families, friends, coworkers, and societies. The COVID pandemic has created a new sense of urgency to analyze and engage all countries and communities in aiding the prevention and finding an achievable and sustainable solution

to the growing problem. The quarantine lockdowns and restrictions over the past couple of years, may have led to a heightened sense of loneliness, anxiety, fear, and depression in many people across the globe. Suicides are preventable and a lot can be done to improve the international strategy efforts for suicide prevention. Our mission is to identify the key drivers of suicide, without underscoring that suicide is determined by multiple factors.

Early identification of individuals who are at higher risk of suicide is vital in preventing suicide. According to multiple studies, machine learning and the advantage of Big Data are becoming the new promising approach in this “early identification” objective. But unlike weather forecasting, traffic, or medical predictions, suicide prediction is uniquely different. Especially when surveying and analyzing data on a global scale and considering additional life stressor factors such as the COVID-19 crisis, prolonged quarantine periods, or the resulting strains on the economy and the job market. The topic of predictive analytics being used to possibly identify suicide risk factors has been controversial. Some experts believe the risks should be identified by in-depth clinical evaluations and not risk assessment tools or data models. Making the argument that the risk of misidentification of suicide risk can be associated with inadvertent negative consequences. Still, the power of big data and technology should not be overlooked and, instead, be optimized to work in conjunction with clinical evaluation methods in addressing the suicide problem.

When understanding the motive behind suicide there are many factors that must be noted:

- Geographic Location
- Mental Health Conditions

- Life Stressors
- Relationship Problems
- Financial Problems
- Cultural/Societal Expectations
- Current Events

II. BACKGROUND

Death by suicide is a very complex issue that causes torment and distress for hundreds of thousands of people globally. But with timely intervention and prediction, suicides can be prevented. This is especially important for times of high stress, economic uncertainty, pandemic or natural disasters, and other stress factors that can have a serious effect on mental health and morale.

The WHO compiles and disseminates data on death and morbidity annually. This data is reported by all participating “Member States”, according to the WHO mandates. These Member States are foreign countries, which at first started out as only 11 back in 1950, right after the inception of WHO, then growing to 74 in the year 1985, and reaching 183 in 2019. Thus, this is one of the most complete and thorough mortality data banks that are publicly available, and reported suicide deaths are an integral part of this data bank. However, although this global data bank is from a reputable organization and with specific data classification requirements, it still offers some challenges. It’s very difficult to get a high quality and high accuracy reading of suicide rates around the world. Only about 80 Member States provide good-quality, reliable data that can be used to estimate the risk factors associated with suicide deaths. “This problem of poor-quality mortality data is not unique to suicide but given the sensitivity of suicide – and the illegality of suicidal behavior in some countries – it is likely that under-reporting and misclassification are greater problems for suicide than for most other causes of death.” [1]. In addition, the official statistics relating to attempted suicide are even more skewed, which makes it difficult to analyze and correlate national global suicide trends to suicide attempt trends. “...suicide may be hidden and underreported for several reasons, e.g. as a result of prevailing social or religious attitudes. In some places, it is believed that suicide is underreported by a percentage between 20% and 100%.” [1].

In spite of these dataset collection and classification challenges, the WHO mortality dataset is one of the most comprehensive and reliable data banks in the world. An important role of big data gained from sources such as the World Health Organization (WHO), Electronic Health Records (EHR) [2], or social media regarding suicide, is supporting the objective of prevention. “The potential in this area is tremendous. The suicidal phenotype is characterized by extreme heterogeneity, and potentially suicidal individuals are very often excluded from any clinical trials. Big data could help by combining very complex and large data samples to detect patterns, signaling suicidal inclinations [3] With such a powerful tool as big data available, research teams and analysts should be able to predict high risk factors with a high degree of certainty. So far there have been cases where scientists were successful in implementing tools to survey and predict risk, with 91% accuracy [4], but unfortunately this level of accuracy is only due to the smaller and more controlled groups, such as adolescents in the state of Utah. The challenge still exists, not only for a nationwide population for all age groups, but also a worldwide population and especially during high stress time periods caused by economic downturns.

Big data is already being used in the fields of psychology and psychiatry for a variety of reasons and goals. But the velocity of data acquisition is one of the main challenges because the majority of the data is updated on a periodic basis, and there aren’t many options available for real-time data. “It should also be recognized that as big data gains a more prominent role in psychiatry, issues of governance and security will need to be clearly considered, and that there must be a thorough and open public dialogue on ethical issues.” [3] Even still, big data in the world of psychiatry can offer significant benefits to help with not only treatment of the patients but also comprehension of their disorders and early diagnosis and prevention of suicide or self-harm. Although, it’s important to keep in mind that machine learning, predictive algorithms, and big data needs to also run hand in hand with additional efforts and factors to be fully effective in identifying risk factors and preventing suicide. Medical and psychiatric professionals need

to recognize the potential of big data technological advancements, while at the same time also being careful to maintain a high standard of professionalism, as well as the traditional doctor-patient relationship which is based on trust and confidentiality.

III. RELATED WORK

As mentioned in the introduction as well as in the background, there are various reasons, risk factors, and warning signs that lead to suicide and suicidal thoughts. Many studies have been conducted over the years that focus on the various factors in an attempt to better understand and hopefully lower suicidal rates.

In this paper, we reviewed six literatures that were published between 2012 and 2020. Of the two that were published pre COVID-19, each focused on suicide rates in the USA during economic recession during a different period. The third pre COVID-19 paper discussed the national cost of suicides and suicide attempts in the United States in 2013. The remaining three literatures observed trends and discussed whether the COVID-19 pandemic resulted in suicide rate increases.

In the report titled, "Increase in state suicide rates in the USA during economic recession" by Reeves et al., it mentions evidence from European countries that show a rise in suicides during economic recessions. Among the worst being Greece, where suicides have risen more than 60% since 2007. Using data on suicide mortality rates from 1999 to 2010 from the Centers for Disease Control and Prevention, along with 'unemployment' data from the Bureau of Labor Statistics, the authors extended their previous analyses of recessions and suicides in Europe to assess trends in all 50 US states. Their findings showed that in the years before the recession (from 1999 to 2007), the suicide mortality rate in the USA were rising on average at a rate of 0.12 per 100K per year. During the recession period (2008-2010), the suicide rate increased at a rate of 0.51 deaths per 100K per year. This difference in rate corresponds to an additional 1580 suicides per year [5].

Another study by Harper et al. explores suicide mortality rate in the USA over a 30 year period. Prior to this study, there were several studies that suggested strong associations between eco-

nomie downturns and suicide mortality. This study aimed to provide more robust evidence by using a quasi-experimental design. The researchers analyzed 955K suicides that occurred in the USA from 1980 to 2010 and used a broad index of economic activity in each US state to measure economic conditions. Based on the quasi-experimental and fixed-effects design, and after accounting for secular trends, seasonality, and unmeasured fixed characteristics of states, they found that an economic downturn in magnitude to the 2007 Great Recession increased suicide mortality by 0.14 deaths per 100K population or around 350 deaths. The effects were also stronger for men than women and for those with less than 12 years of education [6].

The third paper reviewed addresses economic costs of suicides in the United States in 2013. Understanding that suicide would not be eliminated from our society anytime, we felt it was important to review some literature on its economic cost to better put things in perspective. According to the authors, there were three previous studies using different approaches to address the matter. For their study, to seek to improve on the previous by addressing the increase in the number of suicides since those publications by incorporating adjustments for underreporting and using additional data. Shepard et al. paper concluded that the national cost of suicides and suicide attempts in the United States in 2013 was \$58.4 billion based on reported numbers. Lost productivity represented most (97.1%) of the cost. When adjusted for under-reporting, it increased the total cost to \$93.5 billion [7].

Similar to understanding how economic downturns effected suicide mortality rates in the first two reviewed literature, we reviewed a few more literatures that were published during the COVID-19 pandemic to observe how and whether the pandemic affected the rate.

In the paper titled "Suicide risk and prevention during the COVID-19 pandemic", Gunnell et al. seek to understand how suicide is likely to become a more pressing concern as the pandemic spreads and its effects on the general population, the economy, and vulnerable groups. Suicide risk might increase among people with mental illness due to self-isolation, exacerbated fear, and physical distancing. Those with psychiatric disorders might

experience worsening symptoms, while others in stressful work environments (such as health care) might develop new mental health problems. During these extraordinary moments, people in suicidal crises need special attention. Not all will seek help in fear that services are overwhelmed or not willing to attend face-to-face appointments [8].

Deterioration in population mental was one of several factors that underpinned the concern that suicide rates may increase during the COVID-19 pandemic. Based on some widely reported studies modelling the effects of the pandemic on suicide rates predicted increases ranging from 1% to 145% [9]. In the study titled “Trends in suicide during covid-19 pandemic”, the authors tracked and reviewed relevant studies for a living systematic review. In the early months of the pandemic, reports suggest either no rise in suicide rates (Massachusetts, USA; Victoria, Australia; England) or a fall (Japan, Norway) in high income countries. Not much is known for low income countries.

IV. PROBLEM DESCRIPTION

The main purpose of this analysis is to identify risk factors in the hopes of mitigating suicide rates and help combat the worldwide phenomena. Identifying these drivers could help communities implement proactive protective practices and decrease the rate of suicide.

We are approaching this problem by identifying which factors effects the trigger of suicide in each individual like their country GDP or their Income or which place they are from and so on. We will be cleaning the data using R/Python and use scatter plot and ggplot packages to form comparison between the data from 1986 to 2016, and data from 2019 to current.

Questions we aim to answer:

- Is a prior suicide attempt a clear risk factor and indicator of another attempt?
- Is there a correlation between a county’s GDP and suicide rate?
- Is there a correlation between gender and suicide rates?
- Is there a correlation between the year and suicide events? If so, is this due to any global events at the time?

- Do the countries that invest more in mental health see a lower suicide rate?
- Do countries that are more impacted by the opioid epidemic see a higher rate of suicide?
- On average how many combined factors does a suicide victim have?
- Do the countries that have easier access to suicide methods (substances, firearms, etc.) have a higher suicide rate?

V. DATA, TOOLS, TECHNIQUES AND APPROACHES

Methodology:

The team began by cleaning and formatting the suicide datasets using Excel. Once the data is in the proper format, utilizing programming languages like R and Python, the data will be analyzed through a series of analytical techniques that will assess how suicide rates have changed from before and during the pandemic. This will include visualizations, charts, and maps showing the patterns seen within the dataset.

How we collected the data:

The team used secondary data gathered from the “1986 - 2016 Suicide Rates” [10] dataset which we obtained from Kaggle. This dataset is a combination of four datasets, linked by time and place to help identify risk factors worldwide. The four datasets were:

- 1) United Nations Development Program. (2018). Human development index (HDI). Retrieved from <http://hdr.undp.org/en/indicators/137506>
- 2) World Bank. (2018). World development indicators: GDP (current US\$) by country:1985 to 2016. Retrieved from <http://databank.worldbank.org/data/source/world-development-indicators>
- 3) Szamil . (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from <https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook>
- 3) World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

Preliminary Results:

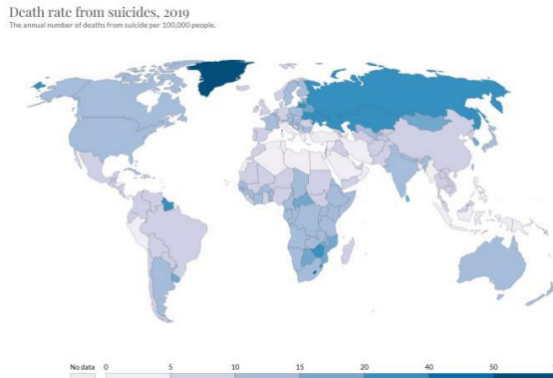


Fig. 1: Death Rate From Suicides

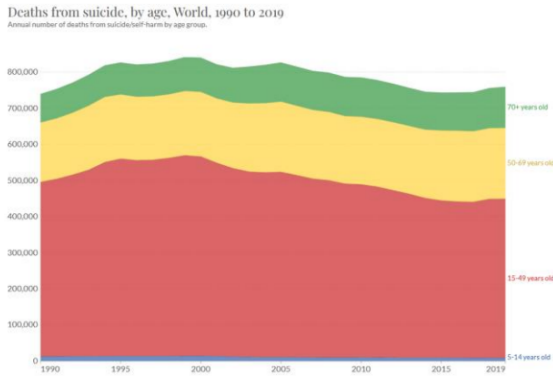


Fig. 2: Death From Suicide, By Age

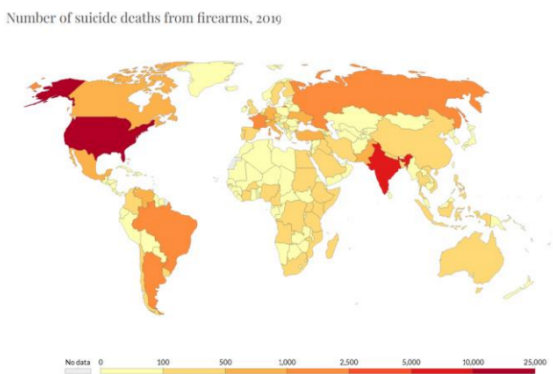


Fig. 3: Number of Death From Firearm

Suicide rate vs. income inequality, 2015

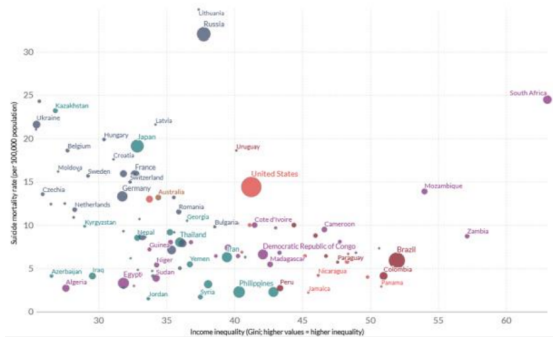


Fig. 4: Suicide Rate vs. Income Inequality

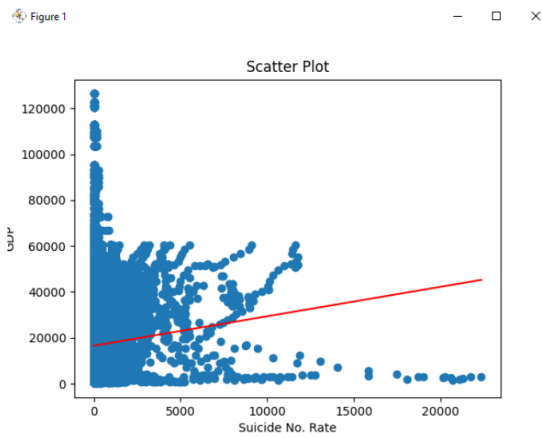


Fig. 5

- Is a prior suicide attempt a clear risk factor and indicator of another attempt?
- Is there a correlation between a county's GDP and suicide rate?
- Is there a correlation between gender and suicide rates?
- Is there a correlation between the year and suicide events? If so, is this due to any global events at the time?
- Do the countries that invest more in mental health see a lower suicide rate?
- Do countries that are more impacted by the opioid epidemic see a higher rate of suicide?
- On average how many combined factors does a suicide victim have?
- Do the countries that have easier access to suicide methods (substances, firearms, etc.)

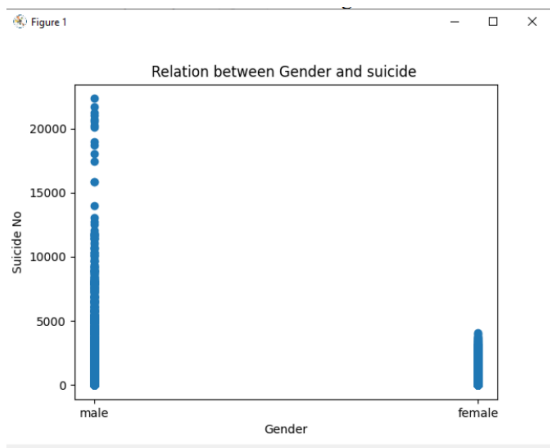


Fig. 6

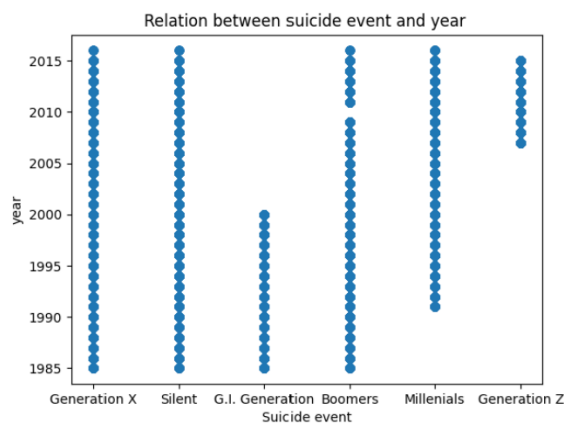


Fig. 7

have a higher suicide rate?

VI. CHALLENGES

Due to limitation of recent and extended periods of worldwide suicidal data from credible sources like WHO or CDC (who only publish suicide figures up to 2019), it is difficult to make clear comparisons of pre and post-pandemic suicidal rates and factors. From our review of recent articles and publication, most do not draw a conclusion that suicidal rates increased during the pandemic, but state that previous pandemics have been associated with increases in suicide rate. The U.S. reported an increase in suicides during the Spanish Flu (1918-19) [11]. Africa did experience an increase in suicides during the Ebola epidemic spanning

2013 to 2016 [12]. Hong Kong did observe an increase in elderly suicides during 2003 SARS outbreak [13]. One common denominator among these publications is that they were all completed years after the epidemic/outbreak or after multiple years of data was available for analysis. There are multiple factors that lead researchers to believe that the current COVID-19 pandemic will share similarities with past events. By the end of 2020, over 76 million people worldwide were infected by the COVID virus, SARS-CoV-2 [14]. Shutdown of businesses and business activities because of lockdown measures affected those living in poverty, relied on hourly wage positions, and many that relied on government programs. As mentioned in the literature review that examined US suicide rates during 2008 as well as a publication by Oyesanya et al. [15], both conclude that economic stress has been associated with higher suicide rates. Another effect of the COVID lockdown is social isolation. Studies such as the one by Christensen et al. have documented that social isolation is associated with increased suicidal thoughts. A 2016 study by VanderWeele et al. show that participating in religious communities is associated with lower suicide rates, however, with churches and community centers closed during lockdowns, social isolation possibly increased suicidal thoughts. One factor that differs greatly from the current pandemic to previous epidemic/outbreaks is the amount of media coverage. With 24-7 news coverage, social media outlets, unlimited SMS service, etc., it's possible that anyone with preexisting mental health conditions will experience intensified anxiety and fear. To make matters worse, during lockdowns, increased restrictions at healthcare facilities were another barrier to those that required mental health treatment.

VII. CONCLUSION

Similar to the report by John et al. that examined suicide rates during the earlier stage of COVID (lockdown, stay-at home period), other reports by Faust et al., Applyby et al., and Qin et al., all [16], [17], concluded that suicide rate for the same period compared to a year ago were similar or did not show significant difference from previous years. Despite these similar conclusions, we find that it

is too early for a study to compare pre and post pandemic suicidal rates due to the lack of data and that the pandemic is still ongoing.

VIII. FUTURE WORK

For a follow on study to be more conclusive, it is best to be done after WHO declares the end of the pandemic and when most countries' citizens have returned to a lifestyle similar to pre pandemic. In the follow on study, instead of focusing on worldwide figures, we suggest to examine certain groups that may be more vulnerable to the effects of the pandemic and experienced increased suicide rates.

- Unemployed: In early 2020, the International Labour Organization (ILO) predicted the pandemic cost 25 million jobs worldwide. Studies of the Great Recession in the early 2000s found an increase of suicide risk by 20-30% between 2000 and 2011 with a peak during 2008 [18].
- Mentally ill: individuals with preexisting mental health conditions were likely affected by interruption in treatment, and experienced increased isolation and intensified anxiety and fear due to the pandemic [19].
- Healthcare workers: In the early stage of the pandemic, medical staff have reported increased hopelessness, guilt, and insomnia. All which can increase the risk for suicide [20].
- Racial minorities: Racial minorities who owned small businesses or worked at hourly waged positions were affected more by lockdowns and shutdown of business activities.
- Youth: Preliminary data from England suggest that child suicide deaths may have increased during the early stages of lockdown, possibly due to disruptions to education, outside activities, and support services [21].
- Elderly: Similar to youth, elderly suffered greatly from social disconnectedness during the pandemic. Social and self-isolation affects elderly who do not have close family and friends, or who have decreased literacy in or access to digital resources [22].

other citations [23]–[28]

REFERENCES

- [1] J. M. Bertolote and A. Fleischmann, "Suicide and psychiatric diagnosis: a worldwide perspective," *World psychiatry*, vol. 1, no. 3, p. 181, 2002. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489848/>
- [2] J. Kent. Machine learning uses ehr data to predict suicide attempt risk. (Retrieved March 23, 2022). [Online]. Available: <https://healthitanalytics.com/news/machine-learning-uses-ehr-data-to-predict-suicide-attempt-risk>
- [3] N. Davies. Big data, big psychiatry — big potential? (October 4, 2016). [Online]. Available: <https://www.psychiatryadvisor.com/home/opinion/big-data-big-psychiatry-big-potential/>
- [4] E. McNemar. Machine learning uses predictive analytics for suicide prevention. (November 8, 2021). [Online]. Available: <https://healthitanalytics.com/news/machine-learning-uses-predictive-analytics-for-suicide-prevention>
- [5] A. Reeves, D. Stuckler, M. McKee, D. Gunnell, S.-S. Chang, and S. Basu, "Increase in state suicide rates in the usa during economic recession," *The Lancet*, vol. 380, no. 9856, pp. 1813–1814, 2012. [Online]. Available: [https://doi.org/10.1016/S0140-6736\(12\)61910-2](https://doi.org/10.1016/S0140-6736(12)61910-2)
- [6] S. Harper, T. J. Charters, E. C. Strumpf, S. Galea, and A. Nandi, "Economic downturns and suicide mortality in the usa, 1980–2010: observational study," *International journal of epidemiology*, vol. 44, no. 3, pp. 956–966, 2015. [Online]. Available: <https://doi.org/10.1093/ije/dyv009>
- [7] D. S. Shepard, D. Gurewich, A. K. Lwin, G. A. Reed Jr, and M. M. Silverman, "Suicide and suicidal attempts in the united states: costs and policy implications," *Suicide and Life-Threatening Behavior*, vol. 46, no. 3, pp. 352–362, 2016. [Online]. Available: <https://doi.org/10.1111/sltb.12225>
- [8] D. Gunnell, L. Appleby, E. Arensman, K. Hawton, A. John, N. Kapur, M. Khan, R. C. O'Connor, J. Pirkis, E. D. Caine *et al.*, "Suicide risk and prevention during the covid-19 pandemic," *The Lancet Psychiatry*, vol. 7, no. 6, pp. 468–471, 2020.
- [9] A. John, J. Pirkis, D. Gunnell, L. Appleby, and J. Morrissey, "Trends in suicide during the covid-19 pandemic," 2020. [Online]. Available: <https://doi.org/10.1136/bmj.m4352>
- [10] R. S. Patil. suicide rates from 1986 to 2016. (Jun 24, 2019). [Online]. Available: <https://www.kaggle.com/datasets/rushirdx/suicide-rates-from-1986-to-2016>
- [11] I. M. Wasserman, "The impact of epidemic, war, prohibition and media on suicide: United states, 1910–1920," *Suicide and Life-Threatening Behavior*, vol. 22, no. 2, pp. 240–254, 1992.
- [12] B. K. Y. Bitanirwe, "Monitoring and managing mental health in the wake of ebola," *Annali dell'Istituto superiore di sanita*, vol. 52, no. 3, pp. 320–322, 2016.
- [13] Y. Cheung, P. H. Chau, and P. S. Yip, "A revisit on older adults suicides and severe acute respiratory syndrome (sars) epidemic in hong kong," *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, vol. 23, no. 12, pp. 1231–1238, 2008.
- [14] Johns Hopkins International Injury Research Unit. Responding to the increasing need for global suicide prevention. (September 16, 2020). [Online].

- Available: <https://www.jhsph.edu/research/centers-and-institutes/johns-hopkins-international-injury-research-unit/news/responding-to-the-increasing-need-for-global-suicide-prevention>
- [15] M. Oyesanya, J. Lopez-Morinigo, and R. Dutta, "Systematic review of suicide in economic recession," *World journal of psychiatry*, vol. 5, no. 2, p. 243, 2015.
- [16] L. Appleby, N. Richards, S. Ibrahim, P. Turnbull, C. Rodway, and N. Kapur, "Suicide in england in the covid-19 pandemic: Early observational data from real time surveillance," *The Lancet Regional Health-Europe*, vol. 4, p. 100110, 2021.
- [17] P. Qin and L. Mehlum, "National observation of death by suicide in the first 3 months under covid-19 pandemic," *Acta Psychiatr Scand*, vol. 143, no. 1, pp. 92–93, 2021.
- [18] C. Nordt, I. Warnke, E. Seifritz, and W. Kawohl, "Modelling suicide and unemployment: a longitudinal analysis covering 63 countries, 2000–11," *The Lancet Psychiatry*, vol. 2, no. 3, pp. 239–245, 2015.
- [19] L. E. Egede, K. J. Ruggiero, and B. C. Frueh, "Ensuring mental health access for vulnerable populations in covid era," *Journal of Psychiatric Research*, vol. 129, p. 147, 2020.
- [20] Q. Chen, M. Liang, Y. Li, J. Guo, D. Fei, L. Wang, L. He, C. Sheng, Y. Cai, X. Li *et al.*, "Mental health care for medical staff in china during the covid-19 outbreak," *The Lancet Psychiatry*, vol. 7, no. 4, pp. e15–e16, 2020.
- [21] D. Odd, T. Williams, L. Appleby, D. Gunnell, and K. Luyt, "Child suicide rates during the covid-19 pandemic in england," *Journal of affective disorders reports*, vol. 6, p. 100273, 2021.
- [22] Z. I. Santini, P. E. Jose, E. Y. Cornwell, A. Koyanagi, L. Nielsen, C. Hinrichsen, C. Meilstrup, K. R. Madsen, and V. Koushede, "Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older americans (nshap): a longitudinal mediation analysis," *The Lancet Public Health*, vol. 5, no. 1, pp. e62–e70, 2020.
- [23] S. C. Curtin, H. Hedegaard, and F. B. Ahmad, "Provisional numbers and rates of suicide by month and demographic characteristics: United states, 2020," *NVSS-Vital Statistics Rapid Release*, 2021.
- [24] R. C. Kessler, S. L. Bernecker, R. M. Bossarte, A. R. Luedtke, J. F. McCarthy, M. K. Nock, W. R. Pigeon, M. V. Petukhova, E. Sadikova, T. J. VanderWeele, K. L. Zuromski, and A. M. Zaslavsky, *The Role of Big Data Analytics in Predicting Suicide*. Cham: Springer International Publishing, 2019, pp. 77–98. [Online]. Available: https://doi.org/10.1007/978-3-030-03553-2_5
- [25] B. Resnick. How data scientists are using ai for suicide prevention. (June 8, 2018). [Online]. Available: <https://www.vox.com/science-and-health/2018/6/8/17441452/suicide-prevention-anthony-bourdain-crisis-text-line-data-science>
- [26] M. R. Hannah Ritchie and E. Ortiz-Ospina, "Suicide," *Our World in Data*, 2015, <https://ourworldindata.org/suicide>.
- [27] Suicide rate by country 2022. (Retrieved March 23, 2022). [Online]. Available: <https://worldpopulationreview.com/country-rankings/suicide-rate-by-country>
- [28] World Health Organization (WHO). Suicide in the world. (Retrieved March 23, 2022). [Online]. Available: <https://www.who.int/publications/i/item/suicide-in-the-world>

Employee Turnover Prediction Model for Garments Organizations of Bangladesh Using Machine Learning Technique

Lutfun Nahar
Dept. of Computer Science & Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
Email: lutfacsecu@gmail.com

Zinnia Sultana
Dept. of Computer Science & Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
Email: zinniaiuuc@yahoo.com

Farzana Tasnim
Dept. of Computer Science & Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
Email: farzanatasnim34@gmail.com

Farjana Akter Tuli
Dept. of Computer Science & Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
Email: farjanatuli2022@gmail.com

Abstract—Garments and textile industry is one of the vital organizations in the modern world. In the world market, Bangladesh is the second largest manufacturer and supplier of clothing industry. Bangladesh has been facing many problems in the clothing industry. Employee turnover means to leave manpower from existing organization to another in the sake of better opportunity, better salary and so on. For this reason, organization faces many problems because of employee turnover. So there need a system or model that can predict the employee turnover rate and help the organization to take necessary step to stop employee turnover by providing necessary demands of employees. In this research, several machine learning algorithms are used to predict the turnover of employee. Among them 98% accuracy has been achieved with the Random Forest and an accuracy of 92% has been achieved with the Gradient Boosting classifier.

Keywords—Employee turnover ; Machine Learning; SVM; Logistic Regression; Random forest; Decision tree classifier; Gradient Boosting classifier.

I. INTRODUCTION

Employee turnover is high especially for Garments organizations in Bangladesh. The client could unserved appropriately thinking about their normal pay with their spending conduct [2] causes employee turnover. Employee turnover is high especially for Garments organizations in Bangladesh.

The issue of Bangladeshi organization's employee turnover rate has become an important part of a company's strategy. In the event of replacing an employee, the human resources team needs around two to eight weeks to find a good candidate for the given role and also a number of interviews must be conducted in order to find the optimal candidate that we rapidly

come to the conclusion that replacing an employee is a costly process. Garments and textile are leading to Bangladesh economy and in future its area is growing largely. Actually, this is manpower based low labor wages organization where employee turnover rate is high. Therefore, organization faces great loss due to this turnover. Since Bangladesh is over populated country, second generation industrialization is going on. Around four million people directly or indirectly depends on garments industry and about 40% of total GDP contributed by garments organization. As the second business hub of garments industry in the world faces many challenges due to employee turnover. Our proposed system identified the employee turnover prediction through machine learning algorithm so that turnover can be resist.

In this paper, we have implemented some of the well-known ML techniques namely Decision Tree, Logistic Regression, Support Vector Machine (SVM), Gradient Boosting classifier and Random Forest on the Human Resources (HR) Employee turn-over by collecting data from Bangladeshi Garments Organizations.

The rest of the paper organized as follows: section II describes the related work. Section III illustrated the methodology where section IV describes the data analysis and results are shown in section V. Finally section VI gives the conclusion with future work.

II. RELATED WORK

Employee turnover to any organization is commonly a bad practice to employee. Here creates a big financial loss to organization. Also, it reduces the working efficiency, productivity remarkably. Some Research has been done to find out the reason that is responsible for employee turnover [1]. In [1] an expert system has been implemented to predict employee churn. Employee turnover firmly related however not

indistinguishable from client beat is comparatively agonizing for an association, prompting disturbances, client disappointment, and time and endeavors lost in finding and preparing substitution. In [2] customer and employee segmentation were developed based on clustering method. In [3], the author explored several machine learning algorithms to predict employee turnover using a data set comprising 1450 records and 35 attributes. They used the following machine learning algorithms: K-Nearest Neighbor, Support Vector Machine, Decision Tree and Random Forest. And found result KNN 86.39%, SVM 86.84%, DT 81.63% and RF 88.43%.

In [4] the author gives an overview of churn and types on churn. In [6], the author described several machine learning algorithms to predict employee turnover using a data set comprising 1470 records and 34 attributes. They used the following machine learning algorithms: NB, Linear Regression, KNN, SVM, RF and found result DT 76.5%, NB 79.1%, Logistic Regression 87.1%, SVM 85.7%, KNN 84.4% and RF 85.0%. In [5], the author used Logistic Regression, LDA, DT, RF to predict employee turnover. And found result Logistic Regression 57.8%, LDA 57.7%, DT 56.3% and RF 59.7% In paper [6] author represented the similitudes between the issues of client turnover and employee turnover. An illustration of an employee turnover model created to utilize traditional AI methods. Representative turnover is the general churn, which alludes to individuals finding employment elsewhere in an association [7]. Representative turnover likewise can be called attrition. Attrition is basically the turnover rate of employees inside an organization. In [9] the author explored several machine learning algorithms to predict employee turnover using data set comprising 10616 records and 39 attributes and got accuracy 98%. by using Logistic Regression.

In [10] author worked on Swedbank employee turnover prediction by using RF, MLP, SVM machine learning methods where RF achieved highest 98.62% accuracy. In [11] they collected data from XYZ company duration of 2015 to 2017, 16,649 instances personal data and applied C4.5 classifier gained 95% accuracy. In [13] HR data of the employee collected from three IT in India and applied SVM algorithms achieved accuracy 85%. In [14] author worked on the effect of employee occasion functions on classifier where overall performance calculated on worker chance to turnover in the meantime preserving the interpretability of classifiers for advantages of retention interventions development.

However, we have implemented some of the well-known ML techniques using new and unique feature to predict employee turn-over by collecting data from Bangladeshi Garments Organizations.

III. METHODOLOGY

Figure 1 demonstrate the major components of the methodology which have been used in this paper to predict the employee turnover.

A. Data collection

The data set have been prepared by collecting data from organization. In data set, there are two types of data, numerical and categorical. There are no null values and for implementation categorical data is converted into numerical data. Finally Random Forest, Support Vector Machine, Decision Tree, Gradient Boosting classifier and Logistic Regression algorithm are implemented. Here precision and recall have been used to evaluate the performance. Moreover, we use 7-fold cross validation for algorithm, which removes to over-fitting.

Then dataset are categorized into two values numerical and categorical values. Then important feature are selected from database. After that feature scaling have been done to preprocess the dataset. Here we use 20 features to implement our model. Different machine learning algorithm is used to predict employee turnover where 70% data are used for training and 30% is used for testing. Here we have used confusion matrices to find the accuracy of the model. In addition, 7 fold cross validation is used to test the model. The flow chart is given below.

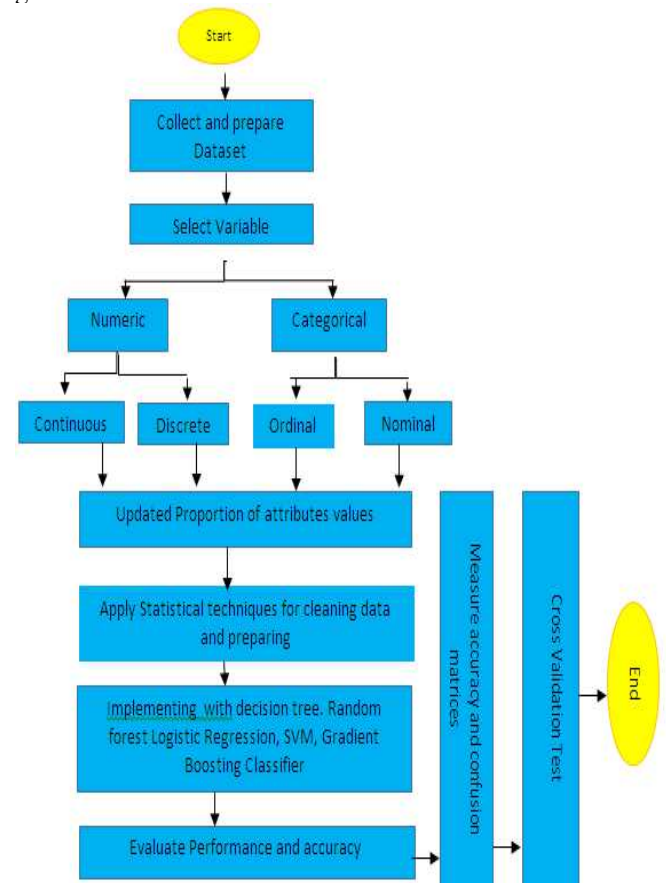


Fig.1. Flowchart of proposed work

B. Classification

In this paper several machine learning algorithm have been used to predict employee turnover. The description is given below.

Logistic Regression

The Logistic Regression is a linear classification algorithm. The models learn the weights of input dimension that separate output into two regions. Logistic regression is easy to implement and work well on linearly separable classes. This algorithm creates a logistic curve, which values lies between 0 and 1.

Random forest

Random forest (RF) is one of the most powerful supervised machine learning algorithms for generating classifications and regressions. Each tree votes for a classification label for a certain data set, then the RF model chooses which class had the most votes from the decision trees. Its output prediction accuracy is very high. If any missing value finds in data set, don't create any impact on accuracy.

Gradient Boosting Classifier

Gradient boosting classifier is also ensemble of decision tree method. The difference between RF and GBT is the gradient boosted tree models learn sequentially in attempts to correct the mistakes of the previous tree in the series. It provides predictive accuracy. There also lots of flexibility which optimize different less of functions and provides different hyper parameter tuning options and no need to preprocessing of data for this. It handles missing data and imputation is not required.

Support vector machine

It is not only used for classification but also use for Regression. SVM will train algorithms with assigned classes by separating each class through a decision boundary, also known as a hyperplane. Some problems are considered nonlinear in so far as it is difficult to draw the decision boundary. However, this can be solved by using a kernel function. This function returns the dot product of the two vectors, where it then maps data points to a new, transformed, high-dimensional space. Moreover, there are several types of kernel function can be used such as linear, Gaussian, and polynomial kernel.

Decision Tree

It is the most used for classification problem. Decision tree is only one single tree. It can be explained into two entities such as decision node and leaves.

C. Evaluation Measurement

To evaluate the accuracy precision, Recall, F-measure is used. We also find out confusion matrices shown in Figure 2 and the correlation matrices is shown in figure 3

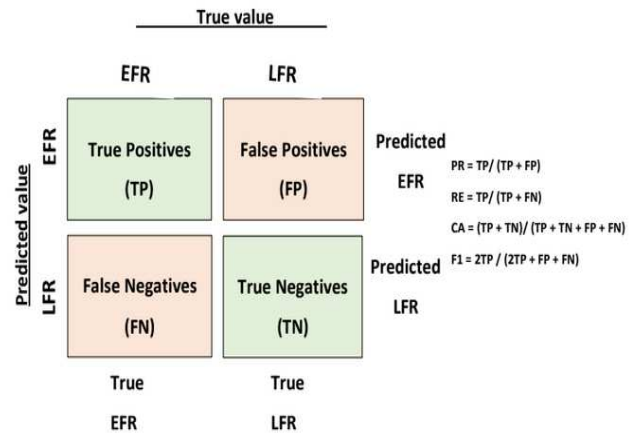


Fig.2. Confusion matrix

IV. DATA SET ANALYSIS

We have collected data from two organization named TRU Febrics and Z&Z organization. We also collected few data by google Form. Table 1. Shown the 19 generated feature which are completely helpful to predict turnover of employee. Some features have been included from Kaggle where analyzed Employee Attrition and Performance [12].

Table 1: Generated Feature

1	Gender	11	On time salary
2	Age	12	Yearly Salary Increment
3	Marital Status (Married / Unmarried)	13	Overtime Bill Allowance
4	Education Field	14	Salary Satisfaction
5	Departments	15	Employee Promotion Status
6	Education Qualification	16	Transport Facility
7	Human Resource Policy	17	Distance From Home(km)
8	Safety Satisfaction	18	Annual Refreshment Facility
9	Job environment Satisfaction	19	Stress Level
10	Job Satisfaction	20	Turnover

This dataset comprises twenty attributes and seven hundred tuples.

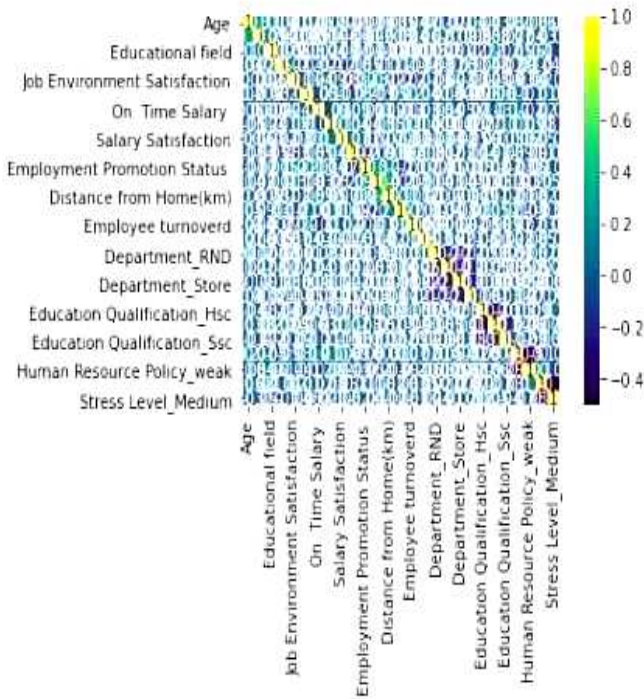


Fig.3 Correlation matrix

V. EXPERIMENTAL RESULTS

Table 2 shows that Random Forest achieved highest precision, recall and f1-score then other different machine learning algorithms.

Table 2: Comparison of different approach

Method	Precision	Recall	F1-score
Random Forest	0.99	0.99	0.99
Gradient Boosting	0.92	0.91	0.92
Decision Tree	0.97	0.96	0.96
SVM	0.91	0.91	0.91
Logistic Regression	0.72	0.71	0.72

Figure 4 has shown that comparison of accuracy of different approaches where Random Forest has higher accuracy over other compared machine learning approaches.

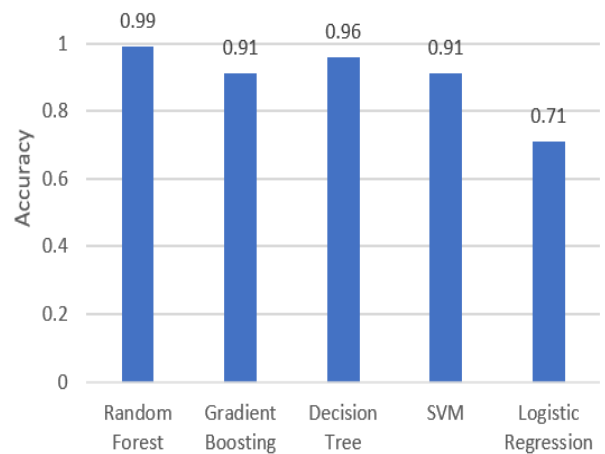


Fig.4. Comparison between different approaches.

Roc curve is a curve where we can see the accuracy of several model and ROC curve defines the comparison of the accuracy among different model. In this curve, we can easily understand that Random forest outperform better that other algorithm.

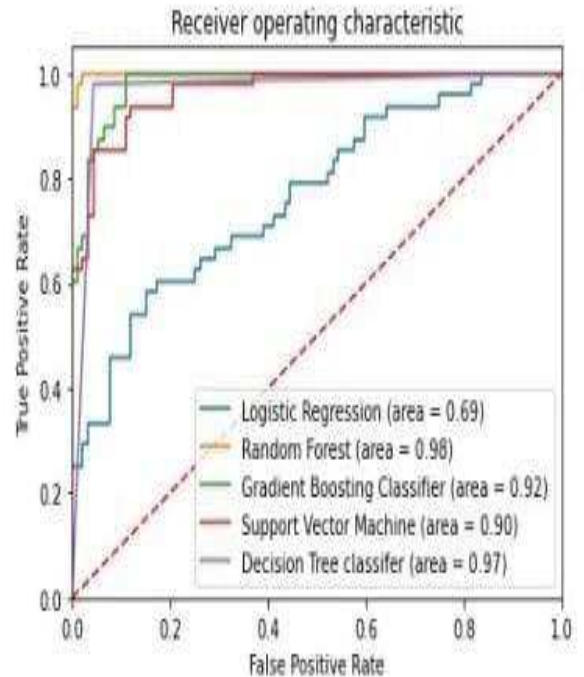


Fig.5. ROC Curve of accuracy model

In addition, a comparative study have been done on survival analysis where feature based analysis have been implemented. It will help the employer to take decision in recruitment. In figure 6 and 7 it is shown that the turnover rate is high for 27-28 years old employee and also if the distance of the organization from home is 7 to 9 km. In figure 8 we have seen that after higher

secondary education the employee turnover rate is low but in secondary education level it is high.

Out[46]: <AxesSubplot:xlabel='Age', ylabel='count'>

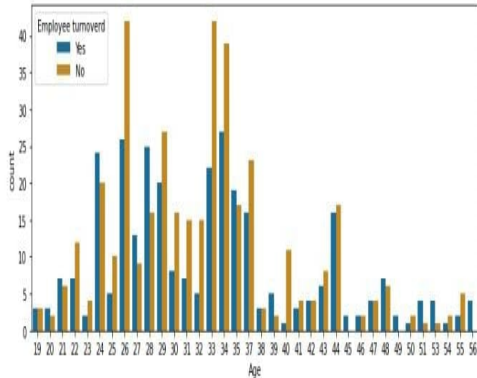


Fig.6: Employee Turnover based on age

Out[47]: <AxesSubplot:xlabel='Distance from Home(km)', ylabel='count'>

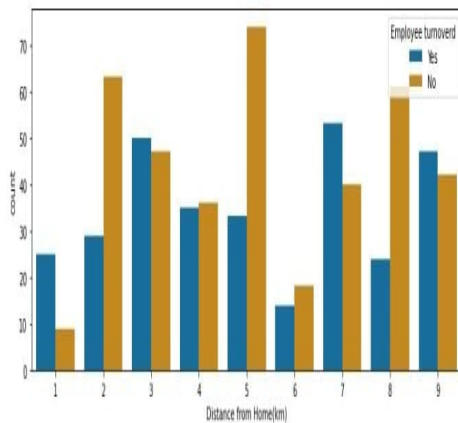


Fig.7: Employee Turnover based on distance.

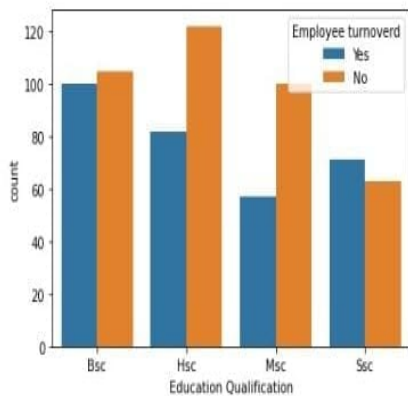


Fig.8: Employee Turnover based on education

VI. CONCLUSION AND FUTURE WORK

In the perspective of Bangladesh, this research has been worked newly though many proposals study and method already been established considering other countries. Collecting data from different organization is very hard. In our proposed work, we do some machine learning approaches to predict employee turnover.

With the machine learning approach, a good accuracy of 98% has been achieved with Random Forest algorithms. In future we will expand our dataset and deep learning approach will apply for more improving more accuracy.

References

- [1] V. Saradhi and G. K. Palshikar, "Employee Churn Prediction," Expert System with Application, vol. 38, no. 3, pp. 1999-2006, 2011.
- [2] Alamsyah and B. Nurriiz, "Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization" International Conference Science and Technology - Computer (ICST), vol. 3, 2017.
- [3] Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule, Suhasini Bhat, "Employee Attrition Predictive Model Using Machine Learning" Volume: 07 Issue: 05 | May 2020.
- [4] Chandrasekhar(chandu)valluri, "The many types of churn and their predictive" (August 12, 2019)
- [5] Ibrahim Onuralp Yiğit and Hamed Shourabizadeh "An Approach for Predicting Employee Churn by Using Data Mining" Conference: International Artificial Intelligence and Data Processing Symposium'17 At: Malatya, Turkey.
- [6] Ibrahim Onuralp Yiğit and Hamed Shourabizadeh "An Approach for Predicting Employee Churn by Using Data Mining" Conference: International Artificial Intelligence and Data Processing Symposium'17 At: Malatya, Turkey.
- [7] L. S. Fischer, "A predictive and prescriptive framework for employee churn," NOVA University of Lisbon, 2019.
- [8] Edouard Ribes, Karim Touahri, Benoît Perthame"Employee turnover prediction and retention policies design: a case study"(July 5, 2017)
- [9] HMN Yousaf, "Analysing which factors are of influence in predicting the employee turnover "(Nov 23, 2016)
- [10] Mari Maisuradze, "Predictive analysis on example of employee turnover "(June 923, 2017)
- [11] Nisrina Salma and Andry Alamsyah, "Employee Churn Prediction Model using C4.5 Classification Algorithm"International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296.
- [12] Kaggle, "HR-Employee-Attrition." [Online]. Available: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>
- [13] Khera, S.N. and Divya, 2018. Predictive modelling of employee turnover in Indian IT industry using machine learning techniques. Vision, 23(1), pp.12-21.
- [14] Juvitayapun, T. Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods, "13th International Conference on Knowledge and Smart Technology (KST)", 181-185.

Machine learning Models to Predict COVID-19 Cases in the DC Metro Area

Michael Thompson
George Mason University
Fairfax, VA, USA
mthomp19@gmu.edu

Maryam Heidari
George Mason University

Ashritha Chitimalla
George Mason University

Fairfax, VA, USA
mheidari@gmu.edu

Fairfax, VA, USA
achitima@gmu.edu

Ghadah Alshabana
George Mason
University Fairfax, VA,
USA galshaba@gmu.edu

Thao Tran
George Mason
University Fairfax, VA,
USA ttran81@gmu.edu

Abstract—Coronavirus can be transmitted through the air by close proximity to infected persons. Commercial aircraft are a likely way to both transmit the virus among passengers and move the virus between locations. The importance of learning about where and how coronavirus has entered the United States will help further our understanding of the disease. Air travelers can come from countries or areas with a high rate of infection and may very well be at risk of being exposed to the virus. Therefore, as they reach the United States, the virus could easily spread. On our analysis, we utilized machine learning to determine if the number of flights into the Washington DC Metro Area had an effect on the number of cases and deaths reported in the city and surrounding area.

Index Terms—coronavirus, Washington DC, virus prediction, machine learning

I. INTRODUCTION

The COVID-19 pandemic has caused a serious impact on the world economy, with many deaths and long-term injuries across the world, an increase in businesses bankruptcy with an associated increase in lost jobs, and an increase in food scarcity as well. Moreover, health care systems, as well as the airline industry, faced enormous financial challenges as 30% of the United States airlines' stock prices decreased during this crisis [1]. During this crisis, many airlines canceled their flights and applied travel restrictions to control the spread of coronavirus, leading to a significant impact on the airline sector. According to [2], [3], the ongoing COVID-19 pandemic is the gravest crisis that happened to the aviation sector, and it may take around five years to recover from it. As thousands of employees were released, international tourism was frequently stopped, and global quarantine restrictions limited the flights that continued.

Given the serious effect COVID-19 has had on the world, there is a significant necessity to understanding how the novel coronavirus is transmitted and the various factors that allow it to rapidly disperse through a community, city, or nation. Among the risk factors to a given location, air travel may well be a factor, or have historically played a factor in allowing additional spread of the coronavirus from more infected regions to those with limited or no prior infections.

Numerous airplanes are supported with High Efficiency Particulate Air (HEPA) filters; however, the filter's effectiveness only applies to the air that goes through it, and it does not necessarily filter every onboard virus; therefore, airlines require people to wear masks during the flight [4]. Additionally, close contact between people is the prime cause of spreading the COVID-19 easily [5] which led many airlines to leave the middle seat empty in their aircraft [6]. However, with all travel restrictions and procedures, COVID-19 can still

be transmitted between people on airplanes, as stated by [7]–[9]. Therefore, our aim is to discover the correlation, if present, between the number of flights and resulting coronavirus cases in the DC area.

The importance of learning about where and how coronavirus has entered the United States will help further our understanding of the disease. According to CDC [10], the first coronavirus case in the US has been identified in Washington state, and that was due to air travel from Wuhan, China. The most common way COVID-19 can spread is by human interaction, through respiratory droplets such as talking, coughing, sneezing, and more. Air travelers can come from countries or areas with a high rate of infection and may very well be at risk of being exposed to the virus. Therefore, as they reach the United States, the virus could easily spread. In our analysis, we intend to use the OpenSky dataset records and combine it with CDC data to determine if the number of flights into or out of the Washington DC metro area may have impacted the number of coronavirus deaths reported in those counties and the region surrounding the respective airports in question.

Other analyses have concluded that coronavirus can travel via flight and there is an inverse relationship between distance to an airport and how many coronavirus cases result from travel into the region. We suspect that the District of Columbia will show different results as a significant portion of business and political activity in the region is focused on the US federal government.

In the next sections of this paper, we make the following contributions. First, we summarize the datasets that have been utilized in this study. Second, we provide an overview of the related work in relation to this study. Third, we discuss the methodology for the analysis of this project. Finally, we present our primary findings followed by the future work.

II. DATASETS

In this paper, we have utilized two dataset sources. The flight dataset was obtained from OpenSky, showing the air traffic during the coronavirus pandemic. Table I summarizes the OpenSky attributes [11]. The coronavirus dataset was obtained from the New York Times and shows the number of cases and death in the United States. Table II summarizes the New York Times [12].

As for the cleaning process. From the OpenSky dataset: the destination variable (filtering for Baltimore International Airport, Dulles International Airport, and Reagan National Airport), and Last seen variable as that gave us an indication of the date the flight was occurring. From the New York times dataset, we used the date, state, and county to filter down to the specific counties surrounding each of the airports mentioned

earlier, and the area surrounding Washington DC itself. We further intend to use the cases and deaths to calculate the number of new cases and deaths occurring each day.

Data cleaning is performed using Python and Tableau Prep. Pandas library in python is used to drop the missing values and the redundant entries and remove the irrelevant columns from the datasets to decrease the processing time and enhance performance and efficiency. Moreover, we resolved the data inconsistencies, filtered the destination variable, and converted the UTC timestamp to date format using the Tableau Prep tool.

TABLE I
OPENSKEY ATTRIBUTES

Variable Name	Description	Type
Callsign	the identifier of the flight	String
Registration	the aircraft tail number	String
Origin	the origin flight airport represented with four letters.	String
Destination	the destination flight airport, represented with four letters.	String
Firstseen	UTC timestamp of the first message received by the OpenSky Network.	String
Lastseen	UTC timestamp of the last message received by the OpenSky Network.	String

TABLE II
NEW YORK TIMES ATTRIBUTES

Variable Name	Description	Type
Date	the date of the reported Covid-19 cases and deaths.	Date
State	the name of the state.	String
County	the name of the county.	String
Fips	standard geographic identifier.	Number
Cases	the total number of cases of Covid-19.	Number
Deaths	the total number of deaths from Covid-19.	Number

III. RELATED WORK

The rapid spread of coronavirus cases across the world motivates us to discover the number of flights effect on coronavirus deaths rate. As in almost every country, the first infection cases of coronavirus were brought by travelers. While travel restrictions have been applied in many countries, they had a modest effect on limiting the spread of coronavirus cases [13], [14]. These restrictions were effective in only delaying the transmission of coronavirus [15].

In order to minimize the transmission of coronavirus during flights, several airports implemented a temperature screening to check for fever. This however was determined to be ineffective, as estimations suggest only 45% of infected people would be detected by temperature screening with an even lower result in young people [10]. Khanh et al. [7] study shows that thermal imaging scans have their limitations at determining if someone certainly has coronavirus. An additional lack of self-disclosure of coronavirus symptoms before and after boarding leads to an increase in the spread of the COVID-19 [7]. Previous studies have also reported that coronavirus can be transmitted before

symptoms appear [8], [9], as people can be infected with coronavirus disease and show no symptoms, or have symptoms develop over a period of several days. Overall, Khanh et al. [7], Bae et al. [8], and Choi et al. [9] concluded that coronavirus could be transmitted on aircraft and consequently increase the infection risk.

A case study was applied in China to calculate the risk index of COVID-19 imported cases from inbound international flights [10]. Through this study, Zhang et al. [16] used global COVID-19 data and international flight data from UMETRIP, and they found that the risk index increases significantly when there are active flights associated with highly infected countries. A research on the effect of the United States local flights on COVID-19 cases indicates a high correlation, i.e., 0.8 between travelers and population and COVID-19 cases at the onset of the pandemic [17].

However, a study conducted and published in 2020 by Desmet and Wacziarg [18] used a cross-sectional regression model to analyze the relationship between COVID-19 cases and deaths and the distance to the closest airport with direct flights from the top five affected countries and found a negative correlation. Throughout this study, Desmet and Wacziarg [18] used the collected COVID-19 cases data from the New York Times and the international flights from the Bureau of Transportation Statistics. A retrospective case series was conducted by Yang et al. [19]. It was clinical data collected from ten patients with no symptom history before the flight. Yang et al. [19] found that coronavirus disease can be transmitted through airplanes and the way how transmitted was unknown. Similarly, a medical evaluation was conducted by Hoehl et al. [20] on twenty-four passengers from an international flight from Israel to Germany. Seven passengers were tested positive for coronavirus (SARS-CoV-2). Hoehl et al. [20] concluded that coronavirus transmission in an airplane depends on other elements such as passenger’s movement, contact, etc. at the same time, wearing a mask during the flight could reduce the transition rate. On the other hand, a study by Schwartz et al. [21] stated that Coronavirus disease 2019 transmission was absent base on patients who traveled an internationalflight with one stop from China (Wuhan) to Canada (Toronto). Although several passengers developed some symptoms after the flight but tested negative from COVID-19.

Another study was conducted to investigate the association between air traffic volume and the spread of COVID-19. This study was conducted using the publicly available of domestic air traffic and passenger data from 2013 to 2018 through CAAC (Civil Aviation Administration of China). This data was used to predict the data for 2019. For the international air traffic and routes, data were derived from Chinese international air traffic from the Official Aviation Guide (OAG) and COVID-19 data from WHO [22]. They had found the continuous measurements using mean ± standard error, statistical significance using t-test, and correlation analysis using linear regression. The analysis indicated a strong direct correlation between domestic covid cases and the number within China itself. The international air traffic analysis also

showed a strong correlation with the cases as it depends on air traffic network [22]. Moreover, a study based on a cross-country regression analysis approach was contacted to find the correlation between coronavirus deaths and cases and global tourism [23]. The COVID-19 cases and deaths data for this study were collected from the European Centre for Disease Prevention and Control while the international tourism data were obtained from the World Bank. [23] found a positive association between coronavirus cases and deaths and international tourism.

Furthermore, a mathematical modelling study was conducted to see whether the spread of COVID-19 was related to international cases imported. The study compared the ratio between the international imported cases and the internal cases in May 2020 and September 2020 in different countries [24]. Since each country has its own regulation implemented on the restriction, the results found at international cases imported made larger impact on May 2020 than September 2020. In fact, imported cases effect the internal spread in May 2020 was reduced compared to September 2020. However, the data collected for this study were from OpenSky data found the impact of international imported cases to internal spread is still little with 10% or less. [24]

IV. METHODOLOGY

The first step in this project was cleaning the datasets. Specifically, the Open Sky data, being crowd sourced, included errors such as duplicate entries or null origin and destination values among other things. We further cut out the irrelevant data from the OpenSky dataset (that being the time period from 2019 until the start of the pandemic) as that information is irrelevant to our analysis. At this stage in the process, the coronavirus data were also filtered to only include the relevant region. Specifically, all cities and municipalities between the airports being studied and Washington, DC itself (including the counties containing each airport).

Our proposed method will involve marking the major airports in the Washington DC Metro Area and their immediately adjacent counties. We will then run a count on all flights arriving at each airport across the United States aggregated by week. Current CDC data suggest that the time from infection to onset of symptoms is 4 – 5 days [25] and research published in the Journal of Medical Virology indicates death from coronavirus has a median of 14.5 days after initial exposure [26]. As a result, based on this information, we intend to pair our flight data with new coronavirus cases averaged over the next 14 days after the flight. This should allow us to run several initial tests, including (assuming linear correlation) a Pearson Correlation Coefficient to determine if a correlation exists between number of flights and number of coronavirus deaths.

We can further take this data and produce scatterplots to visualize any correlation and determine other factors like spread and variance. Depending on the results of this analysis and visualization, we will then use machine learning in MATLAB to experiment with appropriate regression methods and attempt

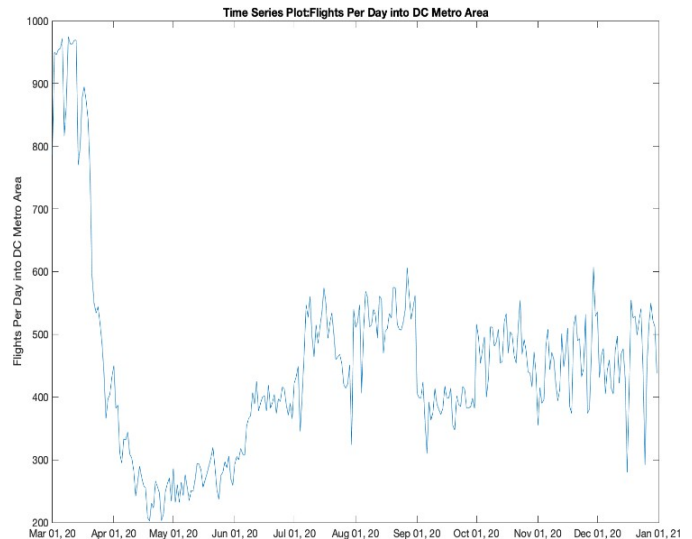


Fig. 1. Flights activity per day in DC Metro Area

to produce a model that can determine (within a degree of confidence) the likely number of additional coronavirus deaths that may result from the ongoing flights into a given airport. Control counties with 0 flights will be added to the model as well, chosen at random, which do not contain airports and are not adjacent to counties with airports.

To allow for public review and availability of our data and ongoing progress, we will be posting the progress of our project online [27].

V. INITIAL ANALYSIS AND VISUALIZATION

When the initial data preparation and cleaning were completed, we began using MATLAB to visualize the data available. To begin, we took a look at the number of flights per day to determine if there are any clear trends, we should look at in the corresponding coronavirus data. As can be seen in Fig. 1 below, we were able to spot several regions of particularly low and high activity relative to the average of 447 flights per day.

Additionally, at this stage we began looking at the increases in cases and deaths per day for each region. Fig. 2 illustrates the cases and deaths reported in Washington DC itself for the time period of March 2020 through March of 2021. Noting that our flight data ends on January 1st of 2021, there are still several clear peaks that can be looked at (at least in this data set) to determine if the number of flights into the DC region have an effect on the resulting cases/deaths.

To accomplish this initial look at how closely correlated flights are with deaths and cases per day, we ran some preliminary analysis and attempted to create scatterplots around our initial concept for the Washington DC area. That is to say that we compared flights with the number of cases reported a week later when we anticipated those would be detected or otherwise reported to the appropriate health agencies. When looking at this data we found unexpectedly that there was very limited negative correlation. Running Pearson's correlation coefficient against the two values yielded a result of -0.29 where the more

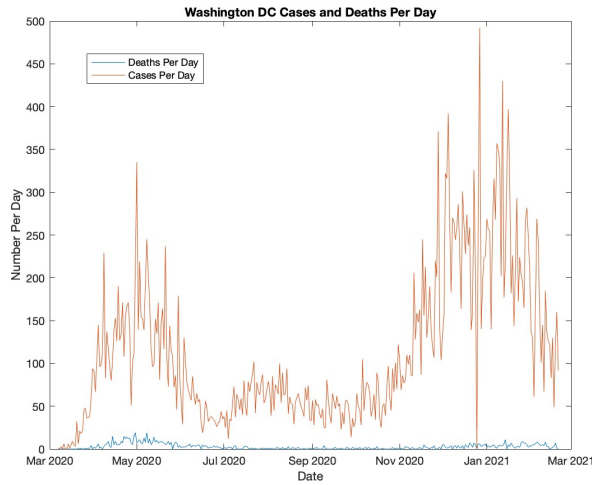


Fig. 2. Washington DC cases and death per day

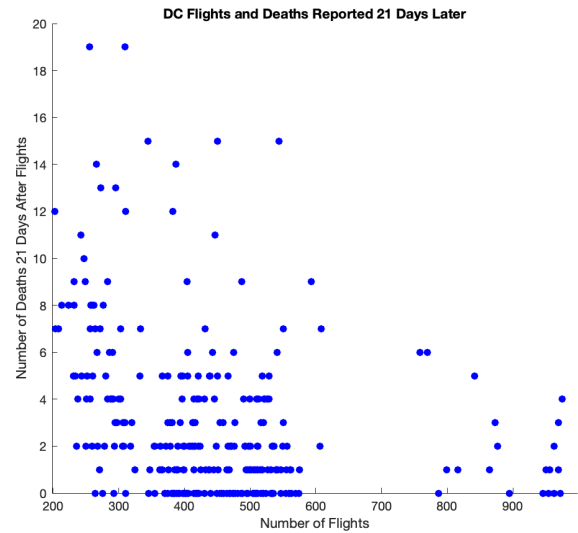


Fig. 4. Washington DC flights and deaths reported three weeks later

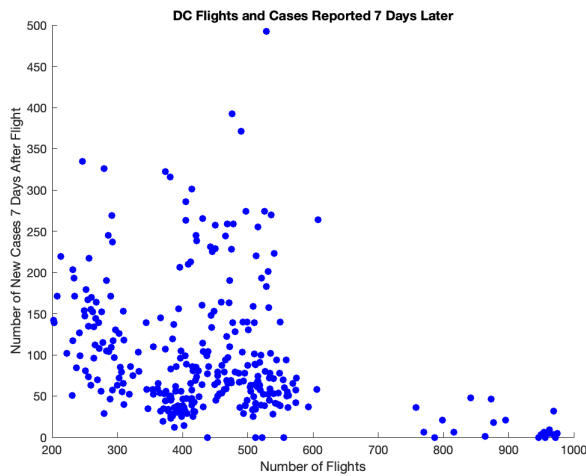


Fig. 3. Washington DC flights and cases reported a week later

flights reported the fewer cases were reported 7 days later. This can be roughly seen in Fig. 3 the scatterplot of these two values with a slight negative trend. Note however that the large accumulation of points near the right end of the graph at 1000 flights per day reflects early flight data towards the beginning of the pandemic and may need to be considered differently as lockdowns and containment procedures by airlines had not yet been implemented at the time.

A similar attempt was made at analyzing the relationship between flights and deaths reported 21 days later. Fig. 4 shows a slightly more pronounced negative correlation of -0.31 and a similarly more pronounced trend in the visible scatterplot of these values. Note that the same limitation of the cases applies, where the cluster of dots at the right side of the graph reflects data from early on in the pandemic before the coronavirus was as widespread.

As our team had initially hypothesized that there would

be a clear positive correlation between these values, it's clear that these unanticipated results may require further review to ensure that our existing methodology is correct, particularly in light of the literature review that we previously conducted. It is however possible that our methodology is correct and that the negative correlation is due to other factors, such as local lockdowns, restrictions and policy which we also have yet to conclusively determine. Additionally, as noted before there is still data in our current analysis and visualizations reflecting the period early in the pandemic where the results of flight into the region may reflect differently on resulting cases than it would later when the virus is more common.

VI. ANALYSIS

After reviewing the results of our initial analysis, we determined that starting our analysis on April 1st (after any government restrictions were put in place) would better reflect the impact of flights on case numbers. Additionally, as our initial method of looking at cases and deaths a number of days past the flight misses the large number of cases reported before and after the chosen day, we determined that a better method would be using an average of the number of new cases reported up to 14 days after the initial flight. Based on these updated methods, here were our analysis results:

A. Loudoun County and Dulles International Airport

When looking at Dulles Airport and the surrounding Loudoun County, we determined that there was a fairly minimal but still present positive trend between the number of flights and new cases over 14 days, with a correlation coefficient of 0.16. In attempting to model this with linear regression however, the following result in Fig. 5 was produced.

The blue dots are a scatterplot of the actual data points, while yellow is the prediction based on the model. With 25% holdout validation, the root mean square error was 35.866.

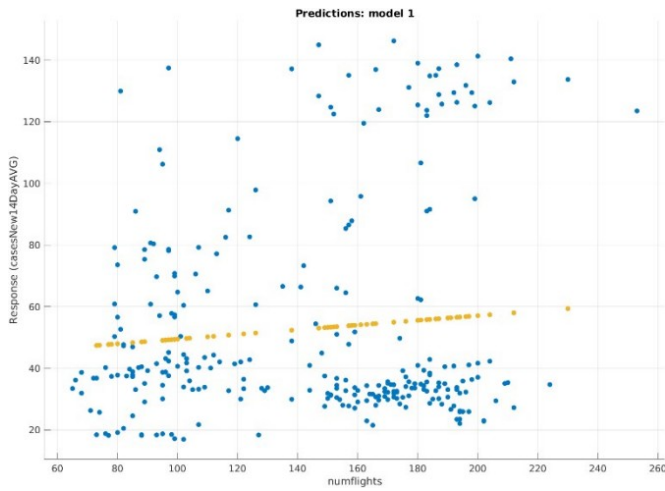


Fig. 5. Linear Regression Model- Flights into Dulles Airport and New Cases in Loudoun County over next 14 days.

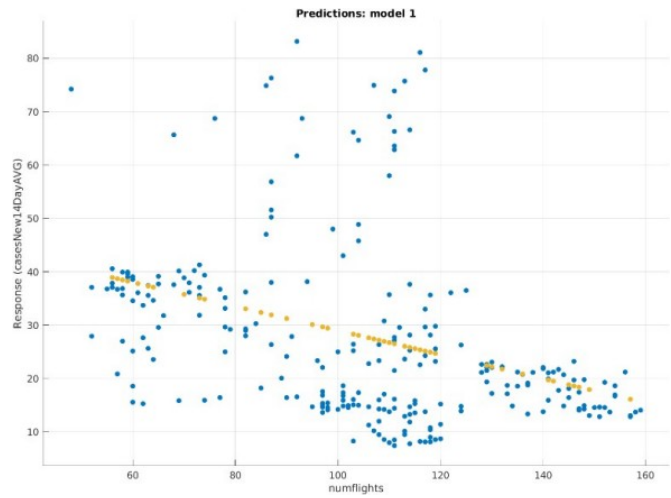


Fig. 6. Linear Regression Model- Flights into Reagan National Airport and New Cases in Arlington County over next 14 days

Based on this, and the visible chart above, it seems apparent that while a positive correlation may exist, there is significant variance in the 14-days average of new cases that limits the predictive power of linear regression.

B. Arlington and Reagan National Airport

A similar result was achieved when looking at Reagan National Airport and the surrounding Arlington County. Based on our data, we unexpectedly observed a negative trend between the number of flights and the number of new cases in Arlington, with a correlation coefficient of -0.376 . Our linear correlation model is illustrated in Fig. 6. Based on the same 25% holdout validation, the root mean square error of this model was 15.084. While a much clearer negative trend exists in this chart, it should however be noted that there may be unaccounted confounding variables, such as the closure of Gate 35X at Reagan National Airport and subsequent opening of a new concourse over the next few months.

C. Anne Arundel County and Baltimore Airport

Anne-Arundel County holds the last of the three major airports in the Washington DC Metro Area. Unlike Reagan National, this airport showed a slight positive correlation between the number of flights and the 14-day average of new cases, with a correlation coefficient of 0.2 . Our resulting model can be seen in Fig. 7 below.

With 25% holdout validation, the root mean square error for this model was 41.43. As can be seen above, while there is a slight positive trend, most 14-day averages are clustered around 50 cases per day, regardless of the number of flights reported.

D. Washington, DC itself

Among the considerations we had when looking at this data, we suspected that a large number of travelers into these airports may be flying into the region to work in Washington, DC as the largest city in the region. As a result, the final

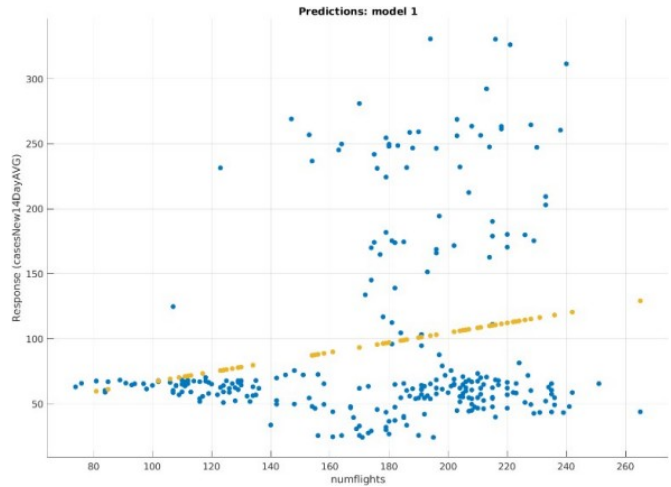


Fig. 7. Linear Regression Model- Flights into Baltimore International Airport and New Cases in Anne-Arundel County over next 14 days

model we produced took all flights into all three airports and, like prior models, paired them to the 14-day average of new cases reported in Washington, DC. Our correlation coefficient showed a slight negative trend of -0.767 , the specific model produced is illustrated below in Fig. 8.

As can be seen above, this was unfortunately the least predictive of our models. Based on 25% holdout validation, the RMSE of this model was a comparatively high 69.238.

VII. CONCLUSIONS

What we determined based on our work in modeling the number of flights and resulting cases into the DC Metro Area is that the number of flights by itself seems to have a very minimal impact on the number of new cases reported over the next 14 days. We suspect this may be due to several reasons. It is possible that the number of flights (which notably non-passenger flights carrying cargo) isn't a good metric for

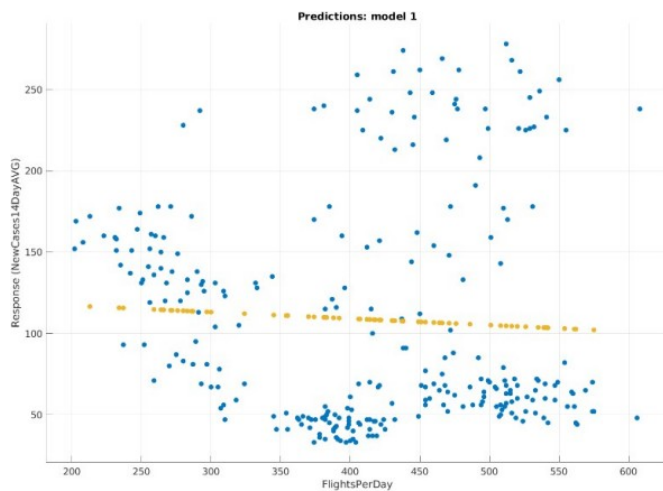


Fig. 8. Linear Regression Model- Flights into IAD, BWI, and DCA, and New Cases in the District of Columbia over the next 14 days

measuring how flights impact the number of reported cases in a region. It is also possible that airline policies have played a significant role in reducing in-flight transmission of coronavirus. We also surmise that a lack of testing requirements either before or after flights coupled with varying testing infrastructure in Maryland, Virginia, and Washington, DC may cause cases resulting from flight to be missed by our sources. Regardless, our conclusion at this time is that the number of flights by itself does not strongly correlate to the resulting number of new cases reported in the surrounding county over the next 14 days and should not be used as the sole predictor in future machine learning systems.

VIII. FUTURE WORK

Although the results showed little correlation because of travel by flight and COVID-19 cases in the DMV region, there are still significant improvements that can be made to our work. As COVID-19 is relatively still new, there are a lot of different experiments and tests that can be done. As a possible example, future algorithms may consider taking the origin airport into consideration. Another plausible research idea would be to see what kind of safety precautions were taken and using hypothesis testing to determine which specific measures reduce coronavirus transmission and to what degree. As a direct improvement to our method, rather than looking at the number of flights another approach for this project would have been to look at the total number of passengers each day, thus excluding the commercial cargo flights. This may allow for a clearer understanding of the correlation between air travel and COVID-19 cases. While it was deemed inappropriate for our usage (as we were anticipating and trying to determine linear correlation), yet another improvement would be attempting to model these data with an array of other unsupervised machine learning algorithms, selecting for the lowest RMSE to find a better model for the data (likely combined with other improvements mentioned above).

REFERENCES

- [1] C. Assis. Airline stocks slammed by coronavirus fears, but experts say reaction may be overdone. (accessed April 25, 2021). [Online]. Available: <https://www.marketwatch.com/story/airline-stocks-slammed-by-coronavirus-fears-but-experts-say-reaction-may-be-overdone-2020-03-06>
- [2] United airlines says coronavirus pandemic is worst crisis 'in the history of aviation'. (accessed April 26, 2021). [Online]. Available: <https://www.marketwatch.com/story/united-airlines-says-coronavirus-pandemic-is-worst-crisis-in-the-history-of-aviation-2020-04-30>
- [3] J. Jolly. Airlines may not recover from covid-19 crisis for five years, says airbus. (accessed April 26, 2021). [Online]. Available: <https://www.theguardian.com/business/2020/apr/29/airlines-may-not-recover-from-covid-19-crisis-for-five-years-says-airbus>
- [4] J. Read. How clean is the air on planes? (accessed April 27, 2021). [Online]. Available: <https://www.nationalgeographic.com/travel/article/how-clean-is-the-air-on-your-airplane-coronavirus-cvd>
- [5] Centers for Disease Control and Prevention (CDC). How covid-19 spreads. (accessed April 27, 2021). [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html>
- [6] R. J. Milne, C. Delcea, and L.-A. Cotfas, "Airplane boarding methods that reduce risk from covid-19," *Safety Science*, vol. 134, p. 105061, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925753520304586>
- [7] N. C. Khanh, P. Q. Thai, H.-L. Quach, N.-A. H. Thi, P. C. Dinh, T. N. Duong, L. T. Q. Mai, N. D. Nghia, T. A. Tu, L. N. Quang, T. D. Quang, T.-T. Nguyen, F. Vogt, and D. D. Anh, "Transmission of SARS-CoV 2 during long-haul flight," *Emerging Infectious Diseases*, vol. 26, no. 11, pp. 2617–2624, Nov. 2020. [Online]. Available: <https://doi.org/10.3201/eid2611.203299>
- [8] S. H. Bae, H. Shin, H.-Y. Koo, S. W. Lee, J. M. Yang, and D. K. Yon, "Asymptomatic transmission of SARS-CoV-2 on evacuation flight," *Emerging Infectious Diseases*, vol. 26, no. 11, pp. 2705–2708, Nov. 2020. [Online]. Available: <https://doi.org/10.3201/eid2611.203353>
- [9] E. M. Choi, D. K. Chu, P. K. Cheng, D. N. Tsang, M. Peiris, D. G. Bausch, L. L. Poon, and D. Watson-Jones, "In-flight transmission of SARS-CoV-2," *Emerging Infectious Diseases*, vol. 26, no. 11, pp. 2713–2716, Nov. 2020. [Online]. Available: <https://doi.org/10.3201/eid2611.203254>
- [10] Centers for Disease Control and Prevention (CDC). First travel-related case of 2019 novel coronavirus detected in united states. (accessed April 27, 2021). [Online]. Available: <https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html>
- [11] X. Olive, M. Strohmeier, and J. Lübke, "Crowdsourced air traffic data from the opensky network 2020," 2021. [Online]. Available: <https://zenodo.org/record/4419079>
- [12] The New York Times, "Nytimes/covid-19-data," <https://github.com/nytimes/covid-19-data/blob/master/us-states.csv>, 2021, March 13.
- [13] M. Bielecki, D. Patel, J. Hinkelbein, M. Komorowski, J. Kester, S. Ebrahim, A. J. Rodriguez-Morales, Z. A. Memish, and P. Schlagenhauf, "Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: A narrative review," *Travel Medicine and Infectious Disease*, vol. 39, p. 101915, Jan. 2021. [Online]. Available: <https://doi.org/10.1016/j.tmaid.2020.101915>
- [14] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, Mar. 2020. [Online]. Available: <https://doi.org/10.1126/science.aba9757>
- [15] A. Adekunle, M. Meehan, D. Rojas-Alvarez, J. Trauer, and E. McBryde, "Delaying the COVID-19 epidemic in australia: evaluating the effectiveness of international travel bans," *Australian and New Zealand Journal of Public Health*, vol. 44, no. 4, pp. 257–259, Jul. 2020. [Online]. Available: <https://doi.org/10.1111/1753-6405.13016>
- [16] L. Zhang, H. Yang, K. Wang, Y. Zhan, and L. Bian, "Measuring imported case risk of COVID-19 from inbound international flights — a case study on china," *Journal of Air Transport Management*, vol. 89, p. 101918, Oct. 2020. [Online]. Available: <https://doi.org/10.1016/j.jairtraman.2020.101918>

- [17] J. A. Ruiz-Gayosso, M. del Castillo-Escribano, E. Hernández-Ramírez, M. del Castillo-Mussot, A. Pérez-Riascos, and J. Hernández-Casildo, "Correlating USA COVID-19 cases at epidemic onset days to domestic flights passenger inflows by state," *International Journal of Modern Physics C*, vol. 32, no. 01, p. 2150014, Nov. 2020. [Online]. Available: <https://doi.org/10.1142/s0129183121500145>
- [18] K. Desmet and R. Wacziarg, "Understanding spatial variation in COVID-19 across the united states," National Bureau of Economic Research, Tech. Rep., Jun. 2020. [Online]. Available: <https://doi.org/10.3386/w27329>
- [19] N. Yang, Y. Shen, C. Shi, A. H. Y. Ma, X. Zhang, X. Jian, L. Wang, J. Shi, C. Wu, G. Li, Y. Fu, K. Wang, M. Lu, and G. Qian, "In-flight transmission cluster of COVID-19: A retrospective case series," Mar. 2020. [Online]. Available: <https://doi.org/10.1101/2020.03.28.20040097>
- [20] S. Hoehl, O. Karaca, N. Kohmer, S. Westhaus, J. Graf, U. Goetsch, and S. Ciesek, "Assessment of SARS-CoV-2 transmission on an international flight and among a tourist group," *JAMA Network Open*, vol. 3, no. 8, p. e2018044, Aug. 2020. [Online]. Available: <https://doi.org/10.1001/jamanetworkopen.2020.18044>
- [21] K. L. Schwartz, M. Murti, M. Finkelstein, J. A. Leis, A. Fitzgerald-Husek, L. Bourne, H. Meghani, A. Saunders, V. Allen, and B. Yaffe, "Lack of COVID-19 transmission on an international flight," *Canadian Medical Association Journal*, vol. 192, no. 15, pp. E410–E410, Apr. 2020. [Online]. Available: <https://doi.org/10.1503/cmaj.75015>
- [22] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, M. Zacharski, J. Bania, and T. Khosrawipour, "The association between international and domestic air traffic and the coronavirus (COVID-19) outbreak," *Journal of Microbiology, Immunology and Infection*, vol. 53, no. 3, pp. 467–472, Jun. 2020. [Online]. Available: <https://doi.org/10.1016/j.jmii.2020.03.026>
- [23] M. R. Farzanegan, H. F. Gholipour, M. Feizi, R. Nunkoo, and A. E. Andargoli, "International tourism and outbreak of coronavirus (COVID-19): A cross-country analysis," *Journal of Travel Research*, vol. 60, no. 3, pp. 687–692, Jul. 2020. [Online]. Available: <https://doi.org/10.1177/0047287520931593>
- [24] T. W. Russell, J. T. Wu, S. Clifford, W. J. Edmunds, A. J. Kucharski, and M. Jit, "Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study," *The Lancet Public Health*, vol. 6, no. 1, pp. e12–e20, Jan. 2021. [Online]. Available: [https://doi.org/10.1016/s2468-2667\(20\)30263-2](https://doi.org/10.1016/s2468-2667(20)30263-2)
- [25] CDC, "Interim clinical guidance for management of patients with confirmed coronavirus disease (covid-19)," <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html>, 16 Feb, 2021.
- [26] W. Wang, J. Tang, and F. Wei, "Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in wuhan, china," *Journal of Medical Virology*, vol. 92, no. 4, pp. 441–447, Feb. 2020. [Online]. Available: <https://doi.org/10.1002/jmv.25689>
- [27] M. Thompson, G. Alshabana, T. Tran, and A. Chitimalla, "Predict covid-19 cases using opensky data," <http://mason.gmu.edu/~ttran81/>, 2021.

Server-Side Distinction of User Mobility Using Machine Learning on Incoming Data Traffic

Hosam Alamleh
Computer Science

University of North Carolina Wilmington
Wilmington, North Carolina
hosam.amleh@gmail.com

Ali Abdullah S. AlQahtani
Computer Systems Technology

North Carolina A&T State University
Greensboro, North Carolina
alqahtani.aasa@gmail.com

Baker Al Smadi
Computer Science

Grambling State Univeresity
Grambling, Louisiana
bakir_smadi@hotmail.com

Abstract—During two decades, there have been a revolution in the field of digital communication and internet access. Today, it became possible for users to access the internet while on the move through an infrastructure of high-speed mobile broadband networks. Technologies such as LTE and 5G became essential. Mobile broadband networks allow mobility; connection reliability drops during movement. Thus, some failure intolerant processes, such as system updates, necessitates the utilization of a reliable connection. This paper introduces a model that predicts whether the user is mobile or stationary. This is done based on the traffic patterns at the server-side. Distinct network technologies entails distinct nature of traffic patterns. In this paper, machine learning is utilized at the server-side to allow differentiating between data transmitted by a stationary user and data transmitted by a mobile user at the server-side. Supervised training is utilized to train the model. Then, the model was tested and prediction accuracy of this model was 92.6 percent. Finally, the proposed system is a novel work and the first of its kind since it is the first to attempt to predict mobile network user's mobility at the server-side by utilizing packets' arrival patterns. The proposed system can be applied at mobile apps and allow them to collect data about the apps users mobility while using this service without needing to access the GPS. Also, it can be used network management and public safety.

Index Terms—machine learning, traffic, mobile, stationery, broadband

I. INTRODUCTION

In recent years, there have been prompt improvements in communication and information technology fields. Since the beginning of this century, these fields have caught a huge revolution that has had profound and powerful influences on humans' lives. At the start of the new millennium, users had to be stationary to be able to access the internet. Landline telephone systems with dial up internet were the main means of communication between individuals. Today, billions of individuals utilize their mobile phones and cellular network infrastructure to access the internet while on the move.

Today, mobile network subscribers have reached 5.31 billion worldwide [1]. Not only has the number of subscribers increased but also the number of devices.

According to Juniper Research, by the end of 2022, the number of devices connected to the internet will reach an astounding 46 billion [2]. Moreover, this communication revolution is linked to a revolution in data-processing, particularly with technologies, such as artificial intelligence, cloud computing, internet of things, robotics, and others.

Mobile broadband networks utilize cellular wireless technologies that allow users to connect their mobile devices (e.g., smartphones, tablets) wirelessly to a broadband internet connection through the mobile cellular network. Such networks offer high-speed internet access provided by the infrastructure of the mobile cellular networks network (e.g., CDMA, GSM, 3G, LTE, and 5G). However, traffic coming from mobile networks is still behind fixed networks' counterparts. As of 2018, the traffic arriving through the fixed broadband infrastructure was four times its mobile counterpart in the US [3]. This is because those mobile networks are less available. Further, the data transmission speed fluctuates due to resource sharing. Mobile networks are more costly compared to fixed networks. Therefore, data usage caps are in place with mobile data plans.

While mobile networks offer users internet access while on the move, this access is subject to several variables and, therefore, might be less reliable. Such variables include the coverage level and the presence of gaps in coverage. Also, in mobile networks, users compete for resources with other users, and when the network is busy, the mobile network assigns more bandwidth to users with better channel conditions, and sometimes the network would disconnect users with bad channel characteristics to serve other users with better channel conditions [4]. Therefore, when mobile devices conduct critical function (e.g., system update) that necessitates the use of a reliable internet connection. Some systems require users to be connected to a fixed broadband network to be updated. On the other hand, access to a mobile network while driving is dangerous

and illegal. This paper proposes an approach that distinguishes whether a user is mobile or stationary. This is done by examining the patterns of packets arriving at the server. Then, a machine-learning algorithm is utilized to differentiate whether the user connected is mobile or stationary. The proposed system has several use cases such as data collection and data mining, surveillance, public safety, network management, and load balancing.

II. MOTIVATION

Access to data networks from anywhere has become necessary in our lives today. The server-side of the application is usually clueless to whether the user of the application is mobile or stationary. Further, if a virtual private network is used it makes it very difficult for the server to determine the type of network used by examining the source IP address. This paper employs machine learning to determine the user's mobility status by utilizing the incoming traffic at the server-side. Knowledge of mobility status can be utilized for security and management applications. For example, limiting certain types of services to stationary users or ensuring that the appropriate kind of connection is employed for important processes such as system updates. Moreover, this can be used for data collection about users whether they use a service while mobile or stationary. Another application would be public safety. For example, if you suspect the user is driving, access to the service can be put on hold.

III. BACKGROUND

A typical objective of analysis of packet patterns is to gauge the quality of service of systems and networks [5][6]. Several research projects examine traffic patterns to extract new knowledge. For instance, a research project examined packets payload to determine the application sending and receiving these packets [7]. Another research project examined HTTP packets for the purpose of detecting attacks such as denial of service attacks [8]. Correspondingly, packet patterns analysis is employed to study user behavior [9]. Generally, there are three categories for packet analysis. The first is port-based analysis, where the port number of the TCP or UDP service is employed to specify the type of traffic [10]. The second category is payload analysis, where the packet's payload is inspected if not encrypted [11], or patterns are inspected if encrypted [12]. The last category is statistical [13], where the statistical patterns of traffic are examined.

Some of the traffic data analysis methods discussed above employ machine learning algorithms such as Bayesian [14] and Ripper [15]. Machine learning has been a paramount instrument to determining traffic types especially for the quality-of-service purposes, as

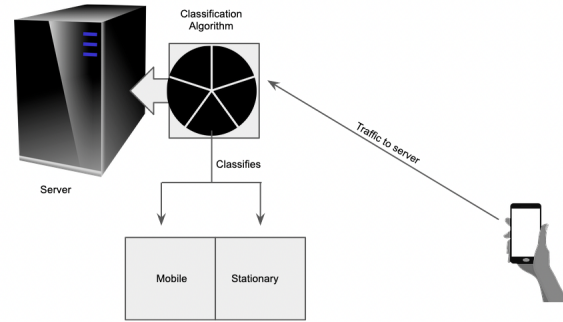


Fig. 1. Overview

detecting dropping nodes [16], or for security purposes as intruders' detection [17]. Several works [18][19][20] used machine learning to predict mobility at the network side. However, it was done using physical layer measurements, which are only accessible by the service provider. As seen from the discussed related work, most of the research projects discussed investigate traffic patterns to specify applications, measure the quality of service, or detect anomalies. Other works use machine learning on the mobile network side to determine the mobility of users. Our study is novel and the first of its kind. It is the first attempt to detect of user's mobility status at the server-side by employing the pattern of the arriving packets.

IV. METHODS

This research project goal is to allow servers to determine whether the user is mobile or stationary. To achieve that, the application server incorporates a machine learning model. This model classifies the incoming traffic and determines whether it is coming from a mobile or a stationary user. This is done by analyzing the arrival patterns of the received packets. The proposed system architecture is illustrated in Fig. 1. As the figure shows, the user sends data through the network medium in-use to the server. At server side, a classification algorithm analyzes the pattern of the incoming packets then classifies this traffic as traffic coming from a connection with a mobile user or from a connection with a stationary user. This work can be combined with our previous work [21], which differentiates between users using fixed(e.g., fiber or cable) networks from users using mobile networks. If the traffic appears to be coming from a mobile network, using the same data set, the classifier above would be able to determine whether the user is mobile or stationary.

As discussed above, the server incorporates a classification model that is used to determine the user's

mobility status using a number of received packets. Supervised machine learning is used to train this model. A label is attached to the training data. The labeled training data includes the following attributes:

- 1) The number of packets received in a window of five seconds.
- 2) The differential time of arrival of the packets received during this time window. The differential time of arrival is the time of arrival of a packet subtracted from the time of arrival of the previous packet. The differential time of the arrival measured in seconds.
- 3) The packet length in bytes. Packet length is usually adjusted based on the bandwidth and channel condition.

Diverse sets of data ought to be utilized in training to improve the classification model accuracy. Thus, improving the overall robustness of the system. A diverse data sets include data from various locations and various mobile carriers. The following section unveils more information about the data used in the training of the classification model. Further, it discusses the classification algorithm used, the test data, and the test results of the trained classification model.

V. EXPERIMENT AND RESULTS

To deploy a testing environment for the proposed system, a cloud Linux server was used. On this server, an HTTP server was installed and configured using Apache. On the HTTP server a large file of the size of 1 Gigabytes was hosted. This file was downloaded by a user using three different mobile network service providers following the scenario below:

- 1) Three hours of data collected while downloading the file discussed above while the data collector is driving at various speeds.
- 2) Three hours of data collected while the data collector is stationary at 9 different locations.

During the time the file is being downloaded generated traffic packets were recorded and monitored at the server. A total of six hours of data collection was done. Packets were collected at the server using Tshark[18].

The collected data is then preprocessed. In the pre-processing, incomplete entries and inconsistent rows are removed. Each five seconds of data was combined in a single record. The number of packets arrived during the five seconds was added to the record. One file combined the recorded data. The file was formatted to include the reflect three types of attributes discussed in the previous section. Then, the records in the file were labeled to define the data class. The collected data was divided into two sets. The first set contains 80 percent of the original data (3,456 records), was used to train the

classification model. The classification model employed the random forest algorithm. After training the model was tested using the second set of data, which consists of 20 percent of the data (864 records), to determine the accuracy of the trained model. The test results. are as shown in Table I.

TABLE I
SUCCESS RATE

N = 864	Actual	
	stationary	Mobile
True	TF = 47.2%	TM = 45.4%
False	FF = 4.1%	FM = 3.3%

In the table above, the confusion matrix displays the accuracy-test results. True stationary(TF) means when the model predicted that the user is stationary and the user was actually stationary. False stationary (FF) means when the model predicted falsely that user is stationary. True Mobile (TM) reflects the times the model predicted the user to be moving and the user was actually mobile. False Mobile(FM) reflects the times the model predicted the user to moving but the user was stationary. Equation 1 below is used to calculate the validation success rate.

$$\frac{True\ Positive + True\ Negative}{N, Total\ number\ of\ a\ dataset} \times 100 \quad (1)$$

As a result, A validation success rate of 92.6% is provided by the proposed scheme.

VI. CONCLUSION

The availability of network access when mobile is becoming essential today. User can access these network while is stationary, moving, or moving at higher speed when driving . Different user mobility status entails different forms of traffic pattern.

A machine learning model at the server-side is employed by this paper to enable the server to distinguish between data generated from a connection with a mobile user and data generated from a connection with a stationary user. This model was trained through supervised training. The test result of the model showed an accuracy of 92.6 percent.

The work in this paper is novel and the first of its kind. It is the first attempt to detect the user’s mobility status at the server side by utilizing the arrival patterns of traffic packets. Machine learning was employed at the server-side to obtain new knowledge. Such knowledge can be utilized to data mine the users’ behavior. Further, it can be employed in applications such as safety and security management. From this point, s next step would be attempting to differentiate between traffic transmitted using different internet service providers. Likewise,

differentiating between traffic transmitted by different mobile networks protocols (e.g., 3G, LTE, 5G).

REFERENCES

- [1] DataReportal. 2022. "Digital Around the World" Retrieved from <https://datareportal.com/global-digital-overview>
- [2] Juniper research. 2021. "Internet of Things' Connected Devices to Triple by 2021, Reaching Over 46 Billion Units" Retrieved from <https://www.juniperresearch.com/press/internet-of-things-connected-devices-triple-2021>
- [3] Claus Hetting. 2018. "New numbers: Wi-Fi share of US mobile data traffic lingers at around 75 percent in Q2". Retrieved from <https://wifinowglobal.com/news-and-blog/new-numbers-wi-fi-share-of-us-mobile-traffic-lingers-at-around-75/>
- [4] Anwasha Mukherjee, Debashis De, Priti Deb. 2018. "Power consumption model of sector breathing based congestion control in mobile network". Digital Communications and Networks, Volume 4, Issue 3., Pages 217-233.
- [5] Sumaira Mustafa, Khubaib Amjad Alam, Bilal Khan, Muhammad Habib Ullah, and Pariwish Touseef. 2019. Fair Coexistence of LTE and WiFi-802.11 in Unlicensed Spectrum: A Systematic Literature Review. In Proceedings of the 3rd International Conference on Future Networks and Distributed Systems (ICFNDS '19). Association for Computing Machinery, New York, NY, USA, Article 37, 1–10. DOI:<https://doi-org.liblink.uncw.edu/10.1145/3341325.3342031>
- [6] Zafar Ayyub Qazi, Melvin Walls, Aurojit Panda, Vyas Sekar, Sylvia Ratnasamy, and Scott Shenker. 2017. A High Performance Packet Core for Next Generation Cellular Networks. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17). Association for Computing Machinery, New York, NY, USA, 348–361. DOI:<https://doi-org.liblink.uncw.edu/10.1145/3098822.3098848>
- [7] Yagi, Shinnosuke & Waizumi, Yuji & Tsunoda, Hiroshi & Jamalipour, Abbas & Kato, Nei & Nemoto, Yoshiaki. (2008). Network Application Identification Using Transition Pattern of Payload Length. 2633 - 2638. 10.1109/WCNC.2008.462.
- [8] Zafar Ayyub Qazi, Melvin Walls, Aurojit Panda, Vyas Sekar, Sylvia Ratnasamy, and Scott Shenker. 2017. A High Performance Packet Core for Next Generation Cellular Networks. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17). Association for Computing Machinery, New York, NY, USA, 348–361. DOI:<https://doi-org.liblink.uncw.edu/10.1145/3098822.3098848>
- [9] M. Mimura and H. Tanaka. "Behavior Shaver: An Application Based Layer 3 VPN that Conceals Traffic Patterns Using SCTP," 2010 International Conference on Broadband, Wireless Computing, Communication and Applications, 2010, pp. 666-671, doi: 10.1109/BWCCA.2010.152. Karagiannis, Thomas; Broido, Andre; Brownlee, Nevil; Claffy, K.C. and Faloutsos, Michalis, Is P2P dying or just hiding?, IEEE Global Telecommunications Conference, November 2004.
- [10] Applications, 2010, pp. 666-671, doi: 10.1109/BWCCA.2010.152. Karagiannis, Thomas; Broido, Andre; Brownlee, Nevil; Claffy, K.C. and Faloutsos, Michalis, Is P2P dying or just hiding?, IEEE Global Telecommunications Conference, November 2004.
- [11] A. Madhukar, C. Williamson, A longitudinal study of p2p traffic classification, in: MASCOTS '06: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation, IEEE Computer Society, Washington, DC, USA, 2006, pp. 179–188. doi:<http://dx.doi.org/10.1109/MASCOTS.2006.6>.
- [12] T. Karagiannis, K. Papagiannaki, M. Faloutsos, BLINC: multilevel traffic classification in the dark, in: SIGCOMM '05: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, ACM Press, New York, NY, USA, 2005, pp. 229–240.
- [13] J. Erman, M. Arlitt, A. Mahanti, Traffic classification using clustering algorithms, in: MineNet '06: Proceedings of the 2006 SIGCOMM workshop on Mining network data, ACM Press, New York, NY, USA, 2006, pp. 281–286
- [14] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005, Banff, Alberta, Canada, June 2005.
- [15] R. Alshammari and A. N. Zincir-Heywood, "Machine learning based encrypted traffic classification: Identifying ssh and skype," in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, July 2009, pp. 1-8.
- [16] N. J. Patel and R. H. Jhaveri, "Detecting packet dropping nodes using machine learning techniques in Mobile ad-hoc network: A survey," 2015 International Conference on Signal Processing and Communication Engineering Systems, 2015, pp. 468-472, doi: 10.1109/SPACES.2015.7058308.
- [17] K. Park, Y. Song and Y. Cheong, "Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm," 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), 2018, pp. 282-286, doi: 10.1109/BigDataService.2018.00050.
- [18] H. Gebrie, H. Farooq and A. Imran, "What Machine Learning Predictor Performs Best for Mobility Prediction in Cellular Networks?," 2019 IEEE International Conference on Communications Workshops (ICC Workshops), 2019, pp. 1-6, doi: 10.1109/ICCW.2019.8756972.
- [19] J. Jeong et al., "Mobility Prediction for 5G Core Networks," in IEEE Communications Standards Magazine, vol. 5, no. 1, pp. 56-61, March 2021, doi: 10.1109/MCOMSTD.001.2000046.
- [20] Ozturk, Metin & Gogate, Mandar & Onireti, Oluwakayode & Adeel, Ahsan & Hussain, Amir & Imran, Muhammad. (2019). A novel deep learning driven low-cost mobility prediction approach for 5G cellular networks: The case of the Control/Data Separation Architecture (CDSA). Neurocomputing. 358. 10.1016/j.neucom.2019.01.031.
- [21] H. Alamleh, K. Waters and B. Al Smadi, "Server-Side Distinction of Incoming Traffic Transmission Medium Using Machine Learning," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0482-0485, doi: 10.1109/UEMCON53757.2021.9666729.
- [22] Tshark. 2021. "The Wireshark network analyzer". retrieved from <https://www.wireshark.org/docs/man-pages/tshark.html>

Data Centric DAO: When blockchain reigns over the Cloud

Ibrahim MEHDI*, Moussaab SBAI†, Mohamed MAZLIN‡ and Kamal AZGHIU§

Data Science & Cloud Computing branch,

École Nationale des Sciences Appliquées Oujda (ENSAO),

UMP - Université Mohammed Premier Oujda (UMP)

Email: *ibrahim.mehdi@ump.ac.ma, †moussaab.sbai@ump.ac.ma, ‡mohamed.mazlin@ump.ac.ma, §k.azghiou@ump.ac.ma

Abstract—Nowadays, data has become more and more important. Most governments have passed laws to control how they are owned and used, such as the GDPR in Europe. However, any law is never perfect. We can see it when a company scattered around the world is hampered by the fact of not being able to exploit customer data outside the borders of a territory. Also, we can add that a data owner is not anymore once he sells them to a third party to be exploited. In this work, we propose a solution based on a permissioned blockchain, namely the Hyperledger Fabric, to allow any stakeholder with data to take part, with value providers, in creating Decentralized Autonomous organizations (DAOs). Once created, a DAO can attract other investors to expand it. The Hyperledger Fabric handles this whole process through channels. The infrastructure needed to run the business of the DAO is generated automatically through special transactions taking place at the blockchain level. In the opposite direction, the cloud infrastructure sends notifications to the Hyperledger for traceability and monitoring purposes. Finally, through simulations of wealth distribution models, we show that to keep control of a DAO based on the proposed architecture, the shares must be negotiated according to a Pareto law.

Index Terms—Data-Marketplace, Cloud, Blockchain, Hyperledger Fabric, IoT, DAO, Secure multi-party computation, Data.

I. INTRODUCTION

We will remember 21st century technology for its data-driven gold rush. In fact, the amounts of data do not decrease but increase exponentially. It all started when Sir Tim Berners Lee invented hyperlinks (WWW), which gave life to data sharing around the world. Today, more and more people are using the Internet and new services are appearing, thus generating even more many and diversified data. Nowadays, companies can provide personalized content to each individual user.

Currently, IoT infrastructures generate a large part of the data in the world [1]. Large enterprises are competing to provide the next generation of smart devices to provide their customers with a new lifestyle, as well as to provide enterprises with real-time monitoring of their business operations [2]. Huge futuristic projects are being carried out thanks to the huge amount of connected smart sensors. Smart cities are the perfect example of this, as we are not just

envisioning but seeing it solving many problems in urban areas [3]. Sectors such as transport, energy, networks of networks up to public administration all benefit from these solutions. Regardless of the IoT-based system, the underlying infrastructure will always generate a dataset to be mined.

But if you get six good things about the data, you have half a dozen challenges around them. Let us cite the enormous storage capacity necessary to keep the data with a view to their exploitation and/or archiving (Big Data paradigm) and the legal aspect of their exploitation which has sounded the alarm of many organizations in order to protect those of privacy which led them to publish laws that dictate their collection, storage and use [4]. However, these laws do not sufficiently respect the economic aspect of the exploitation of data given the absence of mechanisms allowing their use on a large scale while preserving their private nature. Whichever party may regulate the use of data, whatever the nature of the data, it must not impede technological advances by the measures it imposes.

It is in this context that we introduce in this work a solution that gives more rights to the owner of the data. Indeed, deciding how to exploit the data by giving more control to its owner and guaranteeing him the most interesting return on investment possible are the essential objectives of this work. To prevent the use of data without the knowledge of their owner, we offer a solution based on the encryption of the latter using a secure multi-party calculation. The blockchain makes it possible to keep traces, which are very useful in the event of disputes, thanks to the concept of immutable registers. [5]. To preserve the rights of data users, the architecture proposed in this work stipulates an organizational structure according to the paradigm of Decentralized and Autonomous Organizations (DAO). In fact, a DAO is created as soon as a data owner and a value provider agree. New value provider can join an already created DAO by redistributing the benefits and updating its version. The process repeats itself until the DAO stabilizes on a state of equilibrium which may be due, for example, to an unattractive return on investment for new entrants.

We structure this paper as follows: The section II summarizes the technologies used in our system. The section III outlines some of the related work and explains how our approach is unique. We present the proposed architecture and

its components in the section IV. Next, we show how the stakeholder in our system interact with each other in the section V. In the section VII, we cover some aspects of our architecture, along with an explanation of the benefit sharing process. Finally, we end the article with some benefits, general uses of our system, and future work that we can and will implement in our system.

II. BACKGROUND

In this section, we are presenting an overview about various technologies that are the building blocks for our system.

A. Hyperledger Fabric

1) *Overview:* Hyperledger Fabric V1.0 was released by the Linux Foundation in 2017. It is an open-source permissioned blockchain framework. It came to life to provide a secure, scalable and flexible groundwork for industrial blockchain solutions. Fabric got rid of the mining process and made sure that access to the data in the ledger is only to authorized members by creating subnets in the network called Channels. Peers can join a channel (or multiple channels at the same time) by enrolling through MSP (Membership Service Provider) and therefore it will have its ledger and chaincodes (smart contracts) installed. Consensus is achieved in Hyperledger Fabric after three phases: Endorsement, Ordering and Validation.

- **Endorsement phase:** Endorsing peers will simulate and execute transactions in an isolated environment and then either sign it as endorsed or not. The result is sent back to the transaction initiator.
- **Ordering phase:** Ordering service (also called ordering service nodes) will receive the transaction and the endorsement signatures and determine the order of transactions.
- **Validation phase:** Validating the authenticity and correctness of blocks of transactions

2) *Events in Hyperledger Fabric:* There are three sorts of events that can be subscribed to in Hyperledger Fabric:

- **Block events:** Events that are set automatically after committing a block.
- **Transaction events:** Also set automatically after committing transaction.
- **Contract events:** Events explicitly added to the chaincode and is set with the contract invocation.

By listening for these events, the application can respond without having to initiate a transaction. In our system, we will use the contract events by setting up an event listener and handler that will allow us to utilize data included in these notifications to automatically command and control an infrastructure.

B. Secure Multi-party computation

The idea of outsourcing data processing and computing without handing over the keys to it is the basis of our system. We know that fully homomorphic encryption [7], [8], allows such privacy, but we also know that it is fully unpractical. It basically goes like this:

Alice encrypts her data $AliceData$ and sends it $E_{Alice}(AliceData)$ to Bob, does his computation $f(E_{Alice}(AliceData)) = R$ and returns the encrypted result. Alice then decrypts $D_{Alice}(R)$ to find out the results. Multi-party computation [9] on the other hand is a scheme first developed in 1980's that aims to provide techniques allowing entities to collaboratively calculate a function while keeping their inputs private.

Many protocols are being developed to optimise further the performance of such systems. SPDZ [10], a universal multiparty computing protocol that is safe against up to $n-1$ of the n players being corrupted by an active attacker, is an interesting one that can be implemented to ensure data privacy.

C. Decentralized Autonomous Organization

A decentralized autonomous organization (DAO) is a collection of entities that work together using smart contracts [11]. This implies that all corporate processes, definitions, and restrictions are encoded in the blockchain and cannot be changed. This type of organization was inspired by the decentralized cryptocurrencies by not having any central authority controlling all the flow and risking both single points of failure and privacy issues. As a result, investors may now purchase DAO shares and get tokens that reflect their ownership in the organization and let them to vote on future initiatives. DAOs are established in our system when two parties agree to collaborate. It's worth noting that one or both of them might already be a DAO.

III. LITERATURE REVIEW

Numerous blockchain based data marketplaces have been introduced in the last few years, each with its own vision and architecture. In [12]–[15] they proposed a decentralized buy/sell architecture based on blockchain as the new solution to traditional data markets. It enforces the fair play between the parties by a punishment system for dishonest behaviours. Elimination of central trusted authorities so that owners of data can have control over it. The implementations of immutable ledgers and smart contracts enabled the users to browse and pick with who they want to work. And to ensure data privacy during the transaction so that no one but the authorized parties can see it, various approaches are presented; Selling the decryption key to an encrypted database [12], or by using swarms in Ethereum network as a decentralized storage for example [13].

In [16]–[18] systems are designed for IoT devices by providing a decentralized platform to sell data streams. These infrastructures are characterized by their low-resources, computational power and lack of security and privacy which requires some sort of middle-system that helps compensate for these weaknesses so they can serve multiple buyers at once. [19] took a step forward with what was already done. A mix of different technologies as hyperledger fabric as an immutable ledger, Cosmos [20] for token interoperability and Mainflux as an IoT gateway. The trade is done by exchanging tokens

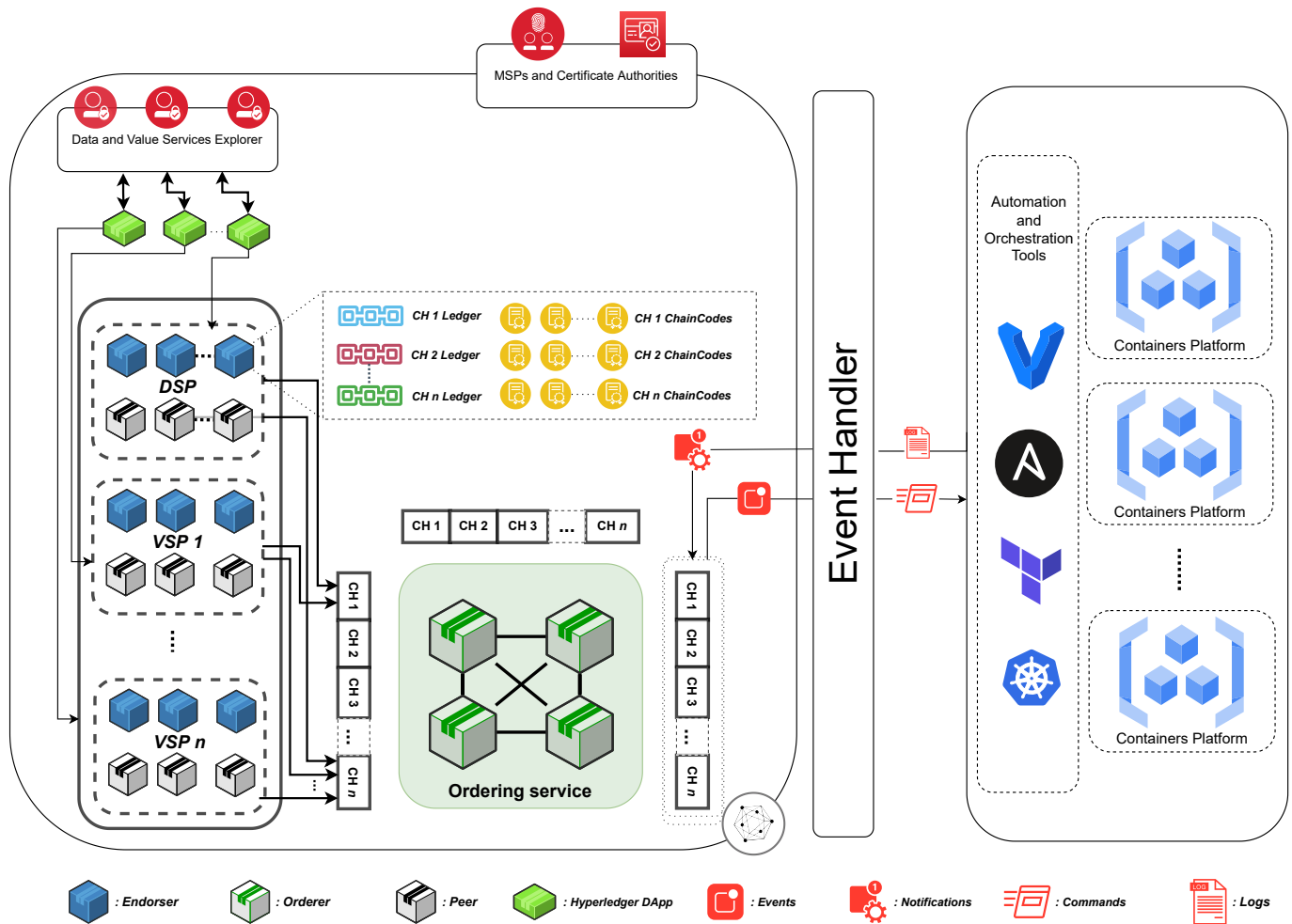


Fig. 1. Data ownership centric architecture

for a Proxied URL with an expiration date determined in the transaction.

We can see in these examples how they all provided a sales marketplace, and at some point the data will be at the other side in plain text with no proof of where it is and how it is being used. Our approach is different. Firstly, we are not interested in selling data as it has already been done before many times. We took the famous sentence ‘Data is the new Oil’ to the letter. Any person having an oil field can either sell it and be satisfied with the one time payment, which is not what people do, or can be part of the business and invest with his land. As for data, it is important to ensure privacy and anonymity in the exploitation phase. Secondly, we want a decentralized control and monitoring over the infrastructure where all the business is running through blockchain.

IV. DATA OWNERSHIP CENTRIC ARCHITECTURE

A. An overview of the proposed architecture

Here, we present an overview of the architecture and its components (Fig. 1). Several parts make up the proposed architecture. Namely, a Hyperledger Fabric module and a

Cloud Infrastructure, as well as an event handler facilitating communication between the two by converting Hyperledger events into commands executed by the Cloud and conveying notifications from the latter to the Hyperledger fabric to be registered IV-C. Our goal is to provide a framework for building data-centric DAOs in a non-trust environment. In this way, a data owner would own shares of the DAO concerned as tokens, allowing him to be a member shaping its evolution.

B. The hyperledger part

The Data and Value Services Explorer is where each service is exposed to the various interested parties, with a description containing all the details needed to conclude a business contract. At this level, the stakeholders sign up to benefit subsequently from a space for negotiation in order to launch new DAOs or to join one or more existing ones.

A business contract can be concluded either between one or more data service providers and one or more value providers, or between one or more value providers and an existing DAO. In case of consensus, the stakeholders create a

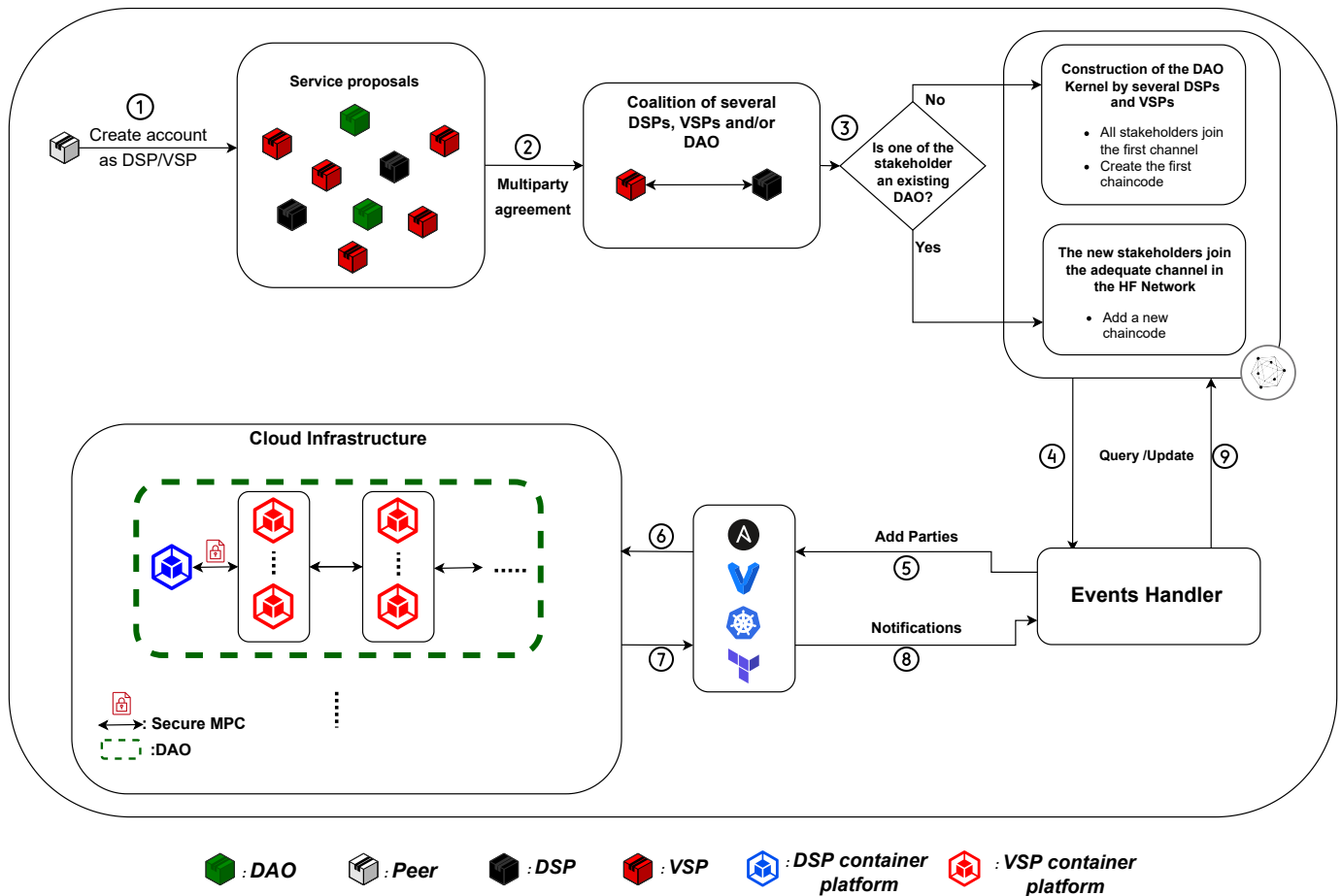


Fig. 2. Interactions between system actors

new DAO or they update the existing DAO.

A concluded contract is materialized by the creation and/or the addition of a new chain in the Hyperledger Fabric where the various transactions will be recorded immutably. Special transactions can auto-generate a cloud infrastructure reflecting the commitment in terms of resources of the stakeholders vis-à-vis the DAO of which they are members. Other types of transactions may concern either the monitoring of the evolution of the DAO or a transfer of rights between an active member and another internal to the DAO or with a new member.

Within each channel, there is a set of blockchain codes that specify the rules of the transactions within it. Then they will send several event notifications to the event handler to command the cloud infrastructure.

A data service provider is the owner of the data, the value provider is any party that can increase the value of the data by subjecting it to a set of operations, such as those relating to machine learning algorithms to derive models for decision making. So the more services above the data layer, the more

valuable the DAO.

Channels play a crucial role in maintaining confidentiality within DAOs. In fact, a data service provider DSP and a value provider VSP_1 which form a DAO will be assigned to a channel $CH_1 = \{DSP, VSP_1\}$. This will regulate the workflow using chaincode: profit sharing, approval policies, contract duration and other parameters regarding the business model and the infrastructure ordered. Any other VSP_2 party wishing to join the current DAO will join a new channel CH_2 with the CH_1 still running. Thus, after this last step, the active channels are: $CH_1 = \{DSP, VSP_1\}$, $CH_2 = \{DSP, VSP_1, VSP_2\}$. Each time a new value provider joins an existing DAO, this process will be repeated. In general, the channel CH_i will always have the form of the equation (1):

$$CH_i = \{DSP\} \cup \{VSP_j \mid 1 \leq j \leq i\}. \quad (1)$$

Any party in a given channel will have its ledger and channel codes installed, as well as subsequent channels. For example, the VSP_2 will have the registry and chaincodes from CH_2 up to CH_n . But not CH_1 since he is not a member. That said, we see how the data service provider will be at the heart of the DAO since it takes part in all channels.

C. The infrastructure part

Part of the transactions that take place at the Hyperledger Fabric level concern orders sent to the Cloud infrastructure. In fact, blockchain-level transactions should be able to automatically establish, update, and maintain cloud infrastructure dedicated to running the underlying DAO business processes.

It is by using automation tools [21] such as Vagrant [22], Ansible [23], Terraform [24], Kubernetes [25] that we can launch the instances working in a cloud environment. If a user wishes to work on his own machines, this will be defined in his contract. If no notification comes from his machine, the system assumes he abandoned the job and the contract will be terminated. In sec. VI we will define a future strategy for how to manage and create confidence in the market.

V. ARCHITECTURE COMPONENTS INTERACTIONS

A. A chart flow for the proposed architecture

We present in this section, see Fig. 2, the set of interactions between the system components, as well as the flow of actions inside the application.

① A user creates an account so he can be identified, then he selects his user type; Data service provider or a value service provider. Offers are posted on the distributed application with a specific description along with terms and conditions imposed by the original poster. These terms though can be changed as negotiations take place. Users browse and pick their desired partner for a specific job.

② Negotiations are a normal phase before any business agreement. Parties interested in each talk about their terms in order to find a mutual ground, each determines a stake which determines their portion of the DAO.

③ Once everything is set and done, the parties are now considered as one DAO. Now if one of the parties is an existing DAO, the new member joins them in the adequate channel. If not, a new DAO is created combining the two users.

④ Hyperledger fabric will be communicating with the cloud infrastructure using an event handler. It basically reads chaincode events and acts accordingly. This allow us to keep track of all the instances created.

⑤ The event handler should have converted events to script files for our automation tools by this step like Ansible, vagrant, kubernetes and terraform. It will give us the ability to systematise the infrastructure orchestration. ⑥ In this step, the script files are executed and everything is in place. The business will be running inside the cloud infrastructure. ⑦ Frequently, notifications about the big picture of the infrastructure will be sent back into the blockchain so that all members can monitor the flow of work.

⑧ The event handler will be the one converting these notification into queries.

⑨ Queries received are executed and new blocks are created reflecting the current state of the cloud infrastructure.

B. Data privacy

As in [26], a combination of Homomorphic encryption and multiparty computation will allow us to achieve a privacy-preserving framework to make our system even more data centric. It will enable our system to keep the data encrypted all the way through the computations and still have valuable results out of it.

VI. SOME ASPECTS OF THE PROPOSED ARCHITECTURE

As previously stated, our main goal is to make data driven projects preserve the right of privacy with no middle trusted authority. Data will never be decrypted during the whole process. Data owners will be able to make a profit from the final project as our vision is not to sell but to invest with data. One example of the possible business model will be a pay per use or subscription based, it all depends on the contract between the providers.

IoT infrastructures will rely significantly on our technology to offer a wide range of data to other parties for analysis and use while maintaining data privacy.

In terms of profit splitting, our system distributes profit π group wise and according to each party's stake as defined in the chaincodes of each channel. See Fig. 3 for a detailed tree graph describing the ownership percentage in a given DAO. The (i) in $DAO^{(i)}$ describes the version of the DAO created in channel CH_i . In order to compute the profit of a specific $DAO^{(i)}$ or VSP_i , we use the equations 2 and 3.

$$\pi[DAO^{(i)}] = \pi[DAO^{(n)}] \times \prod_{k=i+1}^n \beta_k \quad (2)$$

$$\pi[VSP_i] = (1 - \beta_i) \times \pi[DAO^{(i)}] \quad (3)$$

For stakeholders to build or integrating a DAO, the various stakeholders must agree on ratios β_i for the distribution of wealth. This value is used to calculate a member's share of total profit.

To visualize the evolution of the shares of the data owner as well as those of the value providers integrating the DAO step by step, we present in Fig. 4 and Fig. 5 graphs for two different scenarios: (i) In the first scenario we let's simulate the β_i from a uniform law, (ii) In the second scenario the β_i are taken from the equation 3 on which we have imposed a distribution of wealth according to the law of Pareto. We inject the deducted β_i into the equation 2 to generate the profits, as shown in the Fig. 5.

1) *Randomly selected values:* Creating β_i with this method is straightforward. Generate uniformly distributed and sorted values in an interval. We started with 25 – 99% then we increased it gradually and noticed that the value of all β_i should be at least 70% (in our example of 7 VSPs) in order to keep the Data Owner at the top of the pile(see Fig. 4). We also notice how the VSPs will have more leverage if they join a job at the end of it. This will make VSPs refuse to join works

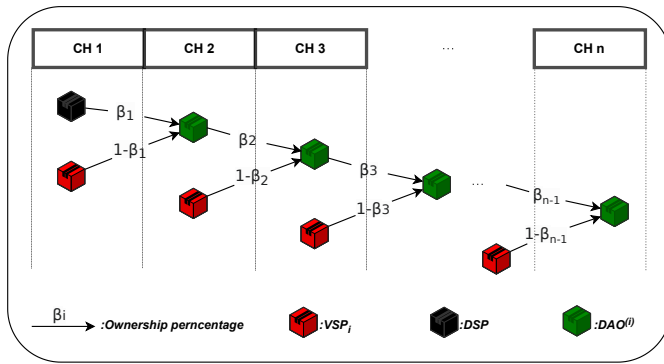


Fig. 3. Ownership percentages of each member in corresponding DAO

in its beginning. It can be solved by changing the range of β_i 80 – 99% but it will bring more problems than it solves as the DSP would monopolise the whole business. In this approach, we created β_i then analyzed the total ownership percentage.

2) *Selecting values according to Pareto Principle:* The Pareto principle [27], named after Vilfredo Pareto, an Italian economist and sociologist. It was developed to describe the distribution of majority of wealth in the hands of a top percentage of the population. It has many uses in multiple areas besides economy as insurance, manufacturing, management and many more. We will be using it to generate numbers following the distribution in question representing the total ownership percentage.

Algorithm Generate β_i fitting the total ownership to a Pareto distribution

Require: $n \geq 2$ ▷ Number of β_i

- 1: Generating n random numbers following Pareto’s distribution.
- 2: Scale (between 0 and 1 to describe percentage) then sort descending these numbers.
- 3: Assign each value with corresponding member starting with DSP and the highest value. ▷
Now we have for each member his total percentage. We can calculate β_i by reversing formulas 3

As we can see in Fig. 5, the percentage of total ownership drops down as we add more VSPs. This is a better result than what we got before. It is logical that as long as we add more VSPs, the value of the DAO will increase which leads to higher share prices. Contrary to what we did on the first one, in this approach we defined the total ownership percentage then deduced the β_i .

VII. CONCLUSION

Our system presents a functioning framework that allows anyone, and especially IoT infrastructure owners to invest their data in a DAO amongst other service providers to further increase the value extracted from that data. This will benefit IoT device manufacturers since our future consists of more gadgets and sensors all over the cities and infrastructures,

analytics services by having broader data point thus reducing the effects of ‘small dataset curse’ also known as overfitting, having real-time data to stay up to date. And the end users of course that will use the final service. Our motivation is to have always our data protected but still making it work in the real world to improve services, applications and overall user experience. As a future development, a user reputation system may be implemented. Parties can offer feedback on their interactions with a specific user by including a quality of experience mechanism, which will help us sanction poor conduct. This will boost user confidence, which is an important component of the online experience.

REFERENCES

- [1] Keyur K Patel, Sunil M Patel, et al. Internet of things-iot: definition, characteristics, architecture, enabling technologies, application & future challenges. *International journal of engineering science and computing*, 6(5), 2016.
- [2] Manlio Del Giudice. Discovering the internet of things (iot) within the business process management: A literature review on technological revitalization. *Business Process Management Journal*, 2016.
- [3] Veronica Scuotto, Alberto Ferraris, and Stefano Bresciani. Internet of things: applications and challenges in smart cities. a case study of ibm smart city projects. *Business Process Management Journal*, 2016.
- [4] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.
- [5] Massimo Di Pierro. What is the blockchain? *Computing in Science & Engineering*, 19(5):92–95, 2017.
- [6] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference*, pages 1–15, 2018.
- [7] Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, et al. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *IEEE Access*, 10:30039–30054, 2022.
- [8] Craig Gentry. *A fully homomorphic encryption scheme*. Stanford university, 2009.
- [9] Oded Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78:110, 1998.
- [10] Ivan Damgård, Marcel Keller, Enrique Larraia, Valerio Pastro, Peter Scholl, and Nigel P Smart. Practical covertly secure mpc for dishonest majority—or: breaking the spdz limits. In *European Symposium on Research in Computer Security*, pages 1–18. Springer, 2013.
- [11] Alexandra Sims. Blockchain and decentralised autonomous organisations (daos): The evolution of companies? 2019.
- [12] Matias Travizano, Carlos Sarraute, Gustavo Ajzenman, and Martin Minnoni. Wibson: A decentralized data marketplace. *CoRR*, abs/1812.09966, 2018.
- [13] Kazim Rifat Özyilmaz, Mehmet Doğan, and Arda Yurdakul. Idmob: Iot data marketplace on blockchain. In *2018 crypto valley conference on blockchain technology (CVCBT)*, pages 11–19. IEEE, 2018.
- [14] Hyunkyung Yoo and Namseok Ko. Blockchain based data marketplace system. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1255–1257. IEEE, 2020.
- [15] Prabal Banerjee and Sushmita Ruj. Blockchain enabled data marketplace - design and challenges. *CoRR*, abs/1811.11462, 2018.
- [16] Pooja Gupta, Volkan Dedeoglu, Salil Kanhere, and Raja Jurdak. Towards a blockchain powered iot data marketplace. pages 366–368, 01 2021.
- [17] Ahmed Suliman, Zainab Husain, Menatallah Abououf, Mansoor Alblooshi, and Khaled Salah. Monetization of iot data using smart contracts. *IET Networks*, 8(1):32–37, 2019.
- [18] Krešimir Mišura and Mario Žagar. Data marketplace for internet of things. In *2016 International Conference on Smart Systems and Technologies (SST)*, pages 255–260, 2016.

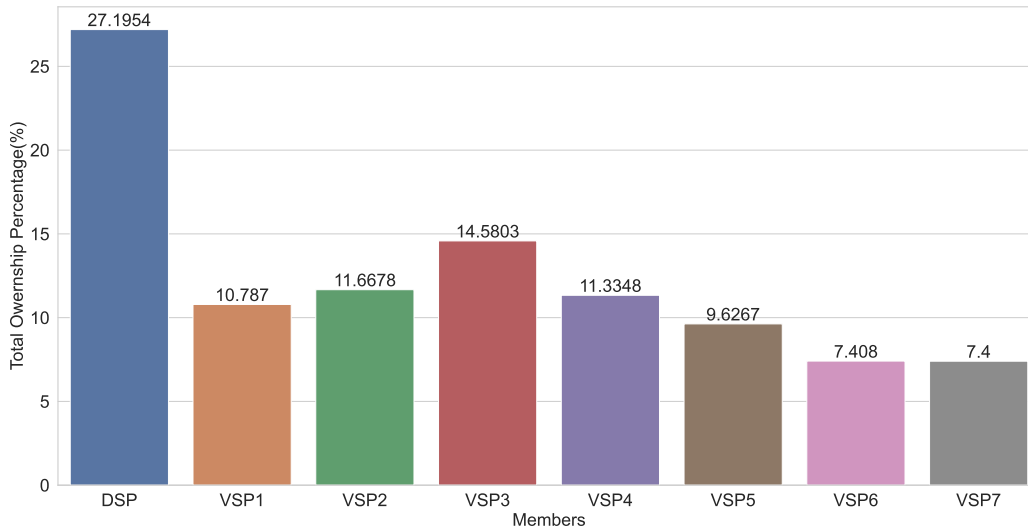


Fig. 4. Total Ownership percentage calculated with uniformly random β_i values using formula 3. Results may vary with randomness in play, but doesn't change the inefficiency of such distribution method.

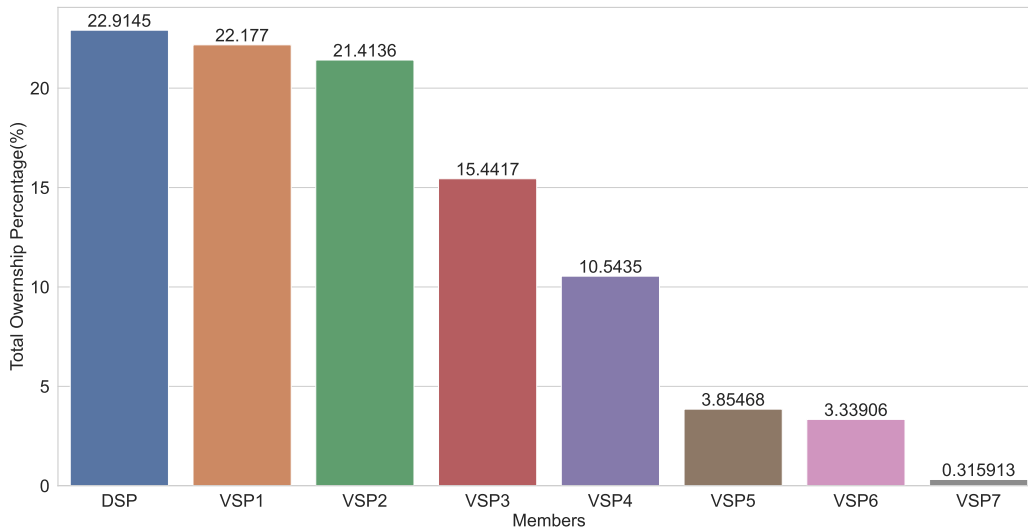


Fig. 5. Total Ownership following the Pareto distribution approach. A remarkable consistency and fairness using β_i fitting to that distribution.

[19] Drasco Draskovic and George Saleh. Datapace-decentralized data marketplace based on blockchain. *Datapace*, 2017.

[20] Jae Kwon and Ethan Buchman. Cosmos whitepaper. *A Netw. Distrib. Ledgers*, 2019.

[21] Loic Houde, Daniel Jacob, Tovo Rabemanantsoa, and Jean-François Rey. *Gestion Automatique d'Environnement Virtuel (GAEV)*. PhD thesis, INRAE, 2021.

[22] Mitchell Hashimoto. *Vagrant: up and running: create and manage virtualized development environments*. " O'Reilly Media, Inc.", 2013.

[23] Lorin Hochstein and Rene Moser. *Ansible: Up and Running: Automating configuration management and deployment the easy way*. " O'Reilly Media, Inc.", 2017.

[24] Yevgeniy Brikman. *Terraform: up & running: writing infrastructure as code*. O'Reilly Media, 2019.

[25] Brendan Burns, Joe Beda, and Kelsey Hightower. *Kubernetes: up and running: dive into the future of infrastructure*. O'Reilly Media, 2019.

[26] Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. Multi-party computation from somewhat homomorphic encryption. In *Annual Cryptology Conference*, pages 643–662. Springer, 2012.

[27] Rosie Dunford, Quanrong Su, and Ekraj Tamang. The pareto principle. 2014.

Algorithmic Analysis of the System Based on the Functioning Table and Information Security

1st Anvar Kabulov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 anvarkabulov@mail.ru

2nd Inomjon Yarashov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 timprivate345gmail.com

3rd Alisher Otakhonov
Faculty of Mathematics and Informatics
Fergana State University
 Fergana, Uzbekistan
 alisherotaxonov91@gmail.com

Abstract—Systems are ideally and rapidly evolving with the development of the Internet and modern web technologies. Example, Ecological monitoring information system has become very important in determining its superiority as a system, so it should be tested. This article describes how to benchmark a system by creating a state change graph to simulate different intuitions on the part of each subsystem, and then comparing the graph to Petri net modeling based on Functioning table (PNMFT). Test cases can be prepared based on the accessibility of trees based on comparable PNMFTs. The existing fact-testing strategy for comprehensive testing of the system will be further generalized to improve the consistent quality of the remaining tests. After that, talk about algorithmic analysis of the system is an important for information security issue. The initial stage of ongoing research to improve the reliability of algorithmic analysis of algorithmic formalization of web systems is described. It may also be advisable to include a set of dynamic coverage criteria for web systems defined for the statistical testing approach used through algorithmic analysis to improve testing quality and algorithmic formalization of the web system through algorithmic analysis.

Index Terms—algorithmic analysis, functioning table, information security, ecological monitoring, algorithmic formalization.

I. INTRODUCTION

The rapidly evolving information and communication technologies are making significant changes in all aspects of our daily life [1], [39]–[42], [44]. Nowadays, the concept of “information” is often used as a special trademark that can be bought, sold, exchanged for another product. Moreover, the cost of information [5]–[9], [43] often exceeds the cost of the computer system in which it is located, several hundred and thousand times. Consequently, there is an urgent need to protect information from unauthorized access, deliberate alteration, theft, loss and other criminal activity. However, the desire of society for a high level of automation makes it dependent on the level of security of the information technologies used [10]–[16].

The importance of information has been known since ancient times. Therefore, in ancient times, various methods were used to protect information [17]–[22], [45]. One of them is a mysterious inscription. No one but the addressee could

read the message. For centuries, this art - the mysterious writing - has not gone beyond the upper strata of society, the residences of state embassies and intelligence missions. Just a few decades ago, everything changed dramatically, that is, information acquired value and became a mass product. Now it is produced, stored, transferred, sold and bought. In addition, it has been stolen, distorted and falsified. Thus, there is a need to protect information.

Not only computers, global and corporate networks (Internet), but also modern, highly efficient means of information technology [23]–[26], [46] and processing of large amounts of information, such as statistical and financial institutions, were selected as the area of computer crimes.

Therefore, it is impossible to imagine the activities of any organization without the use of manual or computer tools for obtaining a variety of information, obtaining certain decisions as a result of the analysis of information and its transmission through communication channels. The computer can be viewed as an object of aggression and as an instrument of aggression.

The complex of measures for information security [27]–[30], [47] should be based on an information protection strategy. It defines the objectives, criteria, principles and procedures needed to build a robust security system.

An important feature of the general information protection strategy is the prohibition of the security system. There are two main areas:

- Analysis of guarantees;
- Detect an attack.

The second problem in the information security hierarchy is policy definition. Its content consists of the most rational means and resources, the goal of the problem and its approach.

This document consists of several pages of text that form the physical architecture of the network, and the information in it determines the choice of security product.

When developing an information security policy, first of all, the protected object [31]–[36], [48] and its functions are determined. Then the level of the enemy’s interest in the object, the possible types of attacks and the damage inflicted are assessed. Finally, the weak points of the object

are identified that do not provide adequate protection against the available countermeasures.

II. LITERATURE REVIEW

Algorithmic analysis methods for rational linear hybrid automata have been used in such tools as HYTECH and POLKA. In addition, when used on a secure system, the inspection procedures of these tools are used to detect security property violations [2]. Provide a way to bridge models at the computational and algorithmic levels and define levels of analysis between computation and algorithm. Often, computational level models involve a deeper introduction of the applied concept of rationality to the algorithmic level [3]. Rational analysis is taken to the algorithmic level [4]. Algorithmic stability was used to analyze the adaptive data [5]. Kuhn-Munkres parallel genetic algorithm and its application to algorithmic analysis of large-scale wireless sensor networks [6].

III. MATERIALS AND METHODS

Individual elements of a Petri net (positions and transitions) can have different properties, on the basis of which the properties of the nets themselves are first determined, and then their classification is built.

The simplest property of a position is the number of labels that can be placed in it. If in any reachable markup the number of labels in a given position is at most one (0 or 1), then such a position is called safe. A Petri net is said to be safe if all its position are safe. In secure networks, the state of each location is described by just one bit, so such networks can be easily implemented in hardware using certain types of switches (triggers). By the way, the original version of the definition of a Petri net, given by Adam Petri himself, just implied that the net is secure. However, for most applications, the requirement for network security is overly stringent. It can be weakened by allowing each location to store some limited number of tags. More strictly, a position is called k-bounded if in any reachable markup at a given position there are at most k labels. Obviously, 1-limited space is safe. A position is called restricted if there is such k such that this positions is k-restricted. Finally, a Petri net is k-bounded if any of its positions are k-bounded, and simply bounded if all of its positions are bounded. Bounded networks also allow for an efficient hardware implementation, in which each location is represented by a counter (for example, a register) of some given capacity. Unlimited networks are, as a rule, of only theoretical interest.

Another property of Petri nets, based on counting the number of labels, is the property of conservatism. A network is called conservative if the number of labels in any reachable markup is kept the same (equal to the number of labels in the initial markup). Such a model is used, for example, in cases where labels represent some system resources that are not destroyed or created. These resources can move from one part of the system to another, but their total amount does not change during the operation of the system. It is easy to show

that any transition that occurs in at least one reachable markup must have the same number of input and output arcs - as many marks he chose, he must put them as many.

A. Formalization of the Functioning table through the Petri net

Petri nets for the Functioning table are classical models for modeling systems that demonstrate parallelism, synchronization and randomness. A Petri net based on a Functioning table is an oriented bipartite graph with two types of nodes called position and transitions. Nodes are connected by directional arcs. Connections between two nodes of the same type are not allowed, such as a simple Petri net. In the graphical representation, positions are represented by circles, and transitions are represented by stripes or rectangles [37], [38].

The position P is called the input transition position t if there is a directional arc from P to t , P is called the output position t if there is a directional arc from t to P . A transition is termed enabled if every of its input positions includes "enough" tokens. The enabled transition may work. Starting the transition t means consuming tokens from input positions and creating tokens for output positions.

Arcs are marked by their weight (positive integers), where a k-weighted arc can be interpreted as a set of k parallel arcs. A marked-up Petri net is shown as an algorithmic formalization.

The markup of the Petri net is a vector of natural numbers $M = (m_1, \dots, m_n)$, where n is the total number of positions. M_i represents the number of tokens in position P_i .

The markup M_n is said to be achievable from the initial markup M_0 if there is a sequence of launches that converts M_0 to M_n in the Functioning table. Then, based on it, it is possible to determine the sequence of markings $\{M_0, M_1, M_2, \dots\}$ and the sequence of transitions $\{t_1, t_2, \dots\}$. The reachability set $R(M_0)$ of the Petri net is the set of all labels achievable from $M(M_0)$ with an algorithmic description. The set of reachability of a Petri net can be algorithmically formalized by a tree. The tree ceases to expand if algorithmic formalization of the system with previously defined markup is achieved. Such a reachability tree can be used to search for dead-end marks in algorithmic analysis, that is, marks in which the transition is not allowed.

IV. RESULT AND DISCUSSION

A. Statistical testing through algorithmic analysis of the web system

This article describes how to use a Petri net model based on a Functioning table generated from a web system to obtain test cases that meet the criterion of all facets. This section summarizes the current statistical testing method for all-k-path and all-edge testing of web systems to improve the reliability of structural testing through algorithmic analysis.

An algorithmic analysis method based on a function table is used to define the specification of a sequence of methods, while this approach is to define the structural aspect of a system as a web system in terms of pathways. The algorithmic analysis method includes the following steps:

TABLE I

28 PATHS OF LENGTH ≤ 7 , CORRESPONDING TO THE REACHABILITY TREE PROPERTY OF FUNCTIONING TABLE SHOWN IN FIGURE 1

Length	Paths
2	$t_4t_{16}t_3t_{15}$,
	$t_2t_{14}t_1t_{13}$
3	$t_4t_8t_{12}$,
	$t_3t_7t_{11}$
	$t_2t_6t_{10}$
	$t_1t_5t_9$
5	$t_4t_{16}t_4t_8t_{12}$,
	$t_3t_{15}t_3t_7t_{11}$
	$t_2t_{14}t_2t_6t_{10}$,
	$t_1t_{13}t_1t_5t_9$
7	$t_4t_{16}t_4t_8t_{12}$
	$t_4t_{16}t_4t_8t_{12}t_3t_{15}$
	$t_4t_{16}t_4t_8t_{12}t_2t_{14}$
	$t_4t_{16}t_4t_8t_{12}t_1t_{13}$
	$t_3t_{15}t_3t_7t_{11}t_4t_{16}$
	$t_3t_{15}t_3t_7t_{11}t_3t_{15}$
	$t_3t_{15}t_3t_7t_{11}t_2t_{14}$
	$t_3t_{15}t_3t_7t_{11}t_1t_{13}$
	$t_2t_{14}t_2t_6t_{10}t_4t_{16}$
	$t_2t_{14}t_2t_6t_{10}t_3t_{15}$
	$t_2t_{14}t_2t_6t_{10}t_2t_{14}$
	$t_2t_{14}t_2t_6t_{10}t_1t_{13}$
	$t_1t_{13}t_1t_5t_9t_4t_{16}$
	$t_1t_{13}t_1t_5t_9t_3t_{15}$
	$t_1t_{13}t_1t_5t_9t_2t_{14}$
	$t_1t_{13}t_1t_5t_9t_1t_{13}$

TABLE II

NUMBER OF TESTS N NEEDED TO ASSURE TEST QUALITY Q

N	q
66	0.91
129	0.991
193	0.9991
256	0.99991

- 1) The system is divided into several controlled subsystems based on the Functioning table.
- 2) The state transition graph is created for the algorithmic model of the behavior of each given subsystem based on the Functioning table.
- 3) Each state transition graph is transformed into an imitation of a Petri net based on the Functioning table, in which each transition has exactly one input and one output arc.
- 4) For each Petri net, based on the Functioning Table, reachability with reachability tree is created.
- 5) The criteria for covering all edges are applied to the reachability tree.

The position P is called the entry point to the transition t if there is a directed arc from P to t, the point P is called the exit point t if there is a directed arc from t to P.

Test cases play an important role in ensuring high quality system testing through algorithmic analysis. Incorrect test cases lead to unreliable web systems. It is also necessary to measure the quality of testing against some coverage criteria. In Denise and Gouraud [1], it is slightly reformalized to carry out statistical testing in congruence with any given graphical formalization of the character of the system under test. The

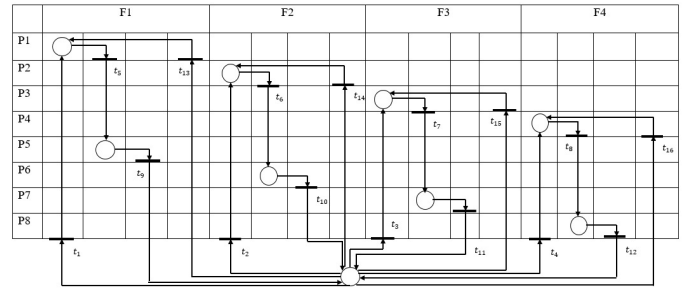


Fig. 1. Subsystem state transition graph based on the Functioning table

formalization of this method is as follows.

Let D be some algorithmic description of the system under test. D can be a specification or an algorithm, depending on the type of test of interest (functional or structural). Based on D , one can determine coverage criteria such as all k-paths, all edges, etc. D . A coverage criterion C is covered with probability $q_{C,N}(D)$ if each element of the corresponding set of elements $E_C(D)$ has probability at least $q_{C,N}(D)$ for N executions with random inputs. Quality $q_{C,N}(D)$ is a measure of test coverage relative to C . The following denotes the algorithmic relationship between test quality and test size N :

$$q_{C,N}(D) = 1 - (1 - q_{C,1}(D))^N \quad (1)$$

It is clear that the probability of reaching an element is equal to one minus the probability of not reaching it N times when N tests are performed.

This method can be used for structural testing of web systems. The completion of this work is shown below. When testing all k-paths on a reachability tree generated from a Petri net model based on a Functioning table of an example of an ecological monitoring subsystem shown with algorithmic formalization, $P \leq n$ and $AP \leq n$ denote a set of such paths, and all-k-criteria for covering paths respectively. In this case, the test quality can be obtained as follows:

$$q_{A,P \leq n,N}(D) = 1 - (1 - q_{C,1}(D))^N \quad (2)$$

In this example, choosing $n = 7$ results in 28 paths of length less than or equal to 7 (see Table 1). Equation (2) takes the form

$$q_{A,P \leq 7,N}(D) = 1 - (1 - \frac{1}{28})^N \quad (3)$$

and can be used to obtain test quality. The result is shown in Table 2. If the criterion for covering all edges is chosen instead of all k-paths, the result will be exactly the same as shown in Denise and Gouraud [1].

V. CONCLUSION

This article proposes a method of end-to-end algorithmic analysis of testing based on a Functioning table for algorithmic formalization of web systems. The web system can be divided into several subsystems depending on their characteristics using a Functioning table. State transition graphs and the corresponding Petri net model based on the Functioning table

can be created on the basis of various subsystems. The criterion for covering all edges is subsequently applied to the reachability tree derived from the Petri model based on the Functioning table for each subsystem to facilitate algorithmic analysis of the algorithm. Scientific novelty: The problem of algorithmic analysis of the system is performed on the basis of Functioning table.

All facets of the algorithmic formalization of the web system can be subjected to algorithmic analysis. The existing method of statistical testing of traditional systems is used to measure the quality of testing and algorithmic formalization of a web system with algorithmic analysis in relation to the criteria for covering all k-paths and all edges.

VI. ACKNOWLEDGEMENT

The scientific results obtained in the article are based on the project BV-M-F4-004 Development of Principles Of Algorithms for Control Of Complex Systems On The Basis Of Algebra Over Functioning Tables.

REFERENCES

- [1] Rashid, Aqsa, Masood, Asif and others, RC-AAM: blockchain-enabled decentralized role-centric authentication and access management for distributed organizations, *Cluster Computing*, vol. 24, pp. 3551–3571, 2021.
- [2] Henzinger, Thomas A., Pei-Hsin Ho, and Howard Wong-Toi. "Algorithmic analysis of nonlinear hybrid systems." *IEEE transactions on automatic control* 43.4 (1998): 540-554.
- [3] Griffiths, Thomas L., Falk Lieder, and Noah D. Goodman. "Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic." *Topics in cognitive science* 7.2 (2015): 217-229.
- [4] Lieder, Falk, and Thomas L. Griffiths. "Advancing rational analysis to the algorithmic level." *Behavioral and Brain Sciences* 43 (2020).
- [5] Bassily, Raef, et al. "Algorithmic stability for adaptive data analysis." *SIAM Journal on Computing* 50.3 (2021): STOC16-377.
- [6] Zhang, Xin-Yuan, et al. "KuhnMunkres parallel genetic algorithm for the set cover problem and its application to large-scale wireless sensor networks." *IEEE Transactions on Evolutionary Computation* 20.5 (2015): 695-710.
- [7] Fugkeaw, Somchart, Sanchol and Pattavee, A Review on Data Access Control Schemes in Mobile Cloud Computing: State-of-the-Art Solutions and Research Directions, *SN Computer Science*, vol. 3, pp. 1–11, 2022.
- [8] L. Vu, C. Bui, Q. Nguyen and D. Rossi, A deep learning based method for handling imbalanced problem in network traffic classification, December 2017, pp. 333339.
- [9] G. Aceto, D. Ciunzo, A. Montieri and P. A. Multi-classification approaches for classifying mobile app traffic, *Journal of Network and Computer Applications*, vol. 57, pp. 131145, 2018.
- [10] P. Wang, C. Xuejiao, Y. Feng and S. Zhixin, A survey of techniques for mobile service encrypted traffic classification using deep learning, *IEEE Access*, vol. 7, pp. 5402454033, 2019.
- [11] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, 2018, pp. 108116.
- [12] P. N. Matheus, F. C. Luiz, L. Jaime and L. P. Mario, Long shortterm memory and fuzzy logic for anomaly detection and mitigation in software-dened network environment, 2020, pp. 8376583781.
- [13] B. Naveen and S. Manu, Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting ddos attacks, *Romanian journal of information science and technology*, vol. 23, no. 3, p. 250 261, 2020.
- [14] S.E.Mahmoud,L.Nhien-An,D.Soumyabrata and D.J.Anca, Ddosnet: A deep-learning model for detecting network attacks, in 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). IEEE, 31 Aug.-3 Sept. 2020, pp. 18.
- [15] M. S. Yin, P. A. Pye and S. H. Aye, A slow ddos attack detection mechanism using feature weighing and ranking, *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore*, pp. 45004509, March. 7-11, 2021.
- [16] A. H. Lashkari, D. G. Gerard, M. M. Mamun and A. A. Ghorbani, Characterization of tor traf using time based features, 2017, pp. 253262.
- [17] A. Kabulov and I. Yarashov, Mathematical model of Information Processing in the Ecological Monitoring Information System, 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4.
- [18] I. Yarashov, Algorithmic Formalization Of User Access To The Ecological Monitoring Information System, 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-3.
- [19] D. Khasanov, K. Khujamatov, B. Fayzullaev and E. Reynnazarov, Wsn-based Monitoring Systems for the Solar Power Stations of Telecommunication Devices, *IJUM Engineering Journal*, 2021, 22(2), p. 98118.
- [20] A. Kamal, K. Ahmad, R. Hassan and K. Khalim, NTRU Algorithm: N^{th} Degree truncated polynomial ring units, *EAI/Springer Innovations in Communication and Computing*, 2021, p. 103115.
- [21] N. G. Zagoruiko, I. A. Borisova and O. A. Kutnenko, Constructing a concise description of data using the competitive similarity function, *Siberian Journal of Industrial Mathematics*, vol.1, no.16, pp.2941, 2013.
- [22] G. Juraev and K. Rakhimberdiev, Modeling the decision-making process of lenders based on blockchain technology, 2021 International Conference on Information Science and Communications Technologies (ICISCT), pp. 16, 2021.
- [23] I. Kalendarov, Algorithm for the Problem of Loading Production Capacities in Production Systems, *Lecture Notes in Networks and Systems*, vol. 246, pp. 887896, 2022.
- [24] Ddos evaluation dataset (cic-ddos2019), 2018. [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>
- [25] I. Sharafaldin, A. H. Lashkari, H. Saqib and A. Ghorban, Developing realistic distributed denial of service (ddos) attack dataset and taxonomy, in *In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICST)*. IEEE, Oct. 1-3, pp. 18.
- [26] A. Kabulov, I. Normatov, E. Urunbaev and F. Muhammadiev, Invariant continuation of discrete multi-valued functions and their implementation, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [27] A. Kabulov and I. Saymanov, Application of IoT technology in ecology (on the example of the Aral Sea region), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [28] A. Kabulov, I. Saymanov, I. Yarashov and F. Muxammadiev, Algorithmic method of security of the Internet of Things based on steganographic coding, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [29] A. Kabulov and M. Berdimurodov, Parametric Algorithm for Searching the Minimum Lower Unity of Monotone Boolean Functions in the Process Synthesis of Control Automates, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-5.
- [30] A. Kabulov and M. Berdimurodov, Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [31] A. Kabulov, I. Saymanov and M. Berdimurodov, Minimum logical representation of microcommands of cryptographic algorithms (AES), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [32] A. Kabulov, I. Normatov, I. Kalendarov and I. Yarashov, Development of An Algorithmic Model and Methods for Managing Production Systems Based on Algebra over Functioning Tables, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [33] A. Kabulov, I. Kalendarov and I. Yarashov, Problems of Algorithmization of Control of Complex Systems Based on Functioning Tables in Dynamic Control Systems, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-4.

- [34] A. Kabulov, E. Urunboev and I. Saymanov, Object recognition method based on logical correcting functions, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2020, pp. 1-4.
- [35] A. Kabulov, A. Babadzhanov and I. Saymanov, Completeness of the linear closure of the voting model, AIP Conference Proceedings, 2022 (accepted).
- [36] A. Kabulov, A. Babadzhanov and I. Saymanov, Correct models of families of algorithms for calculating estimates, AIP Conference Proceedings, 2022 (accepted).
- [37] A. Kabulov, I. Normatov, S. Boltaev and I. Saymanov, Logic method of classification of objects with non-joining classes, Advances in Mathematics: Scientific Journal, 2020, 9(10), p. 8635–8646.
- [38] A. Kabulov, I. Kalandarov and I. Saymanov, Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules, Smart transport conference 2022 Conference, pp. 1-11.
- [39] A. Kabulov, Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions, Cybernetics, 1979, 15(3), p. 305–311.
- [40] A. Kabulov and G. Losef, Local algorithms simplifying the disjunctive normal forms of Boolean functions, USSR Computational Mathematics and Mathematical Physics, 1978, 18(3), p. 201–207.
- [41] A. Kabulov, Local algorithms on yablonskii schemes, USSR Computational Mathematics and Mathematical Physics, 1977, 17(1), p. 210–220.
- [42] M. Shaw, N. Mandal and M. Gangopadhyay, A compact polarization reconfigurable stacked microstrip antenna for WiMAX application, International Journal of Microwave and Wireless Technologies this link is disabled, 2021, 13(9), p. 921-936.
- [43] A. Panda, P. Bhowmick, S.K. Bishnu, A. Ganguly and M. Gangopadhyaya, Derivative based kalman filter and its implementation on tuning PI controller for the van de vusse reactor, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings, 2021, 9422585
- [44] H. Khujamatov, I. Siddikov, E. Reypnazarov, D. Khasanov. Research of Probability-Time Characteristics of the Wireless Sensor Networks for Remote Monitoring Systems // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [45] I. Siddikov, D. Khasanov, H. Khujamatov, E. Reypnazarov. Communication Architecture of Solar Energy Monitoring Systems for Telecommunication Objects // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [46] I. Siddikov, K. Khujamatov, E. Reypnazarov, D. Khasanov. CRN and 5G based IoT: Applications, Challenges and Opportunities // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [47] K. Khujamatov, A. Lazarev, N. Akhmedov, E. Reypnazarov, A. Bekturdiyev. Methods for Automatic Identification of Vehicles in the its System // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [48] S. Tanwar, H. Khujamatov, B. Turumbetov, E. Reypnazarov, Z. Allamuratova. Designing and Calculating Bandwidth of the LTE Network for Rural Areas // International Journal on Advanced Science, Engineering and Information Technology, 2022, 12(2), pp. 437-445.
- [49] H. Zaynidinov, D. Singh, S. Makhmudjanov, I. Yusupov. Methods for Determining the Optimal Sampling Step of Signals in the Process of Device and Computer Integration // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 471–482.
- [50] H. Zaynidinov, D. Singh, I. Yusupov, S. Makhmudjanov. Algorithms and Service for Digital Processing of Two-Dimensional Geophysical Fields Using Octave Method // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 460–470.
- [51] H. Zaynidinov, S. Anarova, J. Jabbarov. Determination of Dimensions of Complex Geometric Objects with Fractal Structure // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 437–448.

Using Algorithmic Modeling to Control User Access Based on Functioning Table

1st Anvar Kabulov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 anvarkabulov@mail.ru

2nd Islambek Saymanov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 islambeksaymanov@gmail.com

3rd Inomjon Yarashov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 timprivate345@gmail.com

4th Anvar Karimov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 anvark87@gmail.com

Abstract—Today, the development of modern digital technologies requires them to solve existing cyber security problems. Thus, in this study, by building an algorithmic model for user access control based on a functional table, information security of existing systems can be used to create imitation of processes occurring in the system, which allows you to apply and analyze specific problems. This study attempted to focus on algorithmic issues.

In this algorithmic model based on Functioning table, the user's access control process is active at all transitions in the Petri net, and cases can be reached and detected with an initial marking, a sequence of transitions leading to a particular state.

The algorithmic model based on the Functioning table of user access control has a strictly formal description that allows you to analyze the process of interaction of three subjects of information relations. This model allows switching the algorithmic model of user access control to the system to software implementation.

Index Terms—algorithmic modeling, functioning table, security, objective attribute, datalogical model, Petri net.

I. INTRODUCTION

Under this century, no one can doubt that smart robots, technologies and other technical devices that humans could not have imagined will be created. The development of information technology will ensure increased productivity, quality and, most importantly, high efficiency for each organization [1]–[5]. Newfangled digital technologies, along with the construction of conveniences, also pose new challenges. Over the last few years, the rapid expansion of information and telecommunications technologies in the world has had a significant impact on the worthy place of states in the global information society [6]–[12].

Today, the threat to the security of information stored in databases and circulating in telecommunications systems is growing rapidly. As a result, the problem of information security [13]–[16] has become a topical issue for the world.

To date, one of the most reliable tools in ensuring information security is cryptographic protection of information. In the world, this direction is developing rapidly. New cryptographic systems, algorithms [17]–[19], standards are being developed and applied in various fields. As noted, the information resources of any organization are one of the factors determining its economic and military potential. The effective [20]–[23] use of this resource will ensure the security of the country and the successful formation of a democratically informed society. In such a society, the speed of information exchange will increase, the application of advanced information and communication technologies [24], [25] for the collection, storage, processing and use of information will be widespread.

An informed society is rapidly evolving. In the world of information, the concept of state borders is disappearing. The global computer network is radically changing public administration.

Regardless of the regional location, in everyday life, a variety of information entered the Internet through an international computer network [26]–[30]. That is why it is important to protect existing information from problems such as illegal access, use and alteration, loss.

The organization's policy in the field of informatization is aimed at creating an information system that takes into account the modern world principles of development and improvement of information resources, information technologies [8] and information systems. The policy of the organization in the data protection sphere will be aimed at regulating social relations in the information sphere and determines the basic duty and

works of state powers and management in the information security sphere of person [31], [32], [44], community and the country.

Data safeguard in workstation systems [33], [42], [45] and grids means the use of various tools and methods, taking measures and implementing measures in order to systematically [34], [43] ensure the reliability of transmitted, stored and processed information.

The basic concept of information security is a comprehensive approach based on the integration into a single system with different communication and security subsystems, common hardware, communication channels, software and databases [35], [41].

Regardless of the form in which the integrated approach is applied, it solves complex and multidisciplinary specific issues with their interrelated solutions. The most pressing of these issues are the restriction of access to information [36], [40], [46], [47], technical and cryptographic protection of information, reducing the level of adjacent radiation of technical means, technical strengthening of objects, their protection and equipped with alarm devices.

The effectiveness of the facilities information security system is critical. For computer systems, this efficiency can be assessed by the choice of hardware and software used in the computer system. Such an efficiency assessment can be made through a growing curve showing that the level of security depends on the strengthening of control [37]–[39], [48] over access rights(Fig 1).

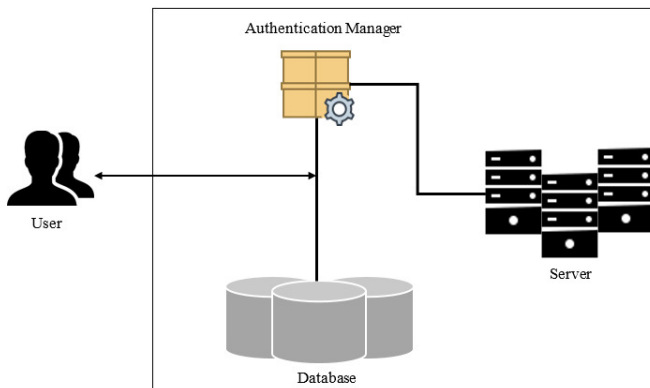


Fig. 1. Scheme for the developed algorithmic model of user access control system

II. MATERIALS AND METHODS

The algorithmic model was intended for use in a system that uses algorithmic modeling to determine the user access control workflow and takes into account some changes in the model based on the Functional Table. The user provides personal information, identifiers and / or is provided before logging in. As a result of algorithmic modeling for the administrator, a corresponding working window appears, which can display a list of active and passive users. When entering information into the database, the fettle of the motionless

/ moving executor swaps and the work that is being done separately begins. It is possible to change the user based on the subjective attribute and / or the objective attribute, it is advocated to apply this parameter only in the event of minor failures in information input(Fig 2).

It may be implemented in in practical analytical work founded on the correct of the note by monitoring the timely completion of the users work. The correct of the outcomes, in order to result in fresh approaches in survey. In this case,naturally, the notes are likened to the antecedent outcomes attained. It is therefore advocated to reserve the notes in concatenated relational databases. Increasing data size causes inconvenience and issues in performing a series of operations on information. It is therefore recommended to apply normalization in information pastures.

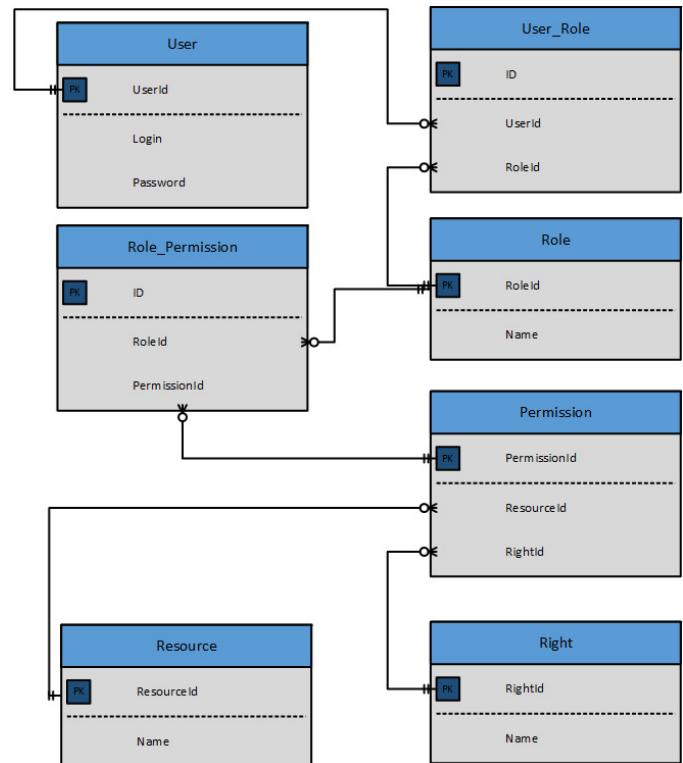


Fig. 2. Datalogical model for the developed algorithmic model of user access control system

III. RESULT AND DISCUSSION

The input label is made appropriate to the JSON Web Token standard. The norm defines JSON-based input labels. The label is made up of three components: header, payload and signature. The label is produced by the structure that provides a service, signed with a private key and conveyed to the users device, which then utilizes this label to prove his personality.

A relational database is used as a data store about the user on the server. It can be described using the datalogical model is shown in Figure 2.

It has ability to made corrections to user, create and destroy users applying the administration interface produced into the

TABLE I

DESCRIPTION OF THE PETRI NETWORK'S POSITIONS THAT IMITATES THE USER ACCESS CONTROL EVENT IN THE ALGORITHMIC MODEL BASED ON THE FUNCTIONING TABLE

p_0	User without access token
p_1	Authentication Manager received the user's login and password
p_2	The structure that provides a service adopted the executor's username and password
p_3	The structure that provides a service did not perceive a ratio in the executor's database
p_4	Authentication Manager adopted a failure information
p_5	The structure that provides a service perceived the executor in the structured set of data held in a computer
p_6	The structure that provides a service produced the executor input label
p_7	Authentication Manager adopted an input label from the structure that provides a service
p_8	User with access token
p_9	Authentication Manager adopted an input label from the executor
p_{10}	Label has been successfully made sure to the Authentication Manager
p_{11}	The label consists of the necessary consents
p_{12}	Permission to access the system
p_{13}	The label does not consist of the necessary consents
p_{14}	Expired label
p_{15}	The token with invalid signature

structure that provides a service. In order to subsequently move on to the program execution of the replica of the constructed algorithmic model based on Functioning table, it will build an algorithmic model for controlling its operation based on the Petri net inside Functioning table(Tab 1,2).

The algorithmic model implemented on the Petri network within the Functioning table is depicted as a six $\langle P, T, I, O, \mu, F_t \rangle$, where P and T are finite position sets and transition sets, I and O are sets of incoming and outgoing functions, μ is a set of state markings. The incoming function I maps the transition t_j to the set of positions $I(t_j)$, F_t to the set of components for operation and the output function O maps the transition t_j to the set of positions $O(t_j)$.

- finite position sets : $P = \{p_0, p_1, p_2, \dots, p_{15}\}$
- finite set of transition: $T = \{t_0, t_1, t_2, \dots, t_{15}\}$
- set of transition incoming positions:
 $I = \{I(t_0), I(t_1), I(t_2), \dots, I(t_{15})\}$.
 $I(t_0) = \{p_0\}, I(t_1) = \{p_1\}, I(t_2) = \{p_3\},$
 $I(t_3) = \{p_4\}, I(t_4) = \{p_5\}, I(t_5) = \{p_6\},$
 $I(t_6) = \{p_7\}, I(t_7) = \{p_8\}, I(t_8) = \{p_1\},$
 $I(t_9) = \{p_9\}, I(t_{10}) = \{p_{10}\}, I(t_{11}) = \{p_{11}\},$
 $I(t_{12}) = \{p_{12}\}, I(t_{13}) = \{p_{13}\}, I(t_{14}) = \{p_{14}\},$
 $I(t_{15}) = \{p_{15}\}.$
- set of transition outgoing positions:
 $O = \{O(t_0), O(t_1), O(t_2), \dots, O(t_{15})\}$.
 $O(t_0) = \{p_2\}, O(t_1) = \{p_3\},$
 $O(t_2) = \{p_4, p_6\}, O(t_3) = \{p_5\},$
 $O(t_4) = \{p_0\}, O(t_5) = \{p_7\},$
 $O(t_6) = \{p_8\}, O(t_7) = \{p_1\},$
 $O(t_8) = \{p_9\}, O(t_9) = \{p_{10}, p_{14}, p_{15}\},$

TABLE II

THE DESCRIPTION OF THE PETRI NETWORK'S TRANSITIONS THAT IMITATES THE USER ACCESS CONTROL EVENT IN THE ALGORITHMIC MODEL BASED ON THE FUNCTIONING TABLE

t_0	Lapsing the login and password to the authentication manager
t_1	Lapsing the login and password to the server
t_2	Searching for a user in the database
t_3	Lapsing the error message to the authentication manager
t_4	Lapsing the failure note to the executor
t_5	Producing an executor input label
t_6	Lapsing an input label from the structure that provides a service to the authentication manager
t_7	Passing the access token to the user
t_8	Passing an access token from a user to an authentication manager
t_9	Verification of access token
t_{10}	Checking access token permissions
t_{11}	Access to the system
t_{12}	Sending a message about successful object access
t_{13}	Lapsing the failure note to the executor
t_{14}	Lapsing the failure note to the executor
t_{15}	Lapsing the failure note to the executor

$$O(t_{10}) = \{p_{11}, p_{13}\}, O(t_{11}) = \{p_{12}\},$$

$$O(t_{12}) = \{p_1\}, O(t_{13}) = \{p_1\}, O(t_{14}) = \{p_1\},$$

$$O(t_{15}) = \{p_1\}.$$

- Inceptive labeling:

$$\mu = \{1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

The execution of the Petri network inside Functioning table is carried out by starting transitions(Fig 4). A transition is induced by detaching labels from its input positions and producing recently developed labels placed at its output positions. A transition is able to only be launched if it is permitted, if every one of its input positions contains amount of labels at least equaling the number of arcs from position to transition.

Evenness in modeling on Petri nets inside Functioning table is determined by the motions deriving position in the mechanism, as well as by the conditions foregoing the supervene of the transitions, after that the cases after the realization of the events. Analysis of the outcomes of the algorithmic model execution is able to talk about the cases in which the algorithmic model was or was not, which cases are not attainable(Fig 3).

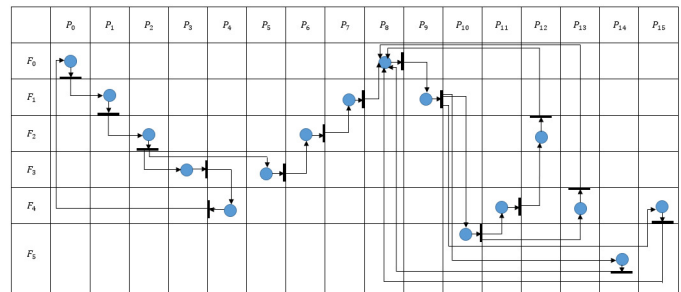


Fig. 3. Algorithmic model based on Functioning table for the developed user access control system.

F_0 - Access marker; F_1 - Authentication Manager; F_2 - Server; F_3 - Database; F_4 - Checking permissions; F_5

- Checking token. The incident matrix for this algorithmic model is as follows (Fig 4):

$$\begin{matrix}
 & t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} & t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\
 p_0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_2 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_3 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_4 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_5 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_6 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
 p_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 p_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
 p_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\
 p_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\
 p_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\
 p_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
 p_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1
 \end{matrix}$$

Fig. 4. The incident matrix for this algorithmic model based on Functioning table

It is simulate the process of user access control in the algorithmic model based on Functioning table. The designations of the elements of this model are as follows:

The obtained algorithmic model is analyzed based on the basic properties of Petri nets: security, constraints, stability, reachability and activity. A description of every one of these features is presented and on their basis the characterization of the built algorithmic model is given(Fig 3).

A position p_i is called secure in a stated inceptive labeling μ if, during the procedure of this algorithmic model, more than k one label (marker or token) never shows up in a given position p_i , that is, $\mu(p_i) \leq 1$. A Petri network inside the Functioning table is called safe if all of its position.

From this definition, it follows that the model represented by the Petri network, which implements the process of user access control in the algorithmic model being suggested, is secure, since there is no elements' gradual gathering in the positions of this algorithmic model.

A position $p_i \in P$ is called limited in a stated inceptive labeling μ if, during the procedure of this algorithmic model, more than k markers never appear at a given position p_i , that is, $\mu(p_i) \leq k$. A Petri network inside the Functioning table is called limited if all of its positions are limited.

The model simulating the process of user access control in the suggested algorithmic model is limited with the value $k = 1$, since the number of labels in each position is considered for the case of $k = 1$.

A Petri net inside the Functioning table is called stable if, for any of its transitions $t_i \in T$, the following condition is satisfied: the case of agitation of this transition is not able to be detached by inducing any other transition. If the algorithmic model has deputy transitions, then it is unsteady.

In the algorithmic model, the process of user access control in the FT, there are alternative transitions, for example, transitions t_2, t_9, t_{10} , therefore, this algorithmic model is erratic.

Labeling μ' is called attainable from some labeling μ if for a stated algorithmic model of the Petri network inside the Functioning table it is achievable to indicate such a pattern of transitions that modifies labeling μ into labeling μ' .

The absence of dead-ends in the algorithmic model shown in Figure 3 allows us to assert that its markings are achievable. A transition $t_i \in T$ is called active in a stated inceptive labeling μ if, for any labeling μ' attainable from μ , it is achievable to indicate a sequence of transitions flashing, which attends to the transition's t_i agitation. The algorithmic model is called active in a stated inceptive labeling μ if whole of its transitions are active.

IV. CONCLUSION

Petri network, which imitates the user access control event in the algorithmic model based on Functioning table being suggested, is active, since all its transitions are active, and the cases are attainable with the inceptive labeling, and it is permissible to designate the pattern of flashing of the transitions such will attend to one or another case.

The constructed algorithmic model based on Functioning table of the user access control has a strict formal description that allows analyzing the process of interaction between three subjects of information relations, namely, a user, an authentication manager and a server. This model makes it possible to go to the program execution of the replica of the user access control algorithmic model for a system.

REFERENCES

- [1] Rashid, Aqsa, Masood, Asif and others, RC-AAM: blockchain-enabled decentralized role-centric authentication and access management for distributed organizations, Cluster Computing, vol. 24, pp. 3551–3571, 2021.
- [2] Fugkeaw, Somchart, Sanchol and Pattavee, A Review on Data Access Control Schemes in Mobile Cloud Computing: State-of-the-Art Solutions and Research Directions, SN Computer Science, vol. 3, pp. 1–11, 2022.
- [3] L. Vu, C. Bui, Q. Nguyen, and D. Rossi, A deep learning based method for handling imbalanced problem in network traf c classification. December 2017, pp. 333339.
- [4] G. Aceto, D. Ciunzo, A. Montieri, and P. A. Multi-classification approaches for classifying mobile app traf c, Journal of Network and Computer Applications, vol. 57, pp. 131145, 2018.
- [5] P. Wang, C. Xuejiao, Y. Feng, and S. Zhixin, A survey of techniques for mobile service encrypted traf c classification using deep learning, IEEE Access, vol. 7, pp. 5402454033, 2019.
- [6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traf c characterization, 2018, pp. 108116.
- [7] P. N. Matheus, F. C. Luiz, L. Jaime, and L. P. Mario, Long shortterm memory and fuzzy logic for anomaly detection and mitigation in software-dened network environment, 2020, pp. 8376583781.
- [8] B. Naveen and S. Manu, Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting ddos attacks, Romanian journal of information science and technology, vol. 23, no. 3, p. 250 261, 2020.
- [9] S.E.Mahmoud,L.Nhien-An,D.Soumyabrata,and D.J.Anca, Ddosnet: A deep-learning model for detecting network attacks, in 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). IEEE, 31 Aug.-3 Sept. 2020, pp. 18.
- [10] M. S. Yin, P. A. Pye, and S. H. Aye, A slow ddos attack detection mechanismusingfeatureweighingandrakingn,Proceedingsofthe11th Annual-InternationalConferenceonIndustrialEngineeringandOperations Management Singapore, pp. 45004509, March. 7-11, 2021.
- [11] A. H. Lashkari, D. G. Gerard, M. M. Mamun, and A. A. Ghorbani, Characterization of tor traf c using time based features, 2017, pp. 253262.
- [12] N. Miloslavskaya, A. Tolstoy, and S. Zapechnikov, Taxonomy for unsecure big data processing in security operations centers, Aug.2224 2016, pp. 154159.

- [13] N. Miloslavskaya and A. Makhmudova, Survey of big data information security, vol. 8, Aug.22-24 2016, pp. 133138.
- [14] D. Khasanov, K. Khujamatov, B. Fayzullaev and E. Reypnazarov, WSN-based Monitoring Systems for the Solar Power Stations of Telecommunication Devices, IIUM Engineering Journal, 2021, 22(2), p. 98118.
- [15] A. Kamal, K. Ahmad, R. Hassan and K. Khalim, NTRU Algorithm: N^{th} Degree truncated polynomial ring units, EAI/Springer Innovations in Communication and Computing, 2021, p. 103115.
- [16] N. G. Zagoruiko, I. A. Borisova, and O. A. Kutnenko, Constructing a concise description of data using the competitive similarity function, Siberian Journal of Industrial Mathematics, vol.1, no.16, pp.2941, 2013.
- [17] G. Juraev and K. Rakhimberdiev, Modeling the decision-making process of lenders based on blockchain technology, 2021 International Conference on Information Science and Communications Technologies (ICISCT), pp. 16, 2021.
- [18] I. Kalandarov, Algorithm for the Problem of Loading Production Capacities in Production Systems, Lecture Notes in Networks and Systems, vol. 246, pp. 887896, 2022.
- [19] R. Ibraimov, M. Sultonova and H. Khujamatov, The Integral distribution function of the kilometric attenuation of infrared radiation in the atmosphere Fergana Region of the Republic of Uzbekistan, Webology, 2021, 18(Special Issue), pp. 316-327.
- [20] I. Sharafaldin, A. H. Lashkari, H. Saqib, and A. Ghorban, Developing realistic distributed denial of service (ddos) attack dataset and taxonomy, in In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICST). IEEE, Oct. 1-3, pp. 18.
- [21] A. Kabulov, I. Normatov, E. Urunbaev and F. Muhammadiev, Invariant continuation of discrete multi-valued functions and their implementation, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [22] A. Kabulov and I. Saymanov, Application of IoT technology in ecology (on the example of the Aral Sea region), International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [23] A. Kabulov, I. Saymanov, I. Yarashov and F. Muxammadiev, Algorithmic method of security of the Internet of Things based on steganographic coding, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [24] A. Kabulov and M. Berdimurodov, Parametric Algorithm for Searching the Minimum Lower Unity of Monotone Boolean Functions in the Process Synthesis of Control Automates, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-5.
- [25] A. Kabulov and M. Berdimurodov, Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal), International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [26] A. Kabulov, I. Saymanov and M. Berdimurodov, Minimum logical representation of microcommands of cryptographic algorithms (AES), International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [27] A. Kabulov, I. Normatov, I. Kalandarov and I. Yarashov, Development of An Algorithmic Model and Methods for Managing Production Systems Based on Algebra over Functioning Tables, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-6.
- [28] A. Kabulov, I. Kalandarov and I. Yarashov, Problems of Algorithmization of Control of Complex Systems Based on Functioning Tables in Dynamic Control Systems, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-4.
- [29] A. Kabulov, E. Urunboev and I. Saymanov, Object recognition method based on logical correcting functions, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2020, pp. 1-4.
- [30] A. Kabulov, A. Babadzhanov and I. Saymanov, Completeness of the linear closure of the voting model, AIP Conference Proceedings, 2022 (accepted).
- [31] A. Kabulov, A. Babadzhanov and I. Saymanov, Correct models of families of algorithms for calculating estimates, AIP Conference Proceedings, 2022 (accepted).
- [32] A. Kabulov, I. Normatov, S. Boltaev and I. Saymanov, Logic method of classification of objects with non-joining classes, Advances in Mathematics: Scientific Journal, 2020, 9(10), p. 8635–8646.
- [33] A. Kabulov, I. Kalandarov and I. Saymanov, Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules, Smart transport conference 2022 Conference, pp. 1-11.
- [34] A. Kabulov, Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions, Cybernetics, 1979, 15(3), p. 305–311.
- [35] A. Kabulov and G. Losef, Local algorithms simplifying the disjunctive normal forms of Boolean functions, USSR Computational Mathematics and Mathematical Physics, 1978, 18(3), p. 201–207.
- [36] A. Kabulov, Local algorithms on yablonskii schemes, USSR Computational Mathematics and Mathematical Physics, 1977, 17(1), p. 210–220.
- [37] M. Shaw, N. Mandal and M. Gangopadhyay, A compact polarization reconfigurable stacked microstrip antenna for WiMAX application, International Journal of Microwave and Wireless Technologies this link is disabled, 2021, 13(9), p. 921-936.
- [38] A. Panda, P. Bhowmick, S.K. Bishnu, A. Ganguly and M. Gangopadhyaya, Derivative based kalman filter and its implementation on tuning PI controller for the van de vusse reactor, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings, 2021, 9422585
- [39] H. Khujamatov, I. Siddikov, E. Reypnazarov, D. Khasanov. Research of Probability-Time Characteristics of the Wireless Sensor Networks for Remote Monitoring Systems // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [40] I. Siddikov, D. Khasanov, H. Khujamatov, E. Reypnazarov. Communication Architecture of Solar Energy Monitoring Systems for Telecommunication Objects // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [41] I. Siddikov, K. Khujamatov, E. Reypnazarov, D. Khasanov. CRN and 5G based IoT: Applications, Challenges and Opportunities // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [42] A. Kabulov and I. Yarashov, Mathematical model of Information Processing in the Ecological Monitoring Information System, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-4.
- [43] K. Khujamatov, A. Lazarev, N. Akhmedov, E. Reypnazarov, A. Bekturdiev. Methods for Automatic Identification of Vehicles in the its System // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021.
- [44] S. Tanwar, H. Khujamatov, B. Turumbetov, E. Reypnazarov, Z. Allamuratova. Designing and Calculating Bandwidth of the LTE Network for Rural Areas // International Journal on Advanced Science, Engineering and Information Technology, 2022, 12(2), pp. 437-445.
- [45] H. Zaynidinov, D. Singh, S. Makhmudjanov, I. Yusupov. Methods for Determining the Optimal Sampling Step of Signals in the Process of Device and Computer Integration // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 471–482.
- [46] I. Yarashov, Algorithmic Formalization Of User Access To The Ecological Monitoring Information System, International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, pp. 1-3.
- [47] H. Zaynidinov, D. Singh, I. Yusupov, S. Makhmudjanov. Algorithms and Service for Digital Processing of Two-Dimensional Geophysical Fields Using Octave Method // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 460–470.
- [48] H. Zaynidinov, S. Anarova, J. Jabbarov. Determination of Dimensions of Complex Geometric Objects with Fractal Structure // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13184 LNCS, pp. 437–448.

Sensory Data Fusion Using Machine Learning Methods For In-Situ Defect Registration In Additive Manufacturing: A Review

1st Javid Akhavan
Mechanical Engineering
Stevens Institute of Technology
Hoboken, USA
jakhavan@stevens.edu
0000-0002-6485-5986

2nd Souran Manoochehri
Mechanical Engineering
Stevens Institute of Technology
Hoboken, USA
smanooch@stevens.edu
0000-0002-7189-6356

Abstract—In-situ control to predict and mitigate defects in Additive Manufacturing (AM) could significantly increase these technologies' quality and reliability. Thorough knowledge of the AM processes is needed to develop such a controller. Recent studies utilized various methods to acquire data from the process, build insight into the process, and detect anomalies within the process. However, each sensory method has its unique limitations and capabilities. Sensor fusion techniques based on Machine Learning (ML) methods can combine all the data acquisition sources to form a holistic monitoring system for better data aggregation and enhanced detection. This holistic approach could also be used to train a controller on top of the fusion system to master the AM production and increase its reliance. This article summarizes recent studies on sensor utilization, followed by ML-based sensor fusion and control strategies.

Index Terms—Sensor Fusion, Data Fusion, Machine Learning, Additive Manufacturing (AM), In-situ Defect Detection, In-situ Process Control, Defect Classification

I. INTRODUCTION

Additive Manufacturing (AM) technologies are known for providing cheap and fast prototyping methods. However, these technologies suffer from unreliable part's quality. Extensive research utilizing different sensors and algorithms has been conducted to study the AM processes and sources of part's defects. Each sensory information would

enable the possibility of detecting a limited variety of defects. Therefore, as a tool for combining information from various sources, sensor fusion gained the attention of many researchers.

Holistic defect detection can be achieved by applying ML-based sensory data fusion methods. These methods play a significant role in enhancing the data processing section of the recent studies and lead to a higher generalization by considering various aspects and correlations of data sources within the networks' training. Capabilities of ML methods have been pointed out in many recent scholarly work in various domains such as bio-printing [1], high dimensional system identification [2], and Structural Health Monitoring (SHM) [3], [4]. Implemented sensor fusion can also be combined with control and prediction algorithms leading to a more accurate and reliable control over the process.

This review summarizes different sensing methods and their capabilities and limitations in section II, followed by various machine learning-based sensory data fusion methods in peer's works and their findings in section III. Then control systems by application of the sensory data fusion are reviewed in section IV. Section V provides a short overview of the gaps found in the literature, and the future work is illustrated. Finally, the review's conclusion

is presented, in section VI.

II. SENSING METHODS

A. Offline Sensing Methods

Peer studies used offline sensing methods to determine parts' mechanical properties and structural integrity. These studies are based on destructive measurement methods that make them unsuitable for in-situ sensor fusion aspects. Among recent studies, Sabyrov et al. [5] correlated Ultimate Tensile Stress (UTS) of the parts with process parameters, and literature [5], [6] used the SEM method to study the bounding and strength of the parts. Their work showed the correlation between process quality and part property. Although the method is unsuitable for sensor fusion, their findings can significantly contribute to process control and error mitigation when combined with online sensory methods.

B. Online Sensing Methods

For online control and error manipulation, online sensing methods are crucial. Some of the most common and popular online data acquisition methods are vision-based sensors, thermal profiler sensors, laser surface profilometers (LSP), and Acoustics Emission (AE) receivers. Literature [7]–[9] used visual information about the process, literature [6], [9]–[11] used infrared and pyrometer sensors to record the thermal history of the ongoing process, literature [7], [10]–[14] used AE of the printer as a primary study tool, literature [8]–[11] used LSP technology to study the process. Here we will discuss selected peers' works on these sensory techniques.

1) *Vision-Based Sensors*: Vision-based information about the ongoing process proved to be beneficial to many researchers. Vandone et al. [8] suggested using a camera system to detect the laser's working state in a Direct Energy Deposition (DED) machine; Baumann and Roller [15] suggested utilizing a cheap webcam camera for in-situ major fault detection of the FDM printing process. And Pearce et al. [16] Stepped even further and came up with a better strategy to cover up a few of the previous works' shortcomings. The author suggested using three pairs of cameras situated 120-degrees apart around the printing bed to reconstruct

the part three-dimensionally to detect any horizontal or vertical geometrical deviation.

To make in-situ data acquisition and processing feasible, peers' studies each demonstrated a new idea for processing vision-based sensory data. Vandone et al. [8] suggested applying two levels of thresholds over the intensity of the pixels in each captured image to detect the area of the heat-affected and melted zone. By summing the number of pixels and locating the center of those regions, the author retrieved critical information such as pool diameter, the center of the heat-affected and melt pool zones, and the existence of any sparks. Baumann and Roller [15] Suggested using image processing algorithms such as CV-HoughCircles and CV-CannyEdge to detect Region of Interest (ROI). Then, by following the center and boundary of ROI, major abnormalities such as part detachment, part deformation, and clogged extruder were detected. Pearce et al. [16] proposed a fusion of camera pairs to combine their field of view. By this method, the author reconstructed the part's 360-degree representation. Then, for 3D reconstruction based on the 360-degree imagery input, the author suggested using the images' non-rescale and rectification method. This pre-processing algorithm helped the author compare the printed part's dimension to the CAD data.

Each method suggested above proved that the vision-based study of the process is useful. However, some of the assumptions are not accurate and generalizable to other applications. Vandone et al. [8] suggested a thresholding approach. Although they considered light intensity changes due to the generated fumes, they didn't consider the effect of the printing area's ambient light changes or the reflection properties of the printing material. They should repeat the thresholding in any new environment or different material. The method presented in Baumann and Roller [15] showed an acceptable detection rate, although it relies on the user's manual input and instability toward ambient changes. However, this method is only valid for major faults. For detailed defects such as surface quality and over/underprint, this approach wouldn't work. Pearce et al. [16] proposed a time-efficient and accurate algorithm. But, their method wouldn't work for shallow or overhung shadowed parts.

2) *Thermal Sensor Utilization:* Heating the raw material to form a material bonding is the basis of many AM techniques; FDM machines use a heater to melt down the filament, and DED and PBF machines utilize a high-power laser module to melt down the metal powder. Due to the nature of the process, the material's thermal history proved to help detect abnormalities in the process or the part's quality. Kuznetsov et al. [6] utilized a FLIR thermal camera to study the FDM process and correlate the temperature history of deposited layers with the part's quality. Wang et al. [17] suggested the usage of a FLIR thermal camera to monitor a Large Scale Additive Manufacturing (LSAM). The author used the printing bed's online thermal map and numerical modeling to lower the layer's printing time. Li et al. [18] utilized a FLIR thermal camera to monitor a metal printer. Based on the thermal images gathered, the author predicted any abnormalities in the process parameters, such as high/low printing speed or power.

The thermal history of the material undergoing 3D printing is used in different ways to determine any abnormalities in the goal part. Kuznetsov et al. [6] varied the extruder's temperature and the cooling rate. Each manufactured piece was analyzed by mechanical testing and SEM to gather UTS and inner filaments' bonding conditions. It was shown that the part's UTS linearly increases by the extruder's temperature. They justified this result by referring to the better interlayer bonding of filaments and part's weight increase due to higher fluidity of melted filament in higher temperatures. Wang et al. [17] pointed out that LSAM processes must be within a specific range of temperatures before deposition of the next layer. This limitation is posed to let previous layers be rigid enough to hold and be warm enough to bond with new material. For that reason, by theoretical and experimental data, they lowered the print time of each layer and increased the part's quality by meeting the acceptable bonding temperatures. Li et al. [18] proposed a Deep Convolutional Neural Network (DCNN) for studying abnormal process parameters such as high/low speed and power. To train this network, they gathered a big data set consisting of the thermal images of multiple classes of process abnormalities.

The works mentioned above all showed the benefits of thermal monitoring of the process for defect detection. However, there are some points worth mentioning. The author in [6] didn't consider the effect of flow rate by changing the temperature. They could have stabilized the mass flow rate by fixing the filament flow rate, leading to a better Design Of the Experiment (DOE). Wang et al. [17] lowered the layer time successfully. However, the author changed all other process parameters to adapt to the goal printing time. These changes lead to the appearance of different realms of defects. In Li et al. [18], the author used a fixed value for normal process parameters; however, normal printing processes would be different in other parts and geometries, which means they have to train their network once again for a new part.

3) *Laser Surface Profilometer Utilization:* One of the AM parts' downsides is their rough surfaces, which are caused by the process's layer-by-layer production. Previous works showed that it's possible to advance the final surface roughness qualities by adjusting the parameters. For that reason, Lin et al. [19] utilized an LSP to monitor the printer's top surface quality in-situ. The author aimed to detect under/overprinted regions based on the point cloud processing generated from the top surface. The author obtained the depth map by performing point cloud processing and dimensionality reduction of data. In parallel, by processing the goal object's STL representation, the goal depth map was generated as ground truth for comparison. The region was then labeled by comparing the depth maps and applying upper and lower tolerances. A feedback signal is generated based on the abnormal areas' volume and dimensions to stop the process and prevent further time and material loss. This idea can be further advanced to cover the sides of the part as well. The author could also have considered changing process parameters to overcome the defects detected throughout the algorithm to obtain real-time quality control. Their work is illustrated in "Fig. 1".

Borish et al. [20], [21] suggested mounting an LSP on a Large-Scale Metal AM (LSMAM) and a Large-Scale Polymer AM (LSPAM) machine, respectively. The author performed the top surface scanning layer by layer to acquire the height

map's point cloud. Like Lin et al. [19], they sliced the CAD representation to obtain the goal part's height map ground truth. By comparing the height maps, determining whether the print's state was under/overprint was deemed feasible. The algorithm dynamically increases the underprint state's feed rate and decreases the overprint state's feed rate. Their method, combined with the control over the printer's process parameters, showed the possibility of in-situ defect mitigation. However, their study was limited to feed rate adjustment, which raises future research to involve other parameters combined with feed rate to mitigate the total height error.

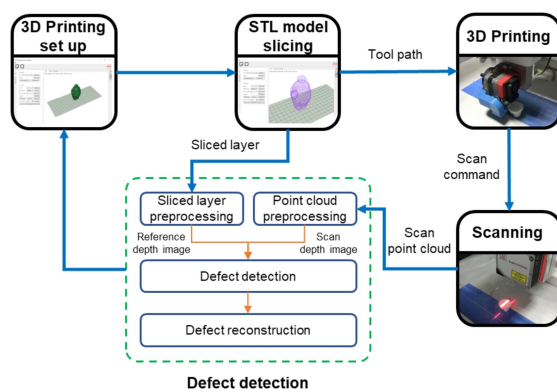


Fig. 1: Defect reconstruction process presented in [19]

4) *Acoustic Emissions And Vibrational Sensors Utilization:* On top of the previous measurement aspects, the Acoustics emissions and Vibrational signals of the printer proved to help determine a new realm of defects. These methods can capture the state of the mechanical systems and pinpoint any abnormalities in the structure. However, processing their signal is expensive and requires complex computations. Zhang et al. [12] proposed utilizing an attitude sensor consisting of 3-axial angular velocity, vibratory and magnetic receivers for dynamic condition monitoring. Li et al. [13] proposed a gearbox fault diagnosis system based on Acoustic Emissions and vibrational sensors. Wu et al. [22] suggested using AE sensors for in-situ state detection of an FDM machine. The author aimed to study normal machine state, idle, material loading, out-of-material, semi-blocked, and blocked extruder head conditions. For data processing, each paper presented a different path. Zhang et al. [12]

utilized sparse autoencoder networks depicted in “Fig. 3” to combine the sensory inputs and study the system. The author implemented nine predefined defect states during each print and aimed to classify the AE readings into nine categories. Li et al. [13] mounted the sensors on the gearbox's body to record the signals of healthy and ten different abnormalities in various system members. The proposed method relies on the Wavelet Packet Transform (WPT) to pre-process the data before feeding them to two Deep Boltzmann Machines (DBM) networks to construct a feature map. In the end, a Deep Random Forest (DRF) fusion method was applied to combine the features to achieve a better generalization. Their network structure is summarized in “Fig. 2” Wu et al. [22] mounted an AE sensor on the extruder head and recorded the FDM machine's performance in all states. After signal processing, they achieved each state's frequency and time domain features, such as ABS energy and RMS values. Then by applying the SVM method, they compared each state with another to distinguish boundaries and classification.

The recent studies mentioned above all proved usefulness of these sensors. Zhang et al. [12] showed that it is possible to detect the deficiencies with reasonable accuracy. However, their work scope is limited to the printer's mechanical faults, and they didn't propose any idea to mitigate the defect when it is detected. Li et al. [13] implied that combining these sensors fulfilled a better prediction of 96.8% accuracy over 11 different classes. Although their work showed a significant improvement in error detection, they still didn't propose an alternative for the system when a fault is detected to overcome the defect. Wu et al. [22] proposed a process resulting in a 95% accuracy in predicting machine states. However, their SVM method is not the best approach due to its linearity. Changing the algorithm with a nonlinear multi-class one could have achieved a better classification. As seen so far, online monitoring is capable of detecting many abnormalities in the process. To conclude the online monitoring section, the peer studies are summarized in Table I.

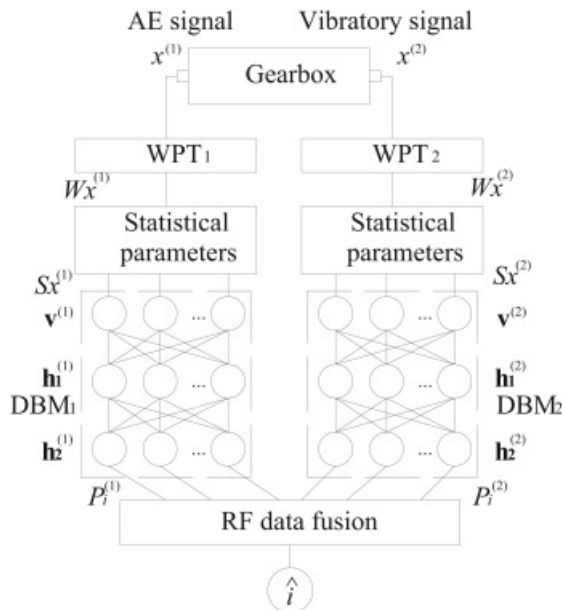


Fig. 2: Summary of data fusion used in [13] for AE signal and Vibrational data using DRF

III. SENSORY DATA FUSION

Each paper discussed so far by utilizing one or a few sensors can cast some light over a limited variety of defects. However, for a general multi-aspect study of the process, a combination of sensory data is needed. But, each sensory data requires a different approach for processing and classification based on its nature. Also, the fusion of information from various sources requires a method to decide on each aggregated data best weight or importance level, which ML methods would come in handy in this part. Previous studies approached sensor fusion from different aspects. Here a summary of them will be discussed.

Data fusion based on Hall and Llinas [23] is divided into three different categories: 1) decision-level, 2) data-level, and 3) feature-level. Some studies fused multiple sensors independently to combine their decisions; Borish et al. [21], among the peer works, utilized an LSP to detect surface quality and control the system's feed rate along with a thermal camera, used to monitor the printed part's temperature. The data from each sensor was processed separately to actuate independent alteration of the process. In some studies, such as

Vandone et al. [8], the author took advantage of the data-level fusion strategy by using multiple sensors to register the data in the time/space domain. They used vision-based sensors and system parameters to register the data in the time domain, the surface profilometer, and the G-code to register data in the space domain. The author then designed a closed-loop data-driven controller to enhance the process and meet the part's geometrical requirements. Their work is represented in "Fig. 4."

Studies like Rao et al. [24] used feature-level sensor fusion by utilizing thermocouples, cameras, and accelerometers all in parallel to achieve informative results. Their setup is shown in "Fig. 5". Zhang et al. [12] used SAE networks to obtain the feature maps; Zhang et al. [7] operated a camera, AE sensor, and Arc Voltage of the machine to predict the welding's quality.

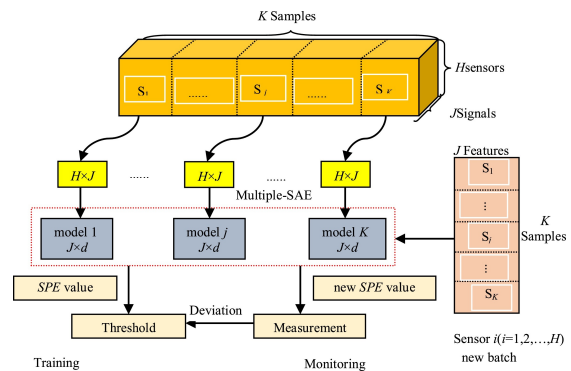


Fig. 3: Data Fusion using sparse autoencoder networks presented in [12]

The author used Support Vector Machine-Cross Validation (SVM-CV) technique to fuse the feature maps. The author compared a single sensor with multisensory data classification. They reported that spectrum-based models achieved 100%-87% accuracy, while for multisensory models, they got 94.5%-93.4%. They concluded that the multisensory model had better repeatability while the spectrum-driven model performed more accurately. But their conclusion is based on utilizing a linear binary classification algorithm (SVM). It might be possible that the author missed some nonlinear behavior of the system. This caused the drop in the accuracy to 94.5%, which could have been increased by implementing a more advanced method

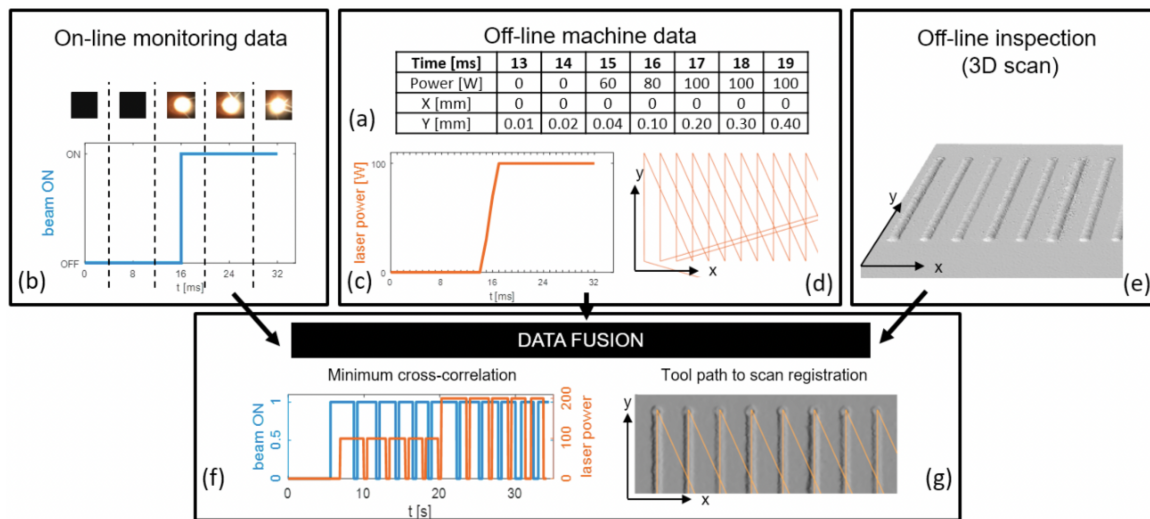


Fig. 4: Data fusion structure presented in [8]. a) The process parameters such as laser power and coordinated of the print are aggregated b) Process observation data c) Laser power schedule and d) Tool path generated from G-code e) Offline inspection results from 3D scanner f) Time alignment of the aggregated data and g) Spatial alignment of the data after fusion.

such as a deep network.

Although sensor fusion enables the ability to extract meaningful feature maps, processing and decision-making require advanced algorithms such as ML methods. Many options are currently available for researchers, from shallow to deep networks, each with many structures. In most prediction cases, the algorithm used requires significant training data, making the study expensive. According to the most recent sensor fusion works, hybrid algorithms can be used to extract information faster with less training data. Among the literature, Lyu et al. [36] suggested a hybrid convolutional autoencoder decoder network structure to automate data feature extraction, which significantly reduced the need for a thoroughly labeled dataset. Li et al. [37] recommended utilizing a one-shot learning algorithm. The author claimed that this approach would change the typical training approaches that rely on extensive training samples to only one instance of each defect. Convolutional Generative Adversarial Encoders for data fusion were suggested by Wang et al. [17] to create a predictive model with only normal condition signals. These studies show that fault detection is possible even when lacking training data or very sparse data is available. However, these studies wouldn't be

practical unless combined with a control system to mitigate the detected anomalies.

IV. QUALITY CONTROL USING SENSORY DATA FUSION

Quality assurance has been a concern of many researchers in all manufacturing domains, especially in AM. A minor flaw in the initial layers builds up to a significant deviation in the top layers. So, early detection and correction of defects would significantly advance the reliability of AM technologies. The fusion of sensory data and process parameters has shown significant value in awareness development and strategizing control signals. The approach in which a flaw can be addressed in AM strategies in this paper, depending on how they aim to fix the deviations, is categorized into the Addressing Future Defects method and the Addressing Existing Defects method.

A. Addressing Future Defects

In addressing the future defects reduction, once the defect is detected, and enough knowledge regarding the coordinates and severity of the anomaly is gathered, then the control module aims to either follow negative defect implementation to minimize older defects' effect on the total part's quality or

TABLE I: Summary of sensor utilization in peer works

Paper	Sensor Type					Performance Testing Tensile Test & SEM	Fabrication Studies Process Parameters	Defect/aspect under the study	Fusion Level	Method
	Vision	Thermal Profliometer Camera	Thermocouple	Vibrational	Acoustics Emission					
[25]	*							Melt pool monitoring and coordinate registration	Data	PCA ¹
[9]	*			*				Welding condition: 1)Burning through 2)Okay 3)Under penetration	Feature	SVM ²
[15]	*							Major error detection in the printing process	Decision	Image Processing
[16]	*							Geometry validation and failure detection	Data	SIFT ³ & RANSAC ⁴
[8]	*					*	*	Target part quality: 1)Superficial roughness 2)3D Geometry	Data	ANN ⁵
[26]	*					*	*	Layer height control using structure light	Data	PCA
[27]	*	*				*	*	Melt pool anomaly detection	Data	Threshold Filters
[9]	*	*	*	*	*	*	*	Target part quality: Surface roughness (Ra)	Data & Decision	DPM ⁶ & ET ⁷
[28]	*	*	*	*	*	*	*	Defect localization and quality assessment using AI	Data & Feature	DCNN ⁸
[29]	*	*	*	*	*	*	*	Porosity detection in laser-beam AM	Data & Decision	LRCN ⁹ DCNN
[30]	*	*	*	*	*	*	*	Strain and temperature profile assessment and simulation	Data	FEA ¹⁰
[10]	*	*	*	*	*	*	*	Target part quality: Surface roughness	Feature	RF ¹¹ , SVM, RR ¹² , LASSO ¹³
[11]	*	*	*	*	*	*	*	Target part quality: Surface roughness	Feature	RF
[21]	*	*	*	*	*	*	*	In-situ quality control for under/overprinting states	Decision	Image Processing
[31]	*	*	*	*	*	*	*	Porosity detection using D-LSTM network	Feature	D-LSTM ¹⁴
[32]	*	*	*	*	*	*	*	Porosity detection using U-net fusion	Feature	U-net
[17]	*	*	*	*	*	*	*	Optimization and control of print based on the temperature history	Feature	Regression Model
[18]	*	*	*	*	*	*	*	Detection of abnormal process parameters in-situ using DCNN	NA	DCNN
[33]	*	*	*	*	*	*	*	Distortion prediction using big data deep models	Data	DCNN
[6]	*	*	*	*	*	*	*	Correlating part's final properties with process parameters	Data & Feature	Regression Model
[22]	*	*	*	*	*	*	*	Detection of printer-state: 1)Idle 2)Blocked 3)Normal	Feature	SVM
[12]	*	*	*	*	*	*	*	Machinery defect detection	Feature	Sparse-AE
[13]	*	*	*	*	*	*	*	Gearbox fault diagnosis	Decision	RF
[34]	*	*	*	*	*	*	*	Fault detection using IoT and Digital Twine	Data & Feature	ANN
[14]	*	*	*	*	*	*	*	Target part quality: Surface roughness	Feature	Fuzzy Reasoning
[35]	*	*	*	*	*	*	*	Multi-material network design for process monitoring	Data	DCNN
[36]	*	*	*	*	*	*	*	Smart Anomaly detection using HCAE network	Data	DCNN
[19]	*	*	*	*	*	*	*	Quality monitoring base on under/overprinting states	Data	SIFT
[20]	*	*	*	*	*	*	*	Height control of metal AM	Data	PCP ¹⁵
[5]	*	*	*	*	*	*	*	Enhancing part's UTS by using laser treatment	Data	Manual Comparison

¹Principal Component Analysis (PCA) ²Support Vector Machine (SVM) ³Scale-Invariant Feature Transform (SIFT) ⁴Random Sample Consensus (RANSAC) ⁵Artificial Neural Network (ANN) ⁶Dirichlet Process Mixture (DPM) ⁷Evidence Theory (ET) ⁸Deep Convolutional Neural Network (DCNN) ⁹Long-term Recurrent Convolutional (LRCN) ¹⁰Finite Element Analysis (FEA) ¹¹Random Forest (RF) ¹²Ridge Regression (RR) ¹³Least Absolute Shrinkage & Selection Operator (LASSO) ¹⁴Deep Long and Short Term Memory network (D-LSTM) ¹⁵Point Cloud Processing (PCP)

implement an active control to prevent new occurrences although there has been a defective feature recognized. In fact, this approach wouldn't fix the previously made errors but aims to keep the part's integrity and quality intact.

In line with this method, Borish et al. [20], [21] presented a decision-level fusion for process control of LSMAM and LSPAM methods, respectively. They utilized a surface profilometer and a thermal camera to study the ongoing process. The machine's feed rate passively would be adjusted based on analyzing sensory input from the surface quality once a layer print is done. Based on the temperature maps gathered by the thermal camera, the process would pause until the desired temperature is achieved. Their proposed method would eliminate under/overprint defects and layers' instability due to the high temperatures of previous layers.

Vandone et al. [8] also proposed a data-driven control approach based on multiple online sensory data and offline mechanical test results fusion to decrease the print's deviation from the goal tolerances. Their proposed model, by actively monitoring and controlling the laser profile and melt pool dimensions, aims to eliminate variations from manufacturing standards and minimize the machine's parameters' total changes, also guaranteeing

mechanical properties within expectation.

Garmendia et al. [26] presented a layer height controller based on structured light to ensure the part's overall quality. The author used the point-cloud data gathered with structured light coupled with process plan data extracted from the CAD model and G-codes to determine the deviation from the anticipated partially manufactured artifact. Once the deviation is calculated, a negative defect implementer would engage in adjusting the future layer deposition parameters.

B. Addressing Existing Defects

Despite the addressing future defects method, which aims to create appropriate changes to compensate for the previous deficiencies, the addressing existing defects approach focuses on innovative engineering solutions to eliminate the errors. This approach eliminates the defect before the next deposition, which not only implies that the part's quality is within tolerances and acceptable, but also all the layers are manufactured within tolerance and evenly. Addressing existing defects approach lowers uncertainties in parts quality and properties fluctuation and also increases productivity and resiliency of the AM processes. Here a few instances

of the addressing existing defects approach in the peers work are summarized.

Inasaki [14] suggested using fuzzy reasoning and genetic algorithms to control grinding processes. Their sensor fusion and control strategy approach eliminated the need for parts' post-processing due to off-tolerance surface roughness by implementing an intelligent grinding cycle adjuster. Based on the part's surface quality and fuzzy logic, the process is adopted to achieve the desired quality.

Also, Sardinha et al. [38] proposed a method called ironing for ABS material in Fused Filament Fabrication (FFF) printers for quality and property advancement. This method utilizes the available printer's head to perform a thermal treatment for the top layer of the manufactured part. This method not only showed significant capabilities in lowering the surface roughness of the artifact but also aided the process of preventing defects such as warping to happen and residual stress reduction. Integrating this method whenever the surface quality of the manufactured layer deviates would allow a smooth fabrication and predictable results.

Chen et al. [39] integrated a laser module in-situ with an FFF printer. This laser module would act as an interactive agent to mitigate surface roughness anomalies in this case study. The proposed method not only showed effective improvement of surface quality and elimination of surface anomalies but also showed that the laser-based thermal treatment can increase the mechanical properties of the part, such as storage modulus (E') and loss modulus (E''). Laser polishing showed to be helpful in keeping the material above or around the glass transition temperature, which is a crucial factor in AM fabricated parts' quality. This example can also expand to the CNC machines as [40] illustrated the capacities for integration of the CNC machines with 3D printers.

V. FUTURE WORK

One of the common bottlenecks in the peer's work was reliance on extensive training for defect detection. Building the training set, in many cases, is expensive and time-consuming due to the nature of the AM processes. A deep hybrid network could be useful for automating the data training in order to reduce the amount of data needed to

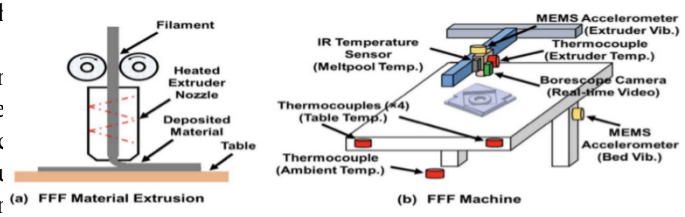


Fig. 5: a) Schematics of FDM machine, b) Schematics of multiple sensors implemented for in-situ data acquisition [24]

train the classifiers and regression required models for anomaly detection. On the other hand, most of the proposed works are designed to minimize the final part's error and deviation by reducing the new defect's occurrence. In other words, most controllers are designed to affect the printed layers while the faulty sections are already deposited and won't change. Nevertheless, if an active, interactive controller were developed, a defect could be mitigated, and a path for future layer deposition could be determined if an anomaly were to be detected. In summary, by benefiting from deep learning capabilities for sensor fusion and online process and part monitoring, a robust control within process time that would focus on future defect prevention and provide alternatives for mitigation or elimination of the detected defects can be achieved.

VI. CONCLUSION

Based on this literature review and the discussions presented, the significance of each sensory module and a summary of a few breakthroughs in AM technologies for abnormality detection have been illustrated. Reviewed literature proposed various ideas to capture process faults in-situ, and prevent additional material and time loss, for example, early stopping the process for parts with unacceptable deviations from the CAD model before fabrication completion. Many steps have also been taken to create a generalized multisensory technique to capture as many defects as possible with higher repeatability and accuracy. However, there is still more room to advance the algorithms' capabilities further to achieve a smarter detection for other varieties of defects and generate a more in-depth correlation between data features. Then,

sensory data fusion approaches and their capabilities and limitations have been discussed. State-of-the-art studies based on sensory data fusion, primarily using Deep Neural Networks, can now predict the system's faults or deviations while having the minimum effort as in training time or training data. Having all the above achievements in the detection field, the author found there are still gaps for error mitigation and compensation field. Sensory data fusion, combined with the control system, can be used for defect mitigation. However, this combination requires much further studies, and there are plenty of opportunities for researchers to take on and expand as in expanding current fusion strategies for control applications. Control algorithms so far mainly focus on future error mitigation and aim to compensate for past deviations with alternations in future depositions. However, a very efficient approach would be utilizing interactive modules to address the anomalies, clear out the part's anomalies to continue the process. As such, ideas of utilizing a CNC robotic arm or a laser module in-situ with the print to remove or mitigate the anomalies would be a great exemplary area to explore.

REFERENCES

- [1] H. Raji, M. Tayyab, J. Sui, S. R. Mahmoodi, and M. Javanmard, "Biosensors and machine learning for enhanced detection, stratification, and classification of cells: A review," *arXiv preprint arXiv:2101.01866*, 2021.
- [2] A. Safari, S. Mehralian, and M. Teshnehlab, "Full-car active suspension system identification using flexible deep neural network," in *2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS)*, pp. 191–198, 2020.
- [3] R. Hassan, S. HekmatiAthar, M. Taheri, S. Cesmecci, and H. Taheri, "Regression model for structural health monitoring of a lab scaled bridge," in *NDE 4.0 and Smart Structures for Industry, Smart Cities, Communication, and Energy*, vol. 11594, p. 115940G, International Society for Optics and Photonics, 2021.
- [4] S. P. H. Athar, M. Taheri, J. Secrist, and H. Taheri, "Neural network for structural health monitoring with combined direct and indirect methods," *Journal of Applied Remote Sensing*, vol. 14, no. 1, p. 014511, 2020.
- [5] N. Sabyrov, A. Abilgazyev, and M. H. Ali, "Enhancing interlayer bonding strength of fdm 3d printing technology by diode laser-assisted system," *International Journal of Advanced Manufacturing Technology*, vol. 108, pp. 603–611, 5 2020.
- [6] V. E. Kuznetsov, A. N. Solonin, A. G. Tavitov, O. D. Urzhumtsev, and A. H. Vakulik, "Increasing of strength of fdm (fff) 3d printed parts by influencing on temperature-related parameters of the process," *Rapid Prototyping Journal*, 2019.
- [7] zhifen Zhang, G. Wen, and S.-B. Chen, *Multisensory Data Fusion Technique and Its Application to Welding Process Monitoring*. 2016.
- [8] A. Vandone, S. Baraldo, and A. Valente, "Multisensor data fusion for additive manufacturing process control," *IEEE Robotics and Automation Letters*, vol. 3, pp. 3279–3284, 10 2018.
- [9] P. K. Rao, J. Liu, D. Roberson, and Z. Kong, "Sensor-based online process fault detection in additive manufacturing," 2015.
- [10] D. Wu, Y. Wei, and J. Terpenney, "Predictive modelling of surface roughness in fused deposition modelling using data fusion," *International Journal of Production Research*, vol. 57, pp. 3992–4006, 2019.
- [11] D. Wu, Y. Wei, and J. Terpenney, "Surface roughness prediction in additive manufacturing using machine learning," *ASME 2018 13th International Manufacturing Science and Engineering Conference, MSEC 2018*, vol. 3, pp. 1–6, 2018.
- [12] S. Zhang, Z. Sun, J. Long, C. Li, and Y. Bai, "Dynamic condition monitoring for 3d printers by using error fusion of multiple sparse auto-encoders," *Computers in Industry*, vol. 105, pp. 164–176, 2 2019.
- [13] C. Li, R. V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mechanical Systems and Signal Processing*, vol. 76–77, pp. 283–293, 8 2016.
- [14] I. Inasaki, "Sensor fusion for monitoring and controlling grinding processes," 1999.
- [15] F. Baumann and D. Roller, "Vision based error detection for 3d printing processes," *MATEC Web of Conferences*, 2016.
- [16] S. Nuchitprasitchai, M. Roggemann, and J. Pearce, *Three Hundred and Sixty Degree Real-Time Monitoring of 3-D Printing Using Computer Analysis of Two Camera Views*, vol. 1. 2017.
- [17] F. Wang, F. Ju, K. Rowe, and N. Hofmann, "Real-time control for large scale additive manufacturing using thermal images."
- [18] X. Li, S. Siahpour, J. Lee, Y. Wang, and J. Shi, "Sciencedirect deep learning-based intelligent process monitoring of directed energy deposition in additive manufacturing with thermal images," 2020.
- [19] W. Lin, H. Shen, J. Fu, and S. Wu, "Online quality monitoring in material extrusion additive manufacturing processes based on laser scanning technology," *Precision Engineering*, vol. 60, pp. 76–84, 11 2019.
- [20] M. Borish, B. K. Post, A. Roschli, P. C. Chesser, L. J. Love, and K. T. Gaul, "Defect identification and mitigation via visual inspection in large-scale additive manufacturing," *JOM*, vol. 71, pp. 893–899, 3 2019.
- [21] M. Borish, B. K. Post, A. Roschli, P. C. Chesser, and L. J. Love, "Real-time defect correction in large-scale polymer additive manufacturing via thermal imaging and laser profilometer," vol. 48, pp. 625–633, Elsevier B.V., 2020.
- [22] H. Wu, Y. Wang, and Z. Yu, "In situ monitoring of fdm machine condition via acoustic emission," *The International Journal of Advanced Manufacturing Technology*, vol. 84, pp. 1483–1495, 2016.
- [23] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," 1997.

- [24] P. K. Rao, J. P. Liu, D. Roberson, Z. J. Kong, and C. Williams, "Online real-time quality monitoring in additive manufacturing processes using heterogeneous sensors," *Journal of Manufacturing Science and Engineering*, vol. 137, p. 061007, 2015.
- [25] *Camera-Based Coaxial Melt Pool Monitoring Data Registration for Laser Powder Bed Fusion Additive Manufacturing*, vol. Volume 2B: Advanced Manufacturing of ASME International Mechanical Engineering Congress and Exposition, 11 2020. V02BT02A045.
- [26] I. Garmendia, J. Pujana, A. Lamikiz, M. Madarieta, and J. Leunda, "Structured light-based height control for laser metal deposition," *Journal of Manufacturing Processes*, vol. 42, pp. 20–27, 2019.
- [27] J. Harbig, D. L. Wenzler, S. Baehr, M. K. Kick, H. Merschroth, A. Wimmer, M. Weigold, and M. F. Zaeh, "Methodology to determine melt pool anomalies in powder bed fusion of metals using a laser beam by means of process monitoring and sensor data fusion," *Materials*, vol. 15, no. 3, 2022.
- [28] Z. Snow, E. W. Reutzler, and J. Petrich, "Correlating in-situ sensor data to defect locations and part quality for additively manufactured parts using machine learning," *Journal of Materials Processing Technology*, vol. 302, p. 117476, 2022.
- [29] Q. Tian, S. Guo, E. Melder, L. Bian, and W. G. Guo, "Deep Learning-Based Data Fusion Method for In Situ Porosity Detection in Laser-Based Additive Manufacturing," *Journal of Manufacturing Science and Engineering*, vol. 143, 12 2020. 041011.
- [30] R. Zou, X. Liang, Q. Chen, M. Wang, M. A. S. Zaghoul, H. Lan, M. P. Buric, P. R. Ohodnicki, B. Chorpening, A. C. To, and K. P. Chen, "A digital twin approach to study additive manufacturing processing using embedded optical fiber sensors and numerical modeling," *Journal of Lightwave Technology*, vol. 38, no. 22, pp. 6402–6411, 2020.
- [31] C. Zamiela, W. Tian, L. Bian, J. Zhipeng, and Z. Tian, "Deep fusion methodology of infrared and ultrasonic data for porosity detection in additive manufacturing," *IIE Annual Conference.Proceedings*, pp. 229–234, 2021. Copyright - Copyright Institute of Industrial and Systems Engineers (IISE) 2021; Last updated - 2021-09-09.
- [32] C. E. Zamiela, *Deep Multi-Modal U-Net Fusion Methodology of Infrared and Ultrasonic Images for Porosity Detection in Additive Manufacturing*. PhD thesis, 2021. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2022-01-27.
- [33] J. Francis and L. Bian, "Deep learning for distortion prediction in laser-based additive manufacturing using big data," *Manufacturing Letters*, vol. 20, pp. 10–14, 2019.
- [34] R. M. Scheffel, A. A. Fröhlich, and M. Silvestri, "Automated fault detection for additive manufacturing using vibration sensors," *International Journal of Computer Integrated Manufacturing*, vol. 34, no. 5, pp. 500–514, 2021.
- [35] V. Pandiyan, R. Drissi-Daoudi, S. Shevchik, G. Masinelli, T. Le-Quang, R. Logé, and K. Wasmer, "Deep transfer learning of additive manufacturing mechanisms across materials in metal-based laser powder bed fusion process," *Journal of Materials Processing Technology*, vol. 303, p. 117531, 2022.
- [36] *In-Situ Laser-Based Process Monitoring and In-Plane Surface Anomaly Identification for Additive Manufacturing Using Point Cloud and Machine Learning*, vol. Volume 2: 41st Computers and Information in Engineering Conference (CIE) of International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 08 2021. V002T02A030.
- [37] C. Li, D. Cabrera, F. Sancho, R.-V. Sanchez, M. Cerrada, and J. V. de Oliveira, "One-shot fault diagnosis of 3d printers through improved feature space learning," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 8 2020.
- [38] M. Sardinha, C. M. Vicente, N. Frutuoso, M. Leite, R. Ribeiro, and L. Reis, "Effect of the ironing process on abs parts produced by fdm," *Material Design & Processing Communications*, vol. 3, no. 2, p. e151, 2021.
- [39] L. Chen and X. Zhang, "Modification the surface quality and mechanical properties by laser polishing of al/pla part manufactured by fused deposition modeling," *Applied Surface Science*, vol. 492, pp. 765–775, 2019.
- [40] M. Lalegani Dezaki, M. K. A. Mohd Ariffin, and M. I. S. Ismail, "Effects of cnc machining on surface roughness in fused deposition modelling (fdm) products," *Materials*, vol. 13, no. 11, p. 2608, 2020.

Unmanned Aerial Vehicle Control Using Hand Gestures and Neural Networks

Jack Nemeec and Dr. Rocio Alba-Flores
 Department of Computer and Electrical Engineering
 Georgia Southern University
 Statesboro, GA

jn08679@georgiasouthern.edu and ralba@georgiasouthern.edu

Abstract—Neural Networks are a series of data manipulations inspired by how neurons perceive information in the brain. This technology is useful for accomplishing tasks that conventional computers do poorly, but people do accurately. Neural Networks are utilized in this project to control an Unmanned Aerial Vehicle (UAV) with hand gestures. For this case the model produced by TensorFlow will take twenty-one different hand points on a user's hand using MediaPipe and distinguish which of eight gestures the user is signaling. This data is received through a camera on the UAV and once ran through the model the flight path will be controlled. The hand points are logged as two-dimensional coordinates in relation to the pixel they are in the frame. This creates a model with forty-two inputs and nine outputs. The model can run at around twenty frames per second due to the low number of inputs. The UAV can handle efficiently due to an acceptable processing time of its commands.

Keywords—Neural Networks, machine learning, artificial intelligence, neurons, Unmanned Aerial Vehicle (UAV).

I. INTRODUCTION

Neural Networks were originally modeled by neurophysiologist Warren McCulloch and mathematician Walter Pitts in 1943 [1]. The idea was to solve computer problems in a similar way that a brain would. The neurons in the brain have one job and that is whether to fire off a signal or not, much like a digital computer. What makes neurons different is how they are connected. One neuron can be attached to many other neurons that control the conditions in which it signals a response. Each of these input neurons have a different “weight” or bias. These weights can be very small or very large depending on the purpose of the neuron. Unlike computer logic which has two conditions, high or low, the brain works much like an analog computer rather than a digital one. The idea that this system of weights and biases can be applied to a computer is what created the idea of a

neural network. Originally neural networks were not very accurate. Two things improved this, loss functions and multiple layered networks. Loss functions take the accuracy of the network from a training set and adjusts the weights of the network based on the error [2]. Originally, networks were just an input layer, the neurons, and the output layer. The neuron layer could be expanded to multiple layers with different numbers of neurons and different weights to improve accuracy. With these two improvements and lots of research neural networks are used in many different computer applications today where a conventional program would not suffice.

To track the hands in this project the software developed by MediaPipe was utilized. MediaPipe is a software that is compatible with JavaScript, Python, C++, IOS, and Android [3]. It has applications that can track body poses, hands, objects, faces, and so much more making it an incredibly useful software. For the purpose of this project the hand tracking was utilized with Python. The software uses twenty-one hand key points starting at the bottom of the palm, to every joint, ending on each fingertip.

To model the network TensorFlow is used. TensorFlow is an end-to-end open-source platform for machine learning [4]. It makes modeling complex data easy, and it used by many well-known companies such as Google, Coca-Cola, Twitter, Airbnb, and intel. TensorFlow has applications for all levels of machine learning. Both TensorFlow and MediaPipe have python libraries increasing the usability.

The UAV used for this experiment is the Tello drone developed by Ryze Robotics. Tello drones are small and simple making them great for indoor work. They use an Intel 14-Core processor and contain features such as automatic takeoff/landing, low battery protection, collision detection, failsafe protection, and a vision positioning system. The Tello drone has a camera on the front that is used for gathering the visual data needed for this project. The drone connects to the use via Wi-fi and can be programmed using the Tello Python library.

The IDE used for this project is PyCharm developed by Jet Brains. The version of Python used

is Python 3.8. PyCharm is a very easy use IDE that has predictive text and quick error fixing. This IDE also allows for easy download of libraries. It will open the library documentation within the IDE which allows for ease of use when reading library codes.

The gestures used for this project are shown in Fig. 1. There are eight different gestures that control eight different commands. Gesture one and two commands the drone to move forward or backward. Gesture three commands the drone stops. Gesture four and five commands the drone to move up or down. Gesture six and seven commands the drone to move left or right. Gesture eight commands the drone to land. To always keep the hand in the center of the frame the drone will rotate left or right depending on where the hand is located. This concludes the motion of the drone.

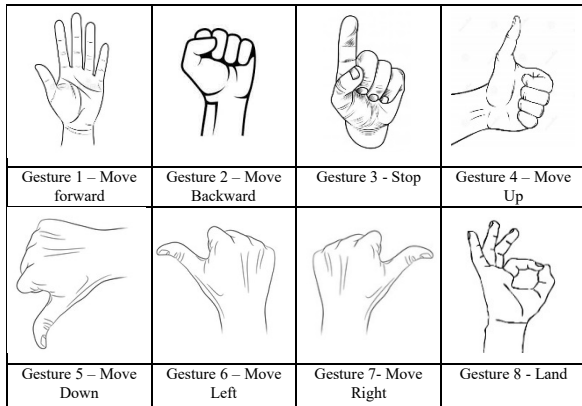
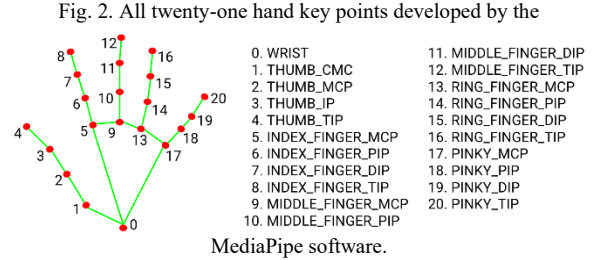


Fig. 1. Eight gestures and their respective commands.

II. METHOD

The MediaPipe software uses twenty-one hand key points as seen in Fig. 2. These are set on the knuckles and tips of the hand. To obtain images of the hand, an external webcam was used for quality purposes. The images were processed using OpenCV which is an image processing software. When the hands are found in the image the MediaPipe software draws points on each landmark, and the lines between them. Then it takes the data on each point's x-coordinate and y-coordinate. The value of which are the pixel they are on in the camera feed. The top left point of the camera feed is the origin. As you move to the right or down the number representing each pixel is increased. To gather the data a 640x480 pixel width times height frame was used. The data in this twenty-one by two list will be used as the data for the neural network. The software can track both hands at once, however, for the purpose of this experiment only one hand was tracked at one time. The data set does contain both hands however as to not discriminate against which hand is being utilized.



To build the training set a code was made that transfers a list of all the hand key point coordinates to an excel file. Whenever the space bar is pressed down the program will append all the hand key points seen in an image to a new list. This allowed many sets of coordinates to be logged without using individual photos. This list contains forty-two coordinates, or an x-coordinate and y-coordinate for each of the twenty-one hand key point. Data is taken from both left and right hands at distances from ten centimeters up to five meters; and orientations to ensure a large variety of data in the training set. Once one gesture is fully logged from all distances and orientations, the data is added to a large excel file that contains all twenty-one coordinates and the gesture that corresponds. A small sample of which can be found in Fig. 3 and Fig. 4. Data was taken from three different people to add to variety in the data. The more variety the data has the more accurate the model will be.

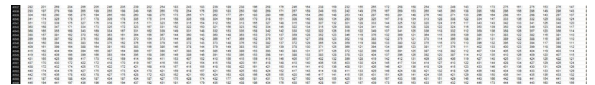


Fig. 3. A sample of the training data set. This contains the last few lines of the set.

0x	0y	20x	20y	Gestures
300	426	350	313	1
300	433	349	321	1

Fig. 4. Example of training set data for x and y coordinates of hand key point 0 and 20. The gesture that corresponds to these points is on the right.

The training set, once completed, is converted into a CSV file. A CSV file is like an excel file but without formatting [5]. Instead, each column is separated by comma instead of a formatted cell. Due to the simple formatting the computer can process the data much quicker.

The TensorFlow model used for this experiment was the Sequential Model which is the most common model used for a plain input. There are forty-two inputs or each of hand key point's x and y coordinate.

The three hidden layers consists of sixty-four neurons, thirty-two neurons, and sixteen neurons. The output has nine neurons which are the eight gestures plus no gesture when a wrong signal is given. The training set is normalized, meaning that TensorFlow converts all the coordinates to values between zero and one. Without this the model would not be able to handle the data given to it. The optimizer used is named ‘Adam’. The Adam optimizer is a stochastic gradient decent method [5]. The loss function used is sparse categorical cross entropy [6]. This loss function is best used when there are two or more label classes. For this experiment there is eight. The model is then fit using sixty-four epochs each with a batch size around twelve thousand. TensorFlow calculates accuracy and loss while training which is displayed to the user for each epoch.

With a trained model real time data can be given to it. Every time the code loops a frame is taken from the drone, twenty-one key points are added to a hand in frame, if there is no hand the code will keep looping, the pixel coordinates of each hand key point is then normalized and sent to the neural network, the network chooses the highest probability or the most likely gesture, then a command is sent to the drone respective of that gesture. This process repeats up to twenty times per second with a usual framerate between fifteen and twenty. Every part of the methodology is done using a python code with many imported libraries and modules such as MediaPipe, OpenCV, TensorFlow, etc.

III. RESULTS AND DISCUSSION

The training set contains three different people’s hands independent of left or right, and up to 37,000 lines of coordinate sets. 8,000 set of coordinates are for gesture one, 3,000 for gesture two, 5,000 for gesture three, 5,500 for gesture four, 5,600 for gesture five, 3,200 for gesture six, 3,500 for gesture seven, and 4,000 for gesture eight. It produces an accuracy of 94% as seen in fig. 5. A testing set was made that contains 200 sets of coordinates for each gesture. The accuracy and loss of the testing set can be seen in fig. 6. The confusion matrix for this testing set can be seen in fig. 7.

loss: 0.2441 - accuracy: 0.9439

Fig. 5. Training result of neural network.

loss: 0.0711 - accuracy: 0.9744

Fig. 6. Testing result of neural network.

1	0	0	0	0	0	0	0
0	0.985	0	0	0	0.005	0.01	0
0	0	1	0	0	0	0	0
0	0	0.005	0.995	0	0	0	0
0	0	0	0	1	0	0	0
0.02	0	0	0	0	0.845	0.135	0
0.02	0	0	0	0	0	0.98	0
0.005	0.005	0	0	0	0	0	0.99

Fig. 7. Confusion Matrix of testing set.

The model has a more difficult time differentiating the hand gestures that are a thumb pointed in a certain direction. Most likely this is because of errors in the data set mainly for gesture six. As seen in Fig. 7. Gesture six is the most inaccurate and conflicts with both gesture two and seven.

The model is increasingly accurate when the hand is at a close distance such as within two meters. It becomes much less accurate at farther distances greater than five meters. The biggest reason for this is because the MediaPipe software struggles to classify the hand at far distances. This leads to errors in data and sometimes no data at all. This is not practical for large UAV operations that require large distances; however, it is useful for indoor operations. Ways to improve upon this in the future is to be able to adjust the image frame to only include the hand no matter the distance or orientation. Also, to improve drone camera quality.

The reason the input to the model is forty-two hand key points instead of a whole frame is due to the speed of the program and processing power of the computer. To process twenty frames per second a GPU or Graphics Processing Unit is a necessity [8]. The frame used in the training set was 640x480 pixels with an area of 307,200 pixels vs 42 coordinates as the network’s input. To run a neural network with this many inputs would be intensive on any computer without a GPU. Most of this project however was done using a laptop with only 16GB of RAM and it could handle processing the images up to twenty frames per second because of the low number of inputs needed for the network. With a GPU it would be much faster. Neural networks are a great way to solve problems that computers typically cannot. However, if your average household computer cannot run it then it is impractical.

REFERENCES

[1] C. Clabaugh, D. Myszewski, and J. Pang, “Neural Networks,” *Neural networks - history*, 2000. [Online]. Available: <https://cs.stanford.edu/people/eroberts/courses/so-co/projects/neural-networks/History/history1.html>. [Accessed: 30-Mar-2022].

- [2] J. Brownlee, "Loss and loss functions for training Deep Learning Neural Networks," *Machine Learning Mastery*, 22-Oct-2019. [Online]. Available: <https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>. [Accessed: 13-May-2022].
- [3] "Home," *mediapipe*, 2020. [Online]. Available: <https://google.github.io/mediapipe/>. [Accessed: 30-Mar-2022].
- [4] "Why tensorflow," *TensorFlow*. [Online]. Available: <https://www.tensorflow.org/about>. [Accessed: 30-Mar-2022].
- [5] "Load CSV data : Tensorflow Core," *TensorFlow*. [Online]. Available: https://www.tensorflow.org/tutorials/load_data/csv v. [Accessed: 13-May-2022].
- [6] "Tf.keras.optimizers.Adam : Tensorflow core v2.8.0," *TensorFlow*. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam. [Accessed: 30-Mar-2022].
- [7] "Tf.keras.losses.SparseCategoricalCrossentropy : Tensorflow core v2.8.0," *TensorFlow*. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/losses/SparseCategoricalCrossentropy. [Accessed: 30-Mar-2022].
- [8] F. F. d. Santos *et al.*, "Analyzing and Increasing the Reliability of Convolutional Neural Networks on GPUs," in *IEEE Transactions on Reliability*, vol. 68, no. 2, pp. 663-677, June 2019, doi: 10.1109/TR.2018.2878387.

Feature Selection Algorithm Characterization for NIDS using Machine and Deep learning

Jyoti Verma
 Department of Computer Science and
 Engineering
 Punjabi University Patiala,
 Punjab, India
 rs_jyoti@csepup.ac.in

Abhinav Bhandari
 Department of Computer Science and
 Engineering
 Punjabi University Patiala,
 Punjab, India
 bhandarinitj@gmail.com

Gurpreet Singh
 Punjab Institute of Technology, Rajpura
 (MRSPTU, Bathinda)
 Punjab, India
 myselfgurpreet@gmail.com

Abstract— Data dimensionality is increasing at a rapid rate, posing difficulties for traditional mining and learning algorithms. Commercial NIDS models make use of statistical measures to analyze feature sets including packet length, inter-arrival time, and flow size, in addition to other internet traffic parameters. Emerging algorithms must deal with diverse data. While multiple deep learning-based solutions exist in the literature, their commercialization is still in its infancy. Currently available machine learning techniques create a large number of false positives. As the quantity of data to be processed has increased in recent years, feature selection (FS) appears to have become a basic requirement for any type of model. The recent advent of promising techniques and different kinds of features advances existing computational research and continuously improves feature selection, seeking to make it applicable to a broader range of applications. This paper intends to provide a fundamental investigation to feature selection throughout NIDS, which will take into account basic concepts, categorization of existing systems, a framework and taxonomy for NIDS, the feature selection methods used by researchers to develop NIDS methods, and a comparison of FS Algorithm classification and Python FS library contents. By examining existing contributions, this study provides an overview of the majority of techniques proposed in the feature selection research. Additionally, we discuss the most recent FS algorithms for NIDS that were established to select the optimal feature subsets. By classification and comparative study, the paper provides a road map for comprehending and constructing the current state of NIDS FS. As a result, a study is presented to help the reader comprehend the research progress, the FS Algorithm's characterization, and the establishment of a new taxonomy for emerging developments and existing challenges.

Keywords— NIDS, Feature selection, Machine learning, Python, Deep Learning

I. INTRODUCTION

In recent years, attention has been focused on FS (Feature Selection) techniques and optimization techniques for selecting the most important features. Feature selection has been demonstrated to be an efficient and effective method for prepping large-dimensional data for machine learning and data mining[1]. All current NIDS (Network Intrusion Detection System) use characteristics of user and system behavior as input to their analysis algorithms, which determine the likelihood of an attack. While numerous network protection methods exist, such as access control, encryption, authentication, and advanced firewalls, there is an urgent need for intelligent NIDS that can automatically detect known and novel attacks[2]. A NIDS is composed of several stages, including data collection, pre-processing, FS, and classification. The FS phase is difficult, as the NIDS must deal with a massive amount of data.

The mathematical equation of FS is as follows: 6-Tuplet FS = D, F, C, S, fs, E, where D denotes a dataset and F denotes a feature set. $D = \{i_1, i_2, \dots, i_m\}$, where m is the number of instances. F is a collection of characteristics. $F = \{f_1, f_2, \dots, f_n\}$ with n features. C is a class that is intended to be used. $C = \{c_1, c_2, \dots, c_k\}$ where k denotes the target classes. S is a search space of the set F that includes all subsets that can be constructed using F. $S = \{s_1, s_2, \dots, s_l\}$ ($l = 2^n - 1$: NP-Hard) with $s_i = \{f_{j_1}, f_{j_2}, \dots, f_{j_n}\}$ ($1 \leq j_1 \neq j_2 \neq \dots \neq j_n \leq n$), where E is evaluation measure and fs represents the feature selection process: $fs: F \rightarrow S$ [3]. Selecting the best feature selection method hinge on on the input and output. This paper intends to provide a basic study of feature selection throughout NIDS, including a framework and taxonomy for NIDS, categorization of existing systems, FS methods used by researchers in developing NIDS methods, and a summary of FS Algorithm categorization and Python FS library contents[4].

The remaining sections of the paper are organized as follows. In section 2, we represent a literature review for feature selection methods used in NIDS. Feature selection framework and taxonomy for intelligent NIDS in section 3. In section 4, a comparison of FS Algorithm characterization and Python FS libraries contents is done. In section 5, we present a conclusion.

II. LITERATURE SURVEY

Feature selection provides an insight into the current methods for dealing with a variety of problem classes[5]. K. Vamsi Krishna et al. presented a Least Variance Feature Elimination approach for feature selection in a Network Intrusion Detection System in a cloud environment to reduce computational time without sacrificing detection rate[6]. Rawaa Ismael Farhan et al. proposed Enhanced BPSO and CFS correlation-based feature selection on DNN classifiers and the CSE-CICIDS2018 dataset [7]. Amrita et al. used a machine learning technique to perform a feature selection and technical evaluation survey on anomaly-based NIDS. Their paper offered an empirical example of how to choose a more appropriate solution [8]. Zhao Yongli et al. describe an enhanced feature selection algorithm based on Distance feature ranking and an enhanced exhaustive search to select a more advantageous combination of features. On the KDD CUP datasets with reduced feature subsets, they used SVM and KNN classifiers [9]. K. Vamsi Krishna et al. proposed a Hybrid Feature Selection algorithm that uses a mutual information-based feature selection algorithm to analytically select the optimal feature for classification. In comparison to state-of-the-art methods, feature selection using Least square Support Vector Machine-based IDS (LSSVM-IDS) achieves higher accuracy at a lower computational cost [10]. Pankaj Kumar Keserwani et al. proposed an anomaly-based NIDS combined with a bald eagle search algorithm for VCN. The classification was performed using a deep sparse auto-encoder (DSAE). Python was used to simulate the proposed system. [11]. Jabali et al. provided an overview of the Intrusion Detection taxonomy and current feature selection techniques, as well as details on IDS design and development issues. It is investigated for dimensionality reduction to determine which methods achieve a higher level of accuracy and workload reduction, followed by existing techniques for comparing classifiers and classifier designs[12]. Gholamreza Farahani et al. proposed a new cross-correlation-based feature selection method and compared it to the cuttlefish algorithm and mutual information-based feature selection methods [13]. Badii Atta et al. described an integrative knowledge representation framework for virtualization, focusing on a topic map-enabled representation of features for pattern recognition within the NIDS domain[14]. Kok-Chin Khor et al. proposed a method for identifying critical features by combining two feature selection algorithms with a filter approach. Domain experts validated the selected features and added additional features to the final proposed feature set [15]. Hebatallah Mostafa Anwer et al. presented a framework for selecting

features for efficient network anomaly detection using a variety of machine learning classifiers based on filter and wrapper methodologies[16]. Mubarak Albarka Umar et al. proposed a method for developing IDS models using the UNSW-NB15 dataset that combines FS and machine learning algorithms. Their method achieved the highest degree of reproducibility (DR) of 97.95 percent[17]. Chaouki Khammassi et al. selected the best subset of features for network intrusion detection systems using a wrapper approach and logistic regression learning algorithm [18].

III. FEATURE SELECTION FRAMEWORK AND TAXONOMY FOR NIDS

FS is the method of selecting unique features based on specific criteria in machine learning[19]. It is a significant and widely used dimension reduction technique in many fields. When selecting features, there are two parts to a selection method: Feature set selection strategy and an evaluation technique for evaluating feature subsets. Listed below are the activities for selecting a subset of features[4]. Because the dependent variables lack sufficient additional information about the classes, they act as noise for the predictor.

TABLE 1: FEATURES TYPES

Feature Type	Description	Accuracy	Training Time
Strongly Relevant	It is necessary for an optimal feature subset	High	Low
Irrelevant or Noisy	It cannot help discriminate between supervised or unsupervised data.	Low	High
Redundant or Weakly Relevant	It can be completely replaced with a set of other features	Low	High
Weakly relevant or non-redundant	It may not always be necessary for an optimal subset and depend on certain conditions.	Low	High
Noisy	It is not relevant to the learning or mining task.	Low	High

We can obtain an overview of the process and enhance the computation necessities and model accuracy by using feature selection techniques. f1 is a pertinent feature that distinguishes between classes 1 and 2. f 2 is an irrelevant feature. f4 is therefore a noisy feature. f6 is a redundant attribute [20]. If f1 is chosen, removing f6 doesn't affect learning performance. The dataset includes insignificant, redundant, and noisy features. Features types are discussed in Table 1. and depicted as a graph in Figure 1.

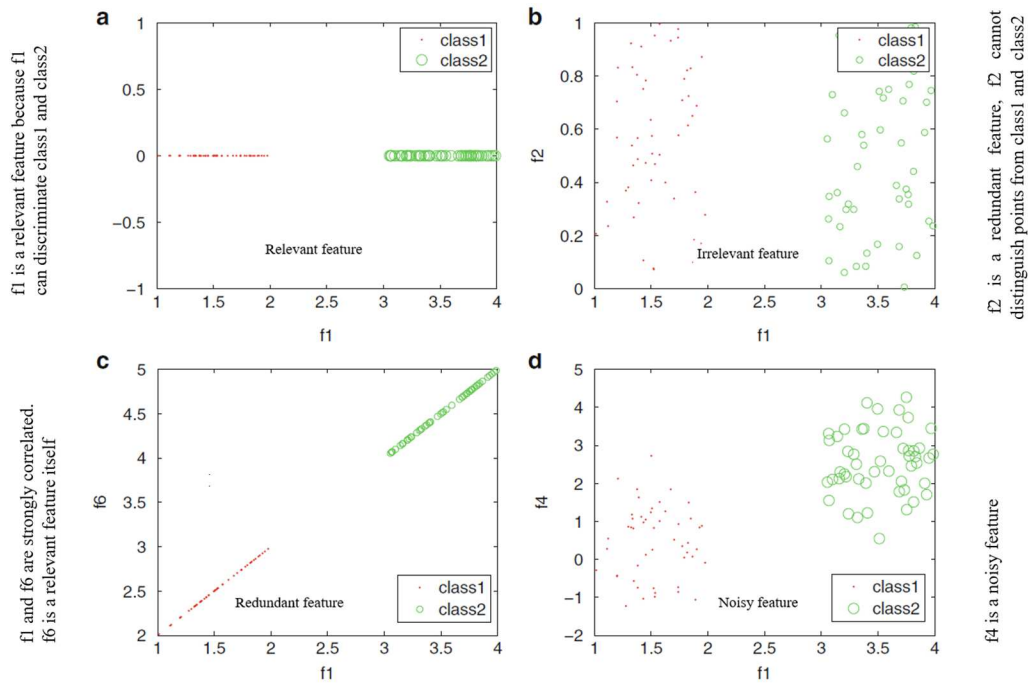


Fig. 1 Relevant, Irrelevant, Redundant, and Noisy features [20]

Feature elimination doesn't always generate new features because it reduces the number of input features. Once the feature selection condition has been chosen, a procedure for locating the subset of important features must be developed. As

the number of features increases, assessing all the feature subsets (2^N) for a data set becomes an NP-hard problem [20]. Classification of NIDS FS methods as shown in Figure 2., described as follows:

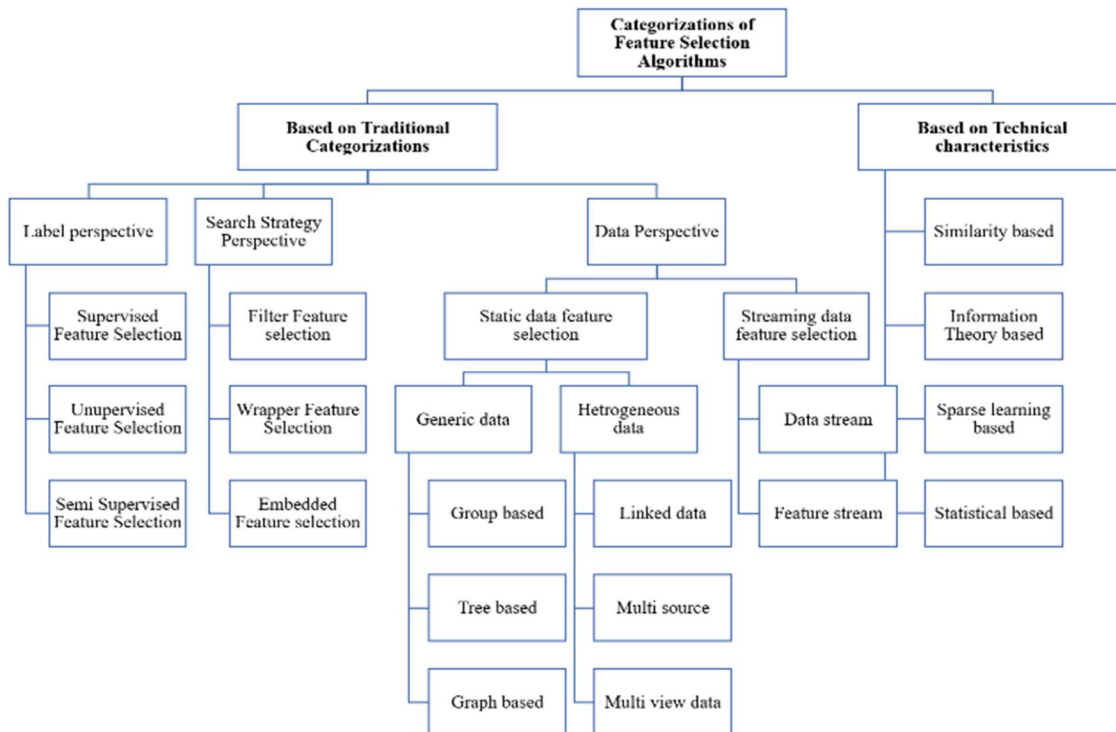


Fig 2 Taxonomy of Feature selection for NIDS

A. Based on Traditional Categorizations

Based on traditional classification feature selection methods can be broadly classified as follows:

1) Label Perspective of Feature Selection Algorithms

The final output is referred to as labels. Based on label perspective feature selection algorithms can be classified as follows:

a) *Supervised Feature Selection*: The goal of the supervised feature selection process is to select a set of features that can distinguish samples collected from different classes and are intended for regression and classification problems. After dividing the train and test sets, classifiers employ supervised feature selection on a subset of features. Based on the selected features, the trained algorithm assumes the class labels of datasets in the test set [21].

b) *Unsupervised Feature Selection*: The Unsupervised feature selection approach is intended for use with clustering problems. Due to the absence of label information to assess feature weights, unsupervised techniques look for alternative criteria to define feature relevance, such as local discriminative information. Unsupervised feature selection makes use of all instances available during the feature selection phase [21].

c) *Semi-Supervised Feature Selection*: For supervised classification, the small amount of labeled data may be inadequate in delivering correlation information about features and for unsupervised methods, class labels may contain useful information for class discrimination. As a result, developing semi-supervised methods that utilize both labeled and unlabeled samples is desirable [22].

2) Search Strategy Perspective Feature Selection Methods

In terms of selection strategies, there are three types of feature selection methods:

a) *Filter methods*: The filter methods are preprocessing procedures that employ variable ranking techniques as the primary criterion for selecting features by ordering the features and applying only the highest-ranked features to a predictor. If a feature is linearly separable from the class labels, it can be considered irrelevant. The process of determining the significance of a feature can be univariate or multivariate. Mutual Information, Correlation, Chi-Square Test, ANOVA Information Gain, Variance Threshold, and Fisher Score are all examples of common filtering techniques [23].

b) *Wrapper methods*: The wrapper method improves the performance of selected features by relying on the predictive accuracy of a predetermined learning algorithm. A typical wrapper method for a given learning algorithm performs two steps: first, the feature set search attribute generates a subset of features; and then the learning algorithm evaluates the quality of these attributes. Wrapper feature selection has several examples, including backward feature elimination, forward feature selection, and recursive feature elimination. Recursive Feature Elimination for Logistic Regression, Boruta, Permutation Importance, and SHAP are all common wrapping techniques [24].

c) *Embedded Methods*: Embedded methods offer mediation between filter and wrapper techniques by combining algorithms with built-in feature selection with model learning that incorporates variable selection during the training process without dividing the test and train sets. Embedded methods are classified into three types: pruning techniques, models with built-in feature selection mechanisms, and regularisation models. LASSO, Elastic Net, and Ridge Regression are all examples of embedded methods. Embedded Random Forest, Embedded LightGBM, and others are frequently used examples [25].

3) Feature selection based on Data Perspective

Based on Data perspective feature selection methods are of the following types:

a) *Static feature selection*: In static data, all data features and instances are known in advance, whereas in streaming data, the number of data instances, features, or both is unknown. Statistical-based techniques entail evaluating the correlation between each input parameter and the target value using statistics and then selecting the input variables with the strongest relationship to the target variable[26]. It can also be classified as generic data or heterogeneous data. Static features could be classified as follows:

i) *Feature Selection on Generic Data*: The focus of previous feature selection techniques for generic data is premised on the false presumption that features are completely independent of one another, oblivious to the intrinsic structure of features. Feature selection methods for generic data can be classified into similarity, information theory, sparse learning, and statistical methods [27].

- *Feature Selection with Group Feature*: Features observe group structures in a wide variety of real-world applications. Each factor is linked to various groups and it can be expressed through the use of a group of dummy features. It is more appealing to consider the group structure of features explicitly when performing feature selection [28].
- *Feature Selection with Tree Feature*: Tree structures can be demonstrated using features that represent a tree, in which the root node, child nodes, and leaf nodes are all represented as features in the spatial locality structure[29].
- *Feature Selection with Graph Feature*: Features exhibit some degree of dependency and can be structured as an undirected graph, with nodes and edges representing pairwise dependencies among features. When there are dependencies between features, we can encode them using an undirected graph $G(N, E)$. Examples include Graph Lasso, GFLasso, and GOSCAR[30].

ii) *Feature Selection with Heterogeneous Data*: Given that the data from each source may be noisy, incomplete, or redundant, determining which sources to use and how to combine them for efficient feature selection becomes a difficult problem. Three

aspects of feature selection techniques for heterogeneous data are as follows:

- *Feature selection for linked data:* In real-world applications, linked data is becoming more ubiquitous. Linked data is distinguished from traditional flat data by the use of various types of links. Eg. Feature Space Selection for Linked Data, Feature Extraction on Network systems [31].
- *Feature selection for multi-source data:* Multi-faceted characterizations of data can enable the outline of several intrinsic patterns hidden insights, the availability of various data sources that allow for the solution of some problems that would otherwise be unsolvable with a single source [31].
- *Feature selection for multi-view data:* Multi-view sources use different features which are innately dependent and high-dimensional, feature selection is asked to undertake these sources for efficient data mining tasks like multi-view clustering [32].

b) Streaming feature selection: By removing redundant and irrelevant, streaming feature selection enables the most insightful features to be selected. In comparison to older methods of feature selection, online feature selection results in models that are easier to interpret for researchers and users, require less training time, avoid issues and challenges associated with dimensionality, and exhibit greater generalization due to reduced overfitting. The selection of streaming features is critical in real-time applications where the action must be taken immediately [31].

i) Feature Selection algorithms for Data Streams: It is critical to develop solutions for dealing with high-dimensional sequential data. One of the most critical characteristics of any effective feature selection approach is its ability to deal with massive amounts of data [31].

ii) Feature Selection Algorithms with Feature Streams: The number of instances is assumed to be constant in the feature selection task with streaming features, while candidate features show up one at a time. Rather than searching the entire feature

space, which is expensive, streaming feature selection (SFS) processes a new feature as it arrives. The process is repeated until no additional features appear [33].

B. Feature Selection based on Categorizations

a) Similarity-based Methods: To define feature relevance a group of methods for determining the significance of features based on their maintaining data similarity is used. These methods are referred to as similarity-based feature selection methods [34].

b) Information Theory-based Methods: Information-theory FS methods comprise a large group of previous feature selection algorithms. Techniques in this family primarily use various heuristic specifications to determine the significance of features. Furthermore, most information-theoretic concepts are only applicable to discrete variables. For example, entropy, conditional entropy, information gain, and conditional information gain [35].

c) Sparse Learning-based Methods: Sparse learning-based FS methods are applied for minimizing fitting errors and sparse regularisation terms. The sparse regularizer makes some feature coefficients small or exactly zero, allowing the corresponding features to be easily eliminated [36].

d) Statistical based Methods: Statistically based FS algorithms examine each feature individually. As a result, feature redundancy is invariably overlooked during the selection stage. T-score, Low Variance-score, Chi-Square Score, T-score, and Gini Index are some examples [36].

IV. EXISTING PYTHON FEATURE SELECTION LIBRARIES

The Python language has grown in popularity in most applications of machine learning [37]. The ML Feature Selection library includes heuristic techniques and algorithms based on machine learning methods and evaluation techniques [35]. Existing open-source repositories in Python with feature selection algorithms are mentioned in Table 2 and a comparison of FS algorithm characterization and Python FS libraries contents in Table 3.

TABLE 2: PYTHON-COMPATIBLE FEATURE SELECTION LIBRARIES FOR NIDS [35]

Machine Learning Library	Full form	Description
Weka	Waikato Environment for Knowledge Analysis	Weka includes different filtering methods and wrappers for feature selection.
MLR	Machine learning libraries for R	It includes a filter, a wrapper, a filter ensemble, and clustering.
ITMO	Information Technologies Mechanics and Optics	This FS library architecture includes wrappers, filters, hybrid, and embedded components.
ASU	Arizona State University library	It is created by Arizona State University's machine learning and data mining Lab.
SKL	scikit-learn library	It is an accessible FS repository based on scikit-learn.
FES	FES book support code	It focuses on evolutionary algorithm
Caret	Classification and Regression Training	Automate the process of developing predictive models in R.

TABLE 3: COMPARISON OF FS ALGORITHM CHARACTERIZATION AND PYTHON FS LIBRARIES CONTENTS[34]

FS Algorithm	Label	Learnin g	Output	Feature Type	ITM O	ASU	SK L	FES	Weka	Caret	MLR
Chi-square	S	ST	FW	DM	✓	✓	✓				
CIFE	S	IT	FW	DM		✓			✓		
CMIM	S	IT	FW	DM		✓					
DISR	S	IT	FW	DM		✓					
FCBF	S	IT	FS	DM		✓					
Fisher Score	S	SI	FW	DM		✓					
F-score	S	ST	FW	DM		✓					
Gini Index	S	ST	FW	DM	✓	✓					
ICAP	S	IT	FW	DM		✓					
JMI	S	IT	FW	DM		✓					
Laplacian Score	S	SI	FW	CD		✓					
Least square loss	S	SL	FW	DM		✓		✓			
Logistic loss	S	SL	FW	DM		✓					
Low variance	U	ST	FS	DB	✓	✓					
MCFS	S	SL	FW	CD		✓					
MIFS	S	IT	FW	DM		✓					✓
MIM	S	IT	FW	DM		✓					✓
MRMR	S	IT	FW	DM	✓	✓					
NDFS	U	SL	FW	CD		✓					
ReliefF	S	SI	FW	DM	✓	✓					
RFS	S	SL	FW	DM							
SPEC	U	SI	FW	CD		✓				✓	
Trace Ratio	S	SI	FW	DM		✓					
T-score	S	ST	FW	DB		✓					✓
UDFS	U	SL	FW	CD		✓			✓		✓

FS Algorithm characterization:
a) Label-S: Supervised; U: Unsupervised
b) Learning-ST: Statistical; IT: Information theory, SI: Similarity, SL: Sparse learning
c) Output-FW: Feature weight; FS: Feature subset
d) Feature Type-DM: Discrete(multi-class); DB: Discrete(binary-class); CD: Continuous and discrete
Box with ✓ means that the algorithm is implemented in that library.

V. CONCLUSION

All current NIDS use characteristics of user and system behavior as input to their analysis algorithms, which determine the likelihood of an attack. There is an urgent need for intelligent NIDS that can automatically detect novel attacks. A NIDS FS phase is difficult, as the NIDS must deal with a massive amount of data. In this paper, we have provided a review of feature selection methods used in NIDS. Feature selection framework and taxonomy for intelligent NIDS along with a comparison of FS Algorithm characterization and Python FS libraries contents are presented.

ACKNOWLEDGMENT

The authors would like to thank other researchers who have contributed to the field of NIDS.

REFERENCES

[1] T. Zaidi, "A Network Intrusion Based Detection System for Cloud Computing Environment," no. April, 2021, doi: 10.20944/preprints202104.0183.v1.
[2] J. Snehi, A. Bhandari, M. Snehi, V. Baggan, and H. Kaur, "AIDAAS: Incident Handling and Remediation Anomaly-based IDaaS for Cloud Service Providers," in 10th International Conference on System Modeling & Advancement in Research Trends, 2021, pp. 356–360, doi: 10.1109/SMART52563.2021.9676296.
[3] S. Maza and M. Touahria, "Feature selection algorithms in intrusion detection system: A survey," KSII Trans. Internet Inf. Syst., vol. 12, no. 10, pp. 5079–5099, 2018, doi: 10.3837/tiis.2018.10.024.
[4] J. Snehi, M. Snehi, V. Baggan, A. Bhandari, and R. Ahuja, "Introspecting Intrusion Detection Systems in Dealing with Security Concerns in Cloud Environment," in 10th International Conference

- on System Modeling & Advancement in Research Trends (SMART), 2021, 2021, pp. 345–349, doi: doi: 10.1109/SMART52563.2021.9676258.
- [5] M. Snehi and A. Bhandari, “Vulnerability retrospection of security solutions for software-defined Cyber–Physical System against DDoS and IoT-DDoS attacks,” *Comput. Sci. Rev.*, vol. 40, p. 100371, May 2021, doi: 10.1016/j.cosrev.2021.100371.
- [6] K. V. Krishna, K. Swathi, and B. B. Rao, “A Novel Framework for NIDS Throuh Fast Knn Classifier on CICIDS 2017 Dataset,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 3669–3675, 2020, doi: 10.35940/ijrte.e6580.018520.
- [7] R. I. Farhan, A. T. Maalood, and N. Hassan, “Hybrid Feature Selection Approach to Improve the Deep Neural Network on New Flow-Based Dataset for NIDS,” *Wasit J. Comput. Math. Sci.*, vol. 1, no. 1, pp. 66–83, 2021, doi: 10.31185/wjcm.voll.iss1.10.
- [8] Amrita and S. Kant, “Machine learning and feature selection approach for anomaly based intrusion detection: A systematic novice approach,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 434–443, 2019.
- [9] Z. Yongli, Z. Yungui, T. Weiming, and C. Hongzhi, “An improved feature selection algorithm based on MAHALANOBIS distance for Network Intrusion Detection,” *Proc. 2013 Int. Conf. Sens. Netw. Secur. Technol. Priv. Commun. Syst. SNS PCS 2013*, pp. 69–73, 2013, doi: 10.1109/SNS-PCS.2013.6553837.
- [10] K. V. Krishna, K. Swathi, and B. B. Rao, “LVFE : A Feature Selection Approach for an Efficient NIDS on Cloud Environment Using Least Variance Feature Elimination,” no. 13396, 2020.
- [11] P. K. Keserwani, M. C. Govil, E. S. Pilli, and P. Govil, “An optimal NIDS for VCN using feature selection and deep learning technique: IDS for VCN,” *Int. J. Digit. Crime Forensics*, vol. 13, no. 6, pp. 1–25, 2021, doi: 10.4018/IJDCF.20211101.0a10.
- [12] V. K. Jabali, M. Rahbari, and A. Kashkouli, “Taxonomy of Feature selection in Intrusion Detection System,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 6, pp. 88–102, 2017.
- [13] G. Farahani, “Feature Selection Based on Cross-Correlation for the Intrusion Detection System,” *Secur. Commun. Networks*, vol. 2020, 2020, doi: 10.1155/2020/8875404.
- [14] B. Atta, “Evolving features-algorithms knowledge map to support NIDS data intelligence and learning loop architecting – a generalised approach to NIDS pattern feature EVOLVING FEATURES-ALGORITHMS KNOWLEDGE MAP TO SUPPORT NIDS DATA INTELLIGENCE AND LEARNING LOOP ARC,” no. April 2008, 2014.
- [15] K.-C. Khor, C.-Y. Ting, and S.-P. Ammuaisuk, “From Feature Selection to Building of Bayesian Classifiers : A Network Intrusion Detection Perspective Kok-Chin Khor , Choo-Yee Ting and Somnuk-Phon Ammuaisuk Faculty of Information Technology ,” *Journal, Am. Sci. Appl. Publ. Sci.*, vol. 6, no. 11, pp. 1948–1959, 2009.
- [16] H. M. Anwer, M. Farouk, and A. Abdel-Hamid, “A framework for efficient network anomaly intrusion detection with features selection,” *2018 9th Int. Conf. Inf. Commun. Syst. ICICS 2018*, vol. 2018-Janua, pp. 157–162, 2018, doi: 10.1109/IACS.2018.8355459.
- [17] M. A. Umar, C. Zhanfang, and Y. Liu, “Network Intrusion Detection Using Wrapper-based Decision Tree for Feature Selection,” pp. 1–8, 2020, [Online]. Available: <http://arxiv.org/abs/2008.07405>.
- [18] C. Khammassi and S. Krichen, “A GA-LR wrapper approach for feature selection in network intrusion detection,” *Comput. Secur.*, vol. 70, pp. 255–277, 2017, doi: 10.1016/j.cose.2017.06.005.
- [19] J. Snehi, A. Bhandari, M. Snehi, U. Tandon, and V. Baggan, “Global Intrusion Detection Environments and Platform for Anomaly-Based Intrusion Detection Systems,” pp. 817–831, 2021, doi: https://doi.org/10.1007/978-981-16-0733-2_58.
- [20] J. Li, K. Cheng, S. Wang, and F. Morstatter, “Feature Selection : A Data Perspective,” vol. 50, no. 6, 2017.
- [21] J. Miao and L. Niu, “A Survey on Feature Selection,” *Procedia Comput. Sci.*, vol. 91, no. Itqm, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111.
- [22] X. Ni, D. He, S. Chan, and F. Ahmad, “Network anomaly detection using unsupervised feature selection and density peak clustering,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9696, pp. 212–227, doi: 10.1007/978-3-319-39555-5_12.
- [23] V. B. N. Sánchez-marroño, “A review of feature selection methods on synthetic data,” pp. 483–519, 2013, doi: 10.1007/s10115-012-0487-8.
- [24] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.
- [25] J. Li, H. Zhang, and Z. Wei, “The Weighted Word2vec Paragraph Vectors for Anomaly Detection over HTTP Traffic,” *IEEE Access*, vol. 8, pp. 141787–141798, 2020, doi: 10.1109/ACCESS.2020.3013849.
- [26] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, “Deep Learning Approach for Intelligent Intrusion Detection System,” *IEEE Access*, vol. 7, no. c, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
- [27] S. Ickin, M. Fiedler, and K. Vandikas, “QoE Modeling on Split Features with Distributed Deep Learning,” *Network*, vol. 1, no. 2, pp. 165–190, 2021, doi: 10.3390/network1020011.
- [28] M. A. Jabbar, K. Srinivas, and S. Sai Satyanarayana Reddy, “A novel intelligent ensemble classifier for network intrusion detection system,” in *Advances in Intelligent Systems and Computing*, 2018, vol. 614, pp. 490–497, doi: 10.1007/978-3-319-60618-7_48.
- [29] J. Verma, A. Bhandari, and G. Singh, “Review of existing data sets for network intrusion detection system,” vol. 9, no. 6, pp. 3849–3854, 2020.
- [30] M. S. Abirami, U. Yash, and S. Singh, “Building an Ensemble Learning Based Algorithm for Improving Intrusion Detection System,” in *Advances in Intelligent Systems and Computing*, 2020, vol. 1056, pp. 635–649, doi: 10.1007/978-981-15-0199-9_55.
- [31] S. Dash and B. Patra, “Feature Selection Algorithms for Classification and Clustering in Bioinformatics,” no. April 2018, 2017, doi: 10.4018/978-1-5225-1759-7.ch085.
- [32] J. L. Epiphany, “A sian R esearch C onsortium Performance Metrics for Feature Selection , Clustering and Classification Algorithms,” no. January, 2016, doi: 10.5958/2249-7315.2016.01179.5.
- [33] M. K. Uçar, “Classification Performance-Based Feature Selection Algorithm for Machine Learning : P-Score,” *IRBM*, vol. 1, pp. 1–11, 2020, doi: 10.1016/j.irbm.2020.01.006.
- [34] N. Pilnenskiy and I. Smetannikov, “Modern Implementations of Feature Selection Algorithms and Their Perspectives,” *Conf. Open Innov. Assoc. Fruct*, pp. 250–256, 2019, doi: 10.23919/FRUCT48121.2019.8981498.
- [35] N. Pilnenskiy and I. Smetannikov, “Feature selection algorithms as one of the python data analytical tools,” *Futur. Internet*, vol. 12, no. 3, 2020, doi: 10.3390/12030054.
- [36] M. Al-qatf, M. Alhabib, and K. Al-sabahi, “Deep Learning Approach Combining Sparse Autoen- coder with SVM for Network Intrusion Detection,” *IEEE Access*, vol. PP, no. c, p. 1, 2018, doi: 10.1109/ACCESS.2018.2869577.
- [37] N. Pilnenskiy and I. Smetannikov, “Feature selection algorithms as one of the python data analytical tools,” *Futur. Internet*, vol. 12, no. 3, pp. 1–14, 2020, doi: 10.3390/12030054.

An Approach to design Keyboard and Mouse assisting device for handicap users

Kamran Hameed*

Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
*khKhawaja@iau.edu.sa

Syed Mehmood Ali

Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
symali@iau.edu.sa

Uzma Ali

Department of Public Health
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
uasali@iau.edu.sa

Abstract— The main idea is to develop an intelligent computing system for disabled people, especially handicapped people who are willing to use a computer for their us-age purpose. Still, due to their disability, they can't use it. Nowadays, engineers try to make some equipment to solve this problem because this century is the computer century. The usage of the computer nowadays is very important. A person needs to use computers for their office purpose, education, and communication. A normal person can do this easily, but a disabled person, especially handicapped people, can't do this. This is a hard situation for a handy cap person because He also needs to communicate with the world. The proposed design can help handicapped peoples use computers for their usage purpose. Intelligent Computing System (ICS) is based on two units one is a head unit which contains a LASER which is used as a remote control, the other unit is ground unit which contains LDR sensor-based keyboard, controlling circuits may assets in performing the whole functionality of the ICS system by providing the instructions through the coding directed by PIC Controller which sending the data to the computer through MAX233 IC. The proposed LDR-based-matrix keyboard system sensor was sensitive enough to detect laser light which was tested on ten different users and shown promising accuracy; for future enhancement in order to use this device as a commercial device, instead of LDR matrix, we can, or the fabrication industry can fabricate LDR matrix during the fabrication process, to minimize the size of the system in order to save the power consumption of the system and may also allow a handicap people to use a computer and communicate with the world and gain knowledge and information just like a normal people.

Keywords—Intelligent computing systems, handy capped assistive device, LDR sensor , matrix keyboard and matrix mouse.

I. INTRODUCTION

This project was motivated by the novella/film "The Diving Bell and the Butterfly" based on the life of Jean-Dominique Bauby. At the age of 42 Bauby, the then French editor of the magazine ELLE, suffered a massive and spontaneous stroke while driving down a road in rural France. After several weeks in a coma, He awoke with what's known as "locked-in syndrome", with full muscular paralysis and with only the control of his left eye remaining. Despite the loss of all His physical ability, His sight, hearing, and especially mental function remained unchanged. He (the butter-fly) was trapped inside His body (the diving bell). With the aid of an

extra-ordinarily patient speech therapist, He gained the ability to communicate again by blinking as she orally read a frequency sorted alphabet to him. In this way, he would slowly form words and then sentences, and over the period of months, with extreme difficulty, he was able to write the account of his experience. This primitive form of communication was excruciatingly tedious to both Bauby and His "interpreter", requiring extreme patience from both individuals as letter entry was quite slow and any mistakes were not easily corrected. Besides this, a lot of work has been carried out to develop keyboard/mouse layouts for non-disable people.

In contrast, a little for physically disabled users, one of the systems designed by the researcher they named BAHDON based on text entry system provided access to the specific user who was unable to use keyboard easily in that system they can interact with the computer using their thumb toe and ankle movement [1], another studied the researcher proposed to design virtual or on-screen keyboards that supported multiple modes of access and had optimum layout based on the frequency of occurrence of the alphabet in English text, the proposed virtual keyboard in this studies designed in visual basic environment which consisted of eight sub circles in a large circle and arrow which were rotated in a clockwise direction each of circles consisted of different keyboard keys [2]. As in previous studies, only researcher focused on keyboard designing layout while this study may also focus on keyboard layout as well as mouse layout to give more access to specific users to use both . The proposed designed is an innovative idea helping the disabled person especially handicap who likes to use pc but due to their disability, they can't use it. In a designed project arrays of LDR are worked on the keyboard principle. The LDR (Light Detecting Resistor) is working on the principle of keyboard containing 26 Alphabets, 0 to 9 Digits, some special keys, scroll keys, enter key, space key, and backspace key. One LDR for mouse mode and one for keyboard mode. Proposed design may use the same principle for mouse usage. The U key for upper movement of mouse, M key for lower movement of the mouse, H key for the mouse's left movement of the mouse, K key for the mouse, P for a single click, O for double click, R for double click, R for right-click. ICS may exceed handy cap person expectations with accurate laser communication and precise sensing. [3-15].

II. DESIGN AND METHODOLOGY

A. Hardware Description

The working principle of the ICS is based on the keypad principle micro-controller, which is PIC16F877A. Another main circuit is the op-amp which uses as a comparator for LDR. The MAX233 is connected to the microcontroller. It amplifies the voltages. The MAX233 IC is connected with an RS232 cable which interfaces the whole project with the computer. We use the same principle for mouse usage, as

shown in Figures 1-2. The U key for upper movement of the mouse, M key for lower movement of the mouse, H key for the mouse's left movement of the mouse, K key for the mouse, P for a single click, O for double click, R for double click, R for right-click. Detection of light using LD. Elimination of daylight which falls on LDR using op-amp comparator. PIC microcontroller receives the data and sends it to the computer through MAX233 IC and RS232 cable. Visual studio.net software receives the serial port data and activates the mouse and keyboard function. A comparator circuit may be designed to respond to continuously varying (analog) or discrete (digital) signals, and its output may be in the form of signaling pulses that occur at the comparison point or in the form of discrete direct-current (dc) levels. [4].

B. Handicap Keyboard and Mouse Pad designed matrix

LDR-based keypad designed, the LDR activated by the LASER, which is mounted on the user's head. 48 arrays of LDR were made on the basic pattern of keyboard keys. These arrays are then connected in 3x16 matrixes consisting of 3 Rows and 16 Columns, as shown in Figure 3. The LDR has 2 pins; one is connected in rows, and the other is connected in columns, as shown in Figure 4. The first pin of LDR has connected in rows; each row relates to the LM741 Operation Amplifier windows comparator as shown in Figure 4, while the LDR signal conditioning circuit is shown in Figure 5. The output of the OpAmp comparator is connected with the PIC microcontroller input port; through programming, we designed non-triggerable one short signal was picked at every 1 second to give LDR matrix sufficient time to respond upon user response to type another letter by moving their head towards the LDR matrix keyboard layout, because some part of LDR matrix is assigned for mouse pad cursor moving the same timing delayed was given for this purpose too. Letter. While the second pin of LDR is connected in columns, and then each column is connected to PIC microcontroller input port pins, as shown in Figures 4 and 5. Now the arrays of LDR are connected to PIC16F877A. The proposed design has three (3) I/O ports used as an input port of the PIC controller. The ports are B, D and E. The columns of LDR are connected at PORTB and D as shown Figure 4-6, PORTB and D both have eight (8 bits); therefore, each pin relates to three (3) LDR's, as shown in Figure 4-5, then the output of the op-amp comparator is connected at PORTE of PIC microcontroller., for this purpose, PORTE bit E0, E1, E2 were used as an input port pins. Now the second pin of LDR is connected in rows, and there are three rows; each row contains 16 LDR arrays, each row relates to LM741 op-amp, which use as a comparator. The

handicapped person may also use and control the mouse with these Keypad LDR matrix pads to give mouse access control to the user; some of the proposed LDR keypad keys were assigned for mouse controlling functions, respectively, as mentioned in Table 1.

TABLE I. LDR BASED MOUSE PAD.

Mouse Keys	Mouse Key Functions
U	For Upward Mouse Movement
M	For Downward Mouse Movement
H	For Left Mouse Movement
K	For Right Mouse Movement
P	For selecting Single Mouse Click
O	For selecting Double Mouse Click
R	For selecting Right Mouse Click

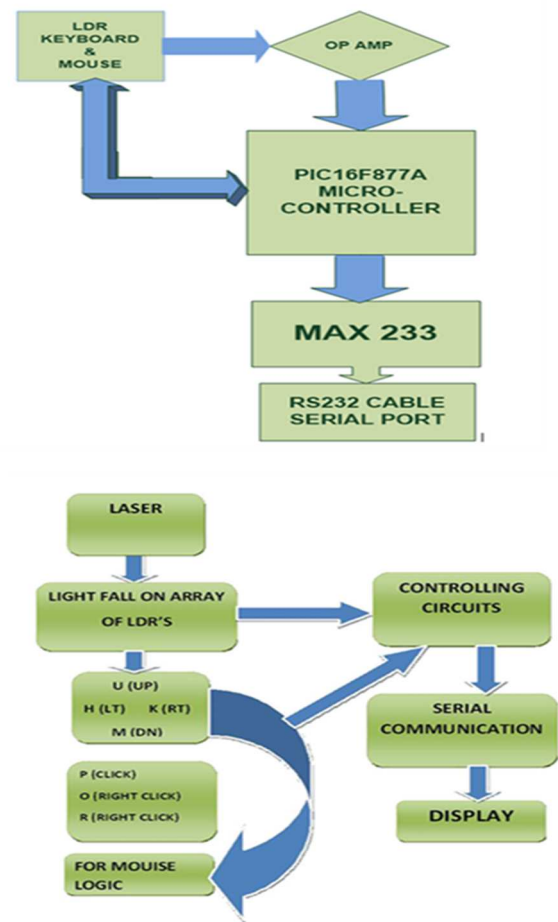


Fig. 1. At Top shows the Hardware components used in proposed LDR-Based Keyboard and Mouse matrix, At Bottom shows the working principle of LDR-based mouse of the proposed system

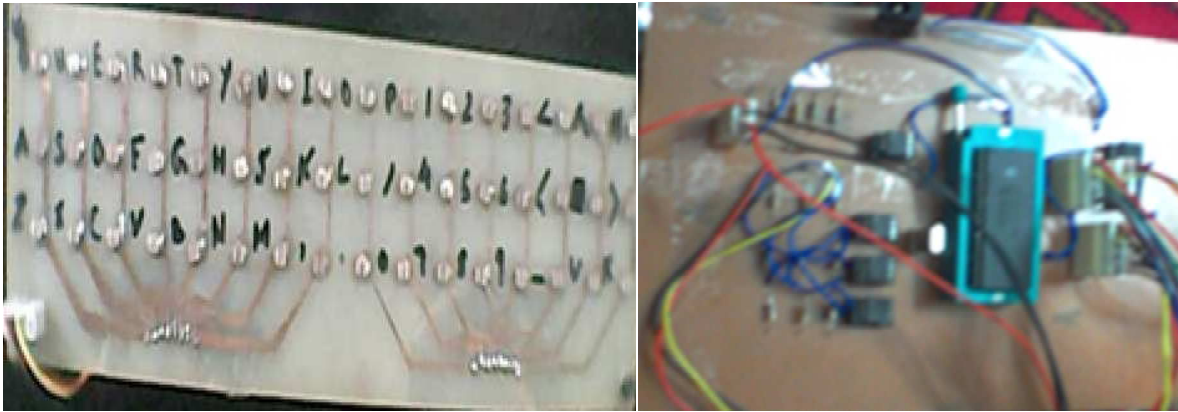


Fig. 2. At Right shows the Proposed LDR-Keyboard/Mouse Matrix. At Left shows the Controlling circuit of the proposed design.

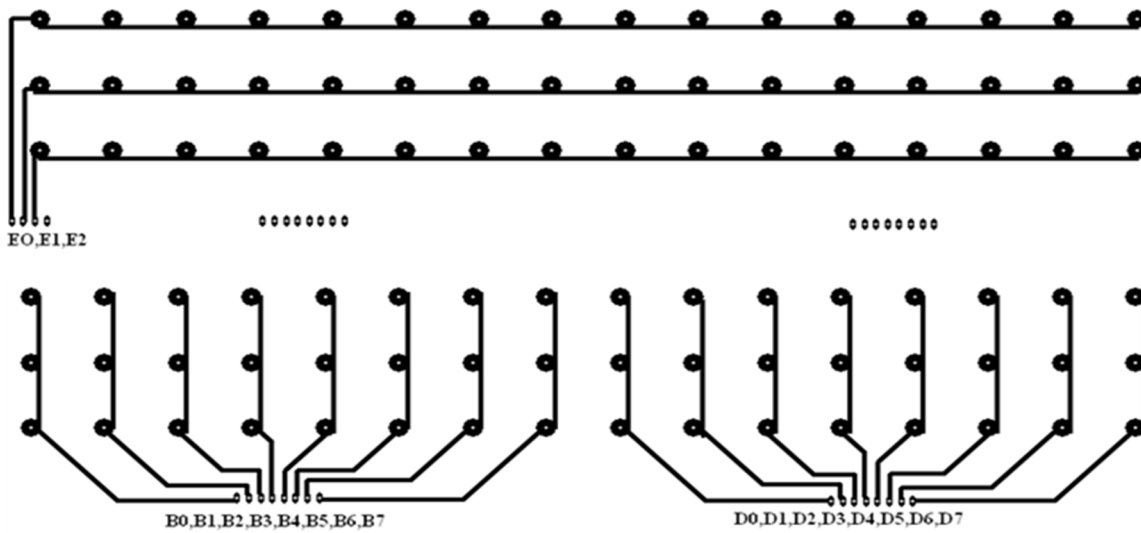


Fig. 3. At Right shows the Proposed LDR-Keyboard/Mouse Matrix. At Left shows the Controlling circuit of the proposed design

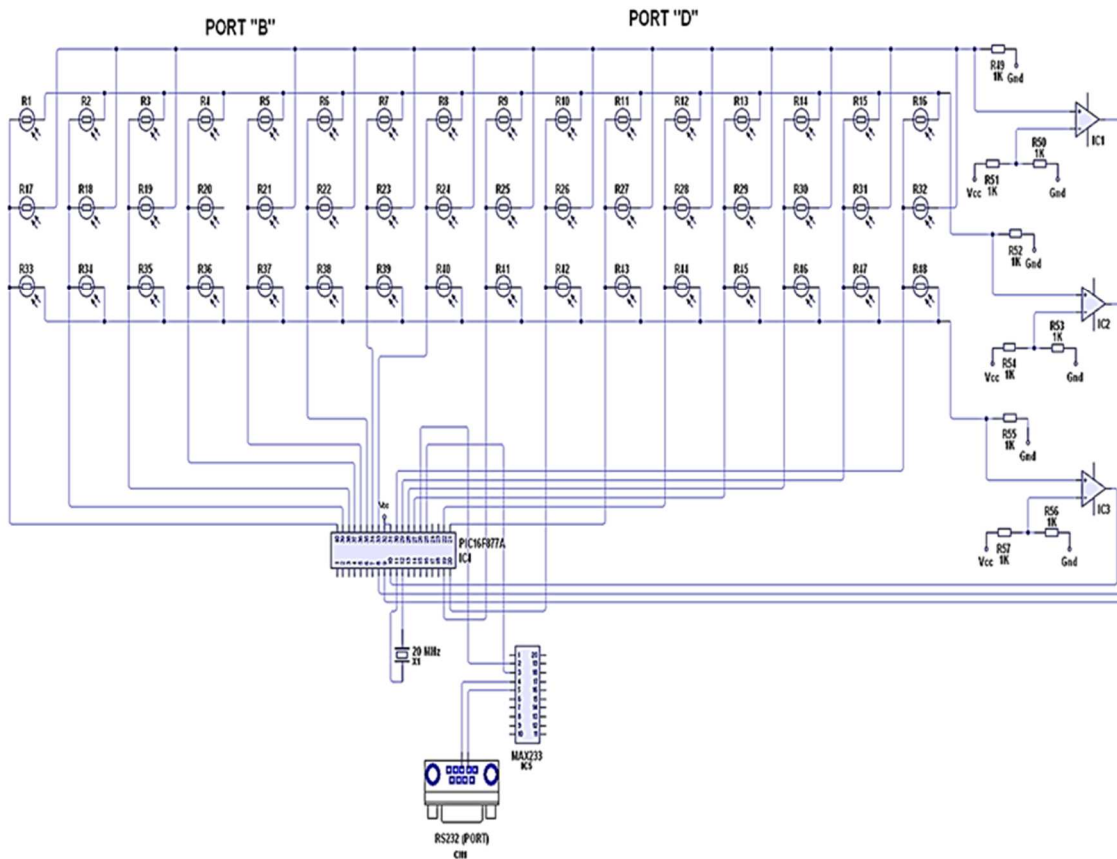


Fig 4. LDR-based Keyboard/mouse LDR connection diagram

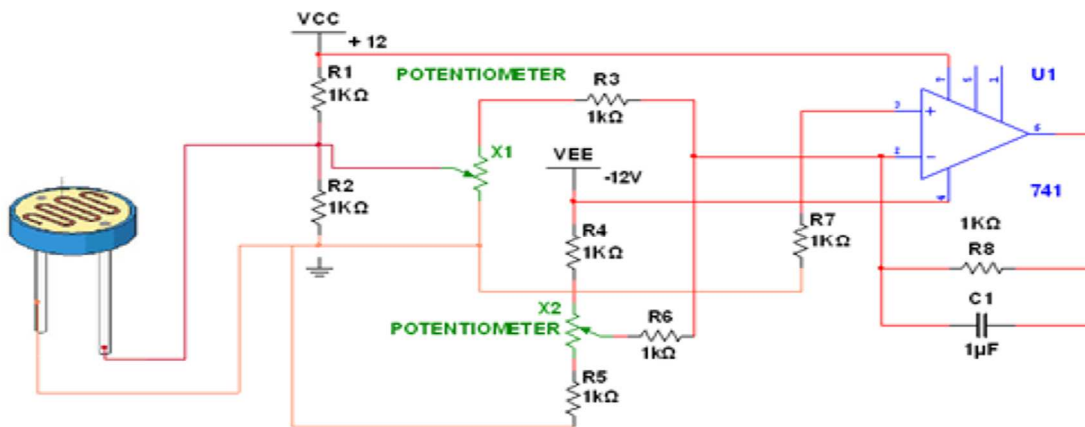


Fig 5. LDR signal conditioning circuit

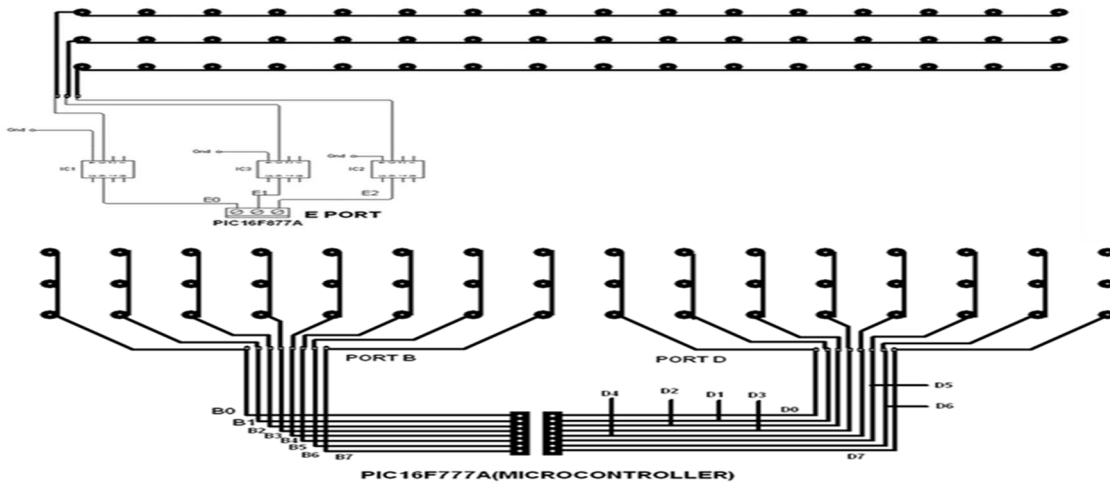


Fig 6. LDR Row/Column wiring diagram

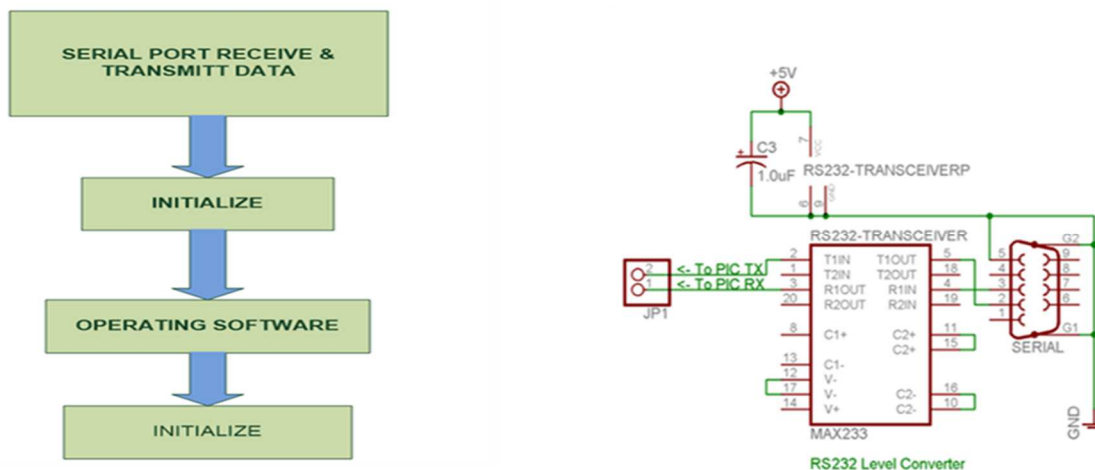


Fig 7. At Right shows the data transmission/receiving flow . At Left shows the connection b/w Pic microcontroller with RS232 connector

C. Data Acquisition

Step by step sending data procedures were needed for sending and receiving LDR data for this purpose Serial port of PIC Microcontroller was also used the designed circuits were operating at 5 V while for interfacing the hardware circuit with user computer MAX233 IC was used, A MAX233 is used to convert the logic level signals of the PIC microcontroller to RS-232 compatible voltage levels. Technically the MAX233 is an RS-232-line driver/receiver, but an easier way to explain its operation is that it allows a PIC microcontroller to

communicate at RS-232 voltage levels with a computer as shown in Figure 7, the MAX233 should nominally be run at 5 Volts.[3] for serial communication to communicate the LDR Matrix circuits with user Pc. To achieve this serial communication within the designed circuitry and pc USART interfacing protocol was adopted. The PIC Microcontroller sends the LDR matrix data through an RS232 cable to the computer. Then the RS232 cable transmits the information to the controller that waits for a while, and the sending data perform the expected task. This basic RS-232 transmit/receive

circuit is necessary for PIC microcontrollers to communicate reliably with a PC serial port.

D. Software programming description

In the coding part IF ELSE condition statement were used to recognized LDR-based keypad detection. When the user moves him/her head bend consisting of Laser LED embedded on these head bend, this laser light will fall towards the character, user willing to type on a computer screen. The LASER activated that particular LDR character where the light falls. The controller checks the whole LDR arrays "if" the LASER activate the "A" letter, LDR then the controller send the "A" letter ASCII code to the computer. If the LASER is not activated the "A" letter LDR, check the other ones. Therefore, IF ELSE condition statements were used in whole programming codes, the same technique was used to recognize all characters.

III. RESULT AND CONCLUSION

The proposed design exceeded our expectations with accurate laser communication and precise sensing. The prototype device was given to 20 users to use it as a keyboard and mouse with their computers, to type the sentence "An Approach to design Keyboard and Mouse assisting device for handicap users" it's showing the promising result with 85 to 90 percent accuracy and the average time taken to type the stated sentences for all these users were between 90 seconds to 2 minutes. It was noticed that the proposed LDR-based-matrix keyboard system sensor was sensitive enough to detect laser light. The distinctive signal provided by the op-amp (window comparator) helped filter out daylight from the laser light used, for future enhancement in order to use this device as a commercial device, instead of LDR matrix, we can, or the fabrication industry can fabricate LDR matrix during the fabrication process, because in this proposed design lot of LDR used and connected in a matrix format may increase the power consumption and may reduce the sensitivity of the system detection. In short, the proposed design will allow disabled person, especially handicapped people, to use computers and communicate with the world and gain knowledge and information just like a normal person. Hope this system may "HELP THE HUMANITY."

- [1] J. Protim et al, "BADHON: A high performing keyboard layout for physically impaired people," in 2019, . DOI: 10.1109/ICSEC47112.2019.8974683.
- [2] V. Prabhu and G. Prasad, "Designing a virtual keyboard with multi-modal access for people with disabilities," in 2011, . DOI: 10.1109/WICT.2011.6141407.
- [3] J. Wilkinson and K. Breneman, "Bridging the Digital and the Physical User Experience: Analysis of Academic Library Floor Plans," *Journal of Web Librarianship*, vol. 14, (1-2), pp. 28-51, 2020.
- [4] G. Rappolt-Schlichtmann, A. R. Boucher and M. Evans, "From deficit remediation to capacity building: Learning to enable rather than disable students with dyslexia," *Language, Speech & Hearing Services in Schools*, vol. 49, (4), pp. 864-874, 2018.
- [5] S. Yang et al, "Design of virtual keyboard using blink control method for the severely disabled," *Computer Methods and Programs in Biomedicine*, vol. 111, (2), pp. 410-418, 2013.
- [6] L. Giancardo et al, "Psychomotor Impairment Detection via Finger Interactions with a Computer Keyboard During Natural Typing," *Scientific Reports*, vol. 5, (1), pp. 9678-9678, 2015.
- [7] F. T. Mohammed and M. T. Mohammad, "Facilitate Access ability and Provide Security for Handicap Person Computers," *International Journal of Computer Science Issues*, vol. 12, (4), pp. 113, 2015.
- [8] I. A. Khan, W. Brinkman and R. Hierons, "Towards estimating computer users ' mood from interaction behaviour with keyboard and mouse," *Frontiers of Computer Science*, vol. 7, (6), pp. 943-954, 2013.
- [9] R. J. P. Damaceno, J. C. Braga and J. P. M. Chalco, "Mobile device accessibility for the visually impaired: Problems mapping and empirical study of touch screen gestures," in 2016, . DOI: 10.1145/3033701.3033703.
- [10] I. Iliiev and I. Dotsinsky, "Assisted living systems for elderly and disabled people: A short review," *Bioautomation*, vol. 15, (2), pp. 131-139, 2011.
- [11] E. M. Holz et al, "User centred design in BCI development," in *Towards Practical Brain-Computer Interfaces* Anonymous Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 155-172.
- [12] C. G. Pinheiro Jr et al, "Alternative communication systems for people with severe motor disabilities: A survey," *Biomedical Engineering Online*, vol. 10, (1), pp. 31-31, 2011.
- [13] T. Felzer and S. Rinderknecht, "Mouse mode of OnScreenDualScribe: Three types of keyboard-driven mouse replacement," in 2013, . DOI: 10.1145/2468356.2468641.
- [14] B. Morales et al, "AsTeRICS: a new flexible solution for people with motor disabilities in upper limbs and its implication for rehabilitation procedures," *Disability and Rehabilitation: Assistive Technology*, vol. 8, (6), pp. 482-495, 2013.
- [15] A. Quezada et al, "Usability Operations on Touch Mobile Devices for Users with Autism," *Journal of Medical Systems*, vol. 41, (11), pp. 1-11, 2017.

Design of Monitoring System for Respiratory Diagnosis

Kamran Hameed*

Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
*khKhawaja@iau.edu.sa

Sana Ijlal Shahrukh

Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
Siateeq@iau.edu.sa

Ijlal Shahrukh Ateeq

Department of Biomedical Engineering
Imam Abdulrahman Bin Faisal
University
Dammam, Kingdom of Saudi Arabia
Lsateeq@iau.edu.sa

Abstract—Respiration is an important vital sign which can help in the indication of human's health status. Sleep apnea, which is defined as a cessation of breathing while sleeping, is recognized as a common and critical popular health problem. By monitoring the patient's respiration rate, sleep apnea can be diagnosed. The received signal from the patient, which is a variation in temperature, is converted into a variation in voltage and then the respiration rate can be calculated. A portable device is proposed that can continuously monitor the breathing rhythm in a manner which is fast, safe and costly effective. The system consists of a nasal-oral thermistor that is mounted inside a nebulizer mask for detecting the irregular respiratory interruption and providing audible and visual alarms when this interrupt indicates an apnea event. Otherwise, if an out-of-range value occurred the alarm also will be initiated. Thus, the device will contribute to reducing the death associated with sleep apnea.

Keywords—Respiratory monitoring system, Sleep apnea, Respiratory rate, Nebulizer mask, Nasal oral thermistor.

I. INTRODUCTION

Respiration is considered as an essential vital sign which gives a clue of the health condition of human. Several health conditions and diseases may affect the respiration rate. Sleep apnea is known as an interruption of breathing during sleep. Sleep apnea is considered as a serious and common disease. Undiagnosed sleep apnea can significantly affect the quality of a person's life. Prolonged apnea reduces oxygen levels in blood and tissue, leading to permanent brain damage and, if not remediated, death. A highly sensitive, accurate, reliable and portable monitoring system is needed for apnea detection. The monitoring system is based on calculating the respiration rate which is the number of breathing cycles per one minute. A portable device is proposed that can continuously monitor the breathing rhythm in a manner which is fast, safe and effective [1][2]. Since this respiratory monitoring device is used by sleep apnea patients at home, it is highly secure and easy to use. Hence, the system is noninvasive [3]

The primary aim of this project was to design a portable respiratory monitoring system that provides a reliable, inexpensive, noninvasive, and faster approach to continuously monitor the respiration rate of the sleep apnea patients for this an airflow-based method was designed in monitoring respiration rate by using a nasal-oral thermistor for detecting and analyzing the respiration signal and

extracting the respiration rate an algorithm has also been developed [3-15].

II. DESIGN AND METHODOLOGY

A. Hardware Description

The nasal-oral thermistor which is mounted in a nebulizer mask is used to detect apneic events by measuring respiration rate. The proposed device presents a noninvasive respiration monitoring system that is appropriate and secure for home utilizing on adults. It proposes an approach to monitor sleep apnea patients prior to, through and after treatment. The system is composed of a nasal-oral thermistor. It detects the breathing rhythm of the patient. A thermistor is placed snugly below the nose and ahead of the mouth. The cool inhaled and warm exhaled air temperatures are sensed by the thermistor. The variance in temperature provides an indication of the continuity of breathing [2]. A certain circuit is designed to receive the sensor's data, then the circuit output is transmitted to the processing unit (Arduino) placed in the same area as the patient. The signal is processed through computational series, keeping track of the occurrence of an interruption in breathing. In healthy adults, the normal range of respiratory rate is between 12 and 18 bpm at rest conditions [4]. That means a breath each 3 to 5 seconds [2]. The normal breathing frequency is falling in the range of 0.2-0.3 Hz [5]. This method is able to detect all types of apnea [6]. The thermistor is considered the most ordinary method for detecting respiratory events [7]. The nasal-oral thermistor should have a low thermal mass and an adequate sensitivity [8]. If the processor detects a signal below the normal range, this will be considered as an apneic event. After that, the action is taken by the processing unit depending on whether the patient's breath is normal, or apnea is detected. In the case of apnea, the alarm will be activated. If the patient starts again 30 breathing unaided, the alarm ends. In case that the patient still does not carry on breathing, the alarm signal remains to warn the caretaker that instant awareness is mandatory. The main components of the system are shown in Figure 1. The respiratory monitoring device requires a highly sensitive thermistor in order to detect a small variation in the temperature. The thermistor is a type of sensors with a variable resistance that changes when the temperature changes, then this change in resistance is translated into voltage. The negative temperature coefficient

(NTC) thermistor was used in this proposed device, since most of the medical applications are used it. The main characteristic of NTC is that the temperature and the resistance of the thermistor are inversely proportional. As the temperature increases, the thermistor's resistance decreases. Equation (1) illustrates the relationship between the thermistor's resistance and temperature.

$$R_{th}(T) = ae^{-bT} + C \quad (1)$$

Where a,b, and c = constants T = the air temperature, °C. It is noticeable that the temperature and resistance have a non-linear relationship [9]. Biomedical applications use a thermistor with a sensitivity in the range of 0.1 to 100 Ω.m [10]. By increasing the thermistor's core temperature, the resistance will exhibit a large decrease but not in a linear relation. To predict this relation accurately, the thermistor calibration stage must be considered. The thermistor circuit is shown in Figure 2. As the patient is breathing, the warm exhaled air changes the R_{th} of the thermistor, therefore, resulting in a proportional change in the output voltage V_{out} . The main idea behind the design circuit in Figure 3 is to detect the variation in the temperature by the 100k NTC thermistor through the designing of the three stages. These stages are: thermistor circuit stage, amplification stage, and low pass filter stage. Furthermore, the output of the circuit, which is taken from the op-amp's pin number 7, is connected to the pin number 8 in the Arduino UNO. Here in Figure 4, the sensor symbol represents the circuit output. After the processing for the input of the Arduino UNO, the output is the respiration rate. Depending on the value of the respiration rate the action will be taken. If it is normal it will be displayed on the LCD. In contrast, if the respiration rate is below the normal, which means below 12, the respiration rate and "Sleep Apnea" text will be displayed on the LCD. At the same time, the audible and visual alarms will be initiated to warn the patient. Moreover, the patient can stop the alarms by pressing the stop button. In case of the respiration rate was not logical, which means above 25, the text "System Error" will be displayed and also the alarm will be initiated.

The thermistor circuit captures the variation in temperature from the subject's air flow and transmits it as a voltage to the second circuit stage. As the temperature increases the thermistor's voltage decreases.

The operational amplifier circuit: It amplifies the thermistor's voltage by a gain value of 3.167. This value is found by dividing the output voltage measured by the amplifier (3.686 V) by the output of the thermistor circuit stage (1.164 V). This stage is shown in Figure 3. The output of the amplifier is transmitted to the third stage which is a low pass filter. In Low-pass filter stage, an active low-pass filter with unity feedback was designed as shown in Figure 3. The cutoff frequency of the filter is calculated based on the values of the capacitors and resistors. The calculations for the cutoff frequency are done by using: $R_1 = R_2$, $C_1 = 2 C_2$.

$$f = \frac{\sqrt{2}}{4\pi RC_2} \quad (2)$$

The cutoff frequency of the active low pass filter is equal to $f=0.938$ Hz. The purpose of this stage is to suppress any frequencies above the cutoff frequency. The suppressed values considered as noise and interruption for the respiration signal.

B. Algorithm Flowchart.

The designed circuit output digital signal is acting as an input for the processing unit. The breathing timer will be initiated with the first inhalation and then it will measure the time for every respiration cycle. This time will be used to calculate the respiration rate in breaths per minute unit. The respiration rate value will be displayed on LCD. As an action, if the respiration rate is less than the minimum of the normal range which is 12 breaths per minute, an audible and visual alarm are initiated indicating an apnea event to warn the patient and wake him/her up [11]. A 'System Error' text will be displayed on LCD in the case of respiration rate value that is not logical (above 25), and also here the alarm will be initiated. Finally, the alarm will continue until someone press the button to stop it.

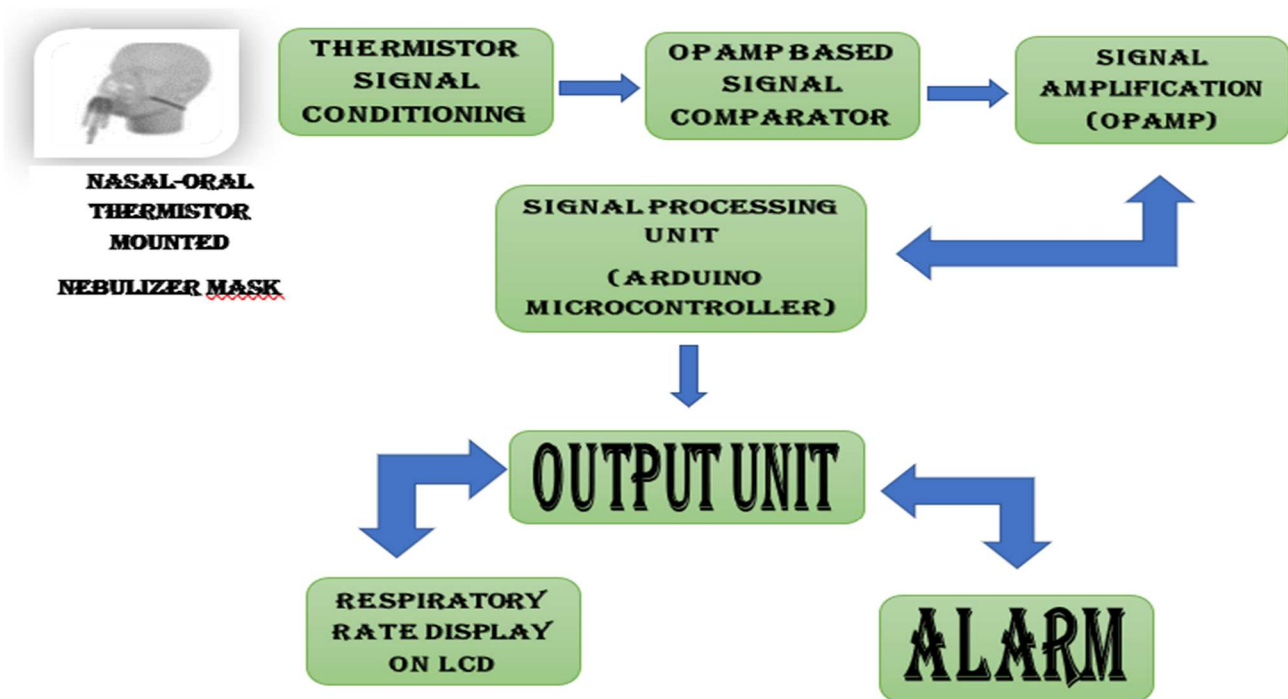


Fig. 1. Main components of the proposed respiration monitoring system

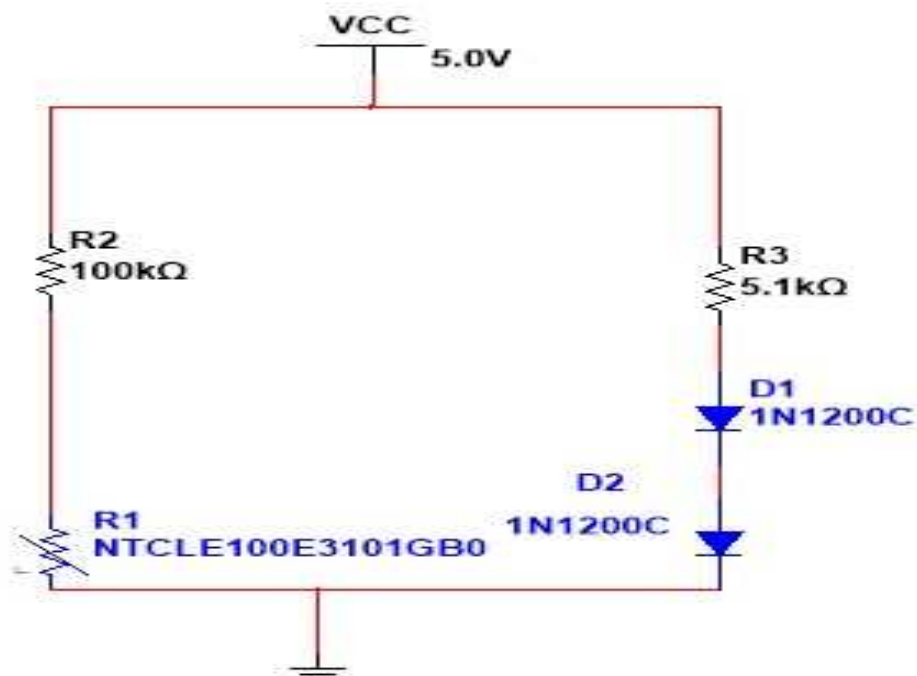


Fig. 2. Thermistor Circuit

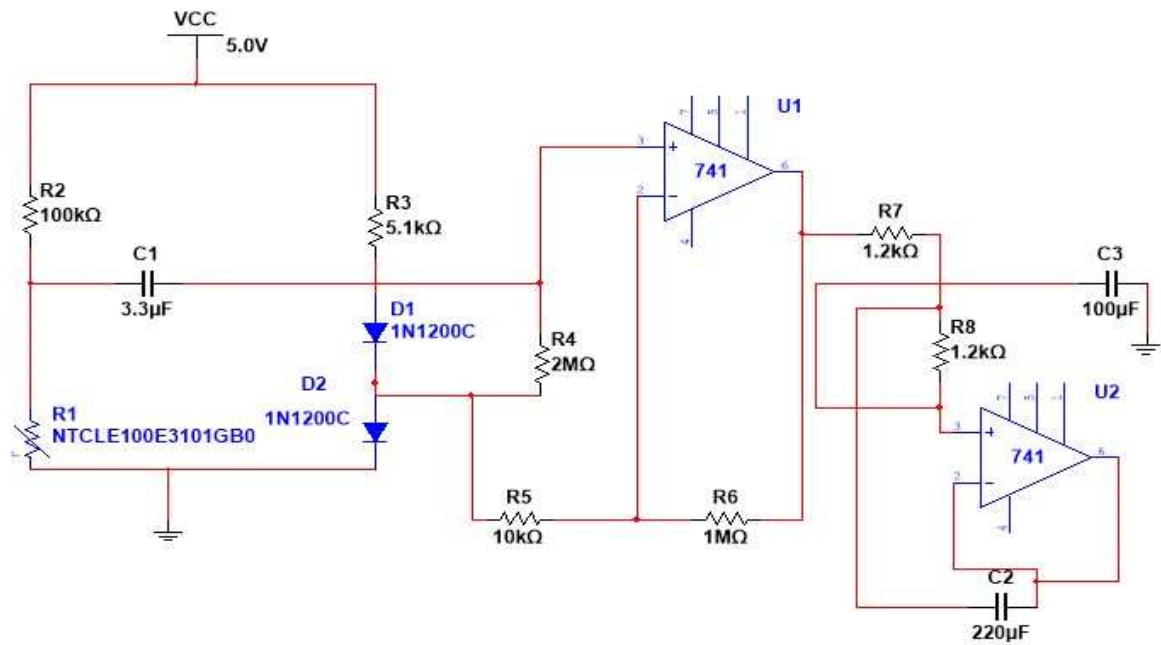


Fig. 3. The circuit design of the system

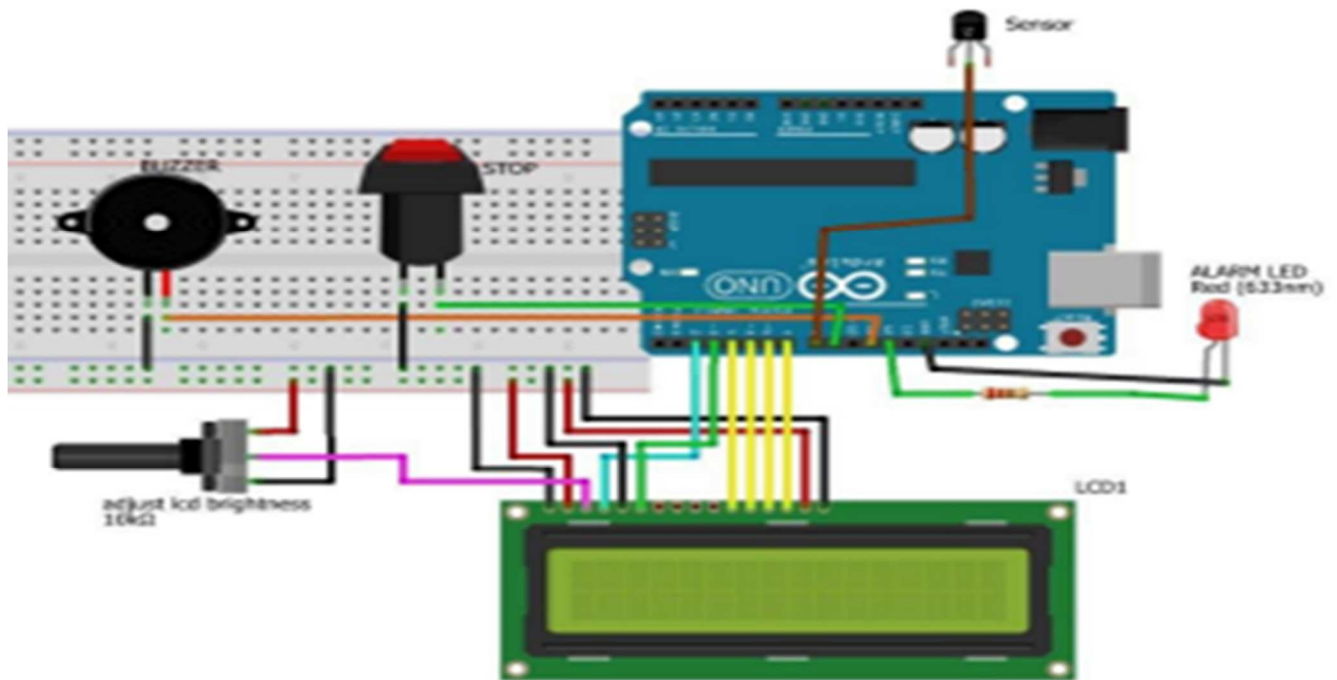


Fig. 4. Processing unit.



Fig. 5. Fig 5.The algorithm flowchart of the proposed monitoring system

C. Hardware / Experimental Setup

The hardware of the proposed device is composed of the circuit design, Arduino UNO, visual alarm, audible alarm and LCD which are shown in Figure 6. The Arduino UNO was powered using a 9 V adapter in order to operate the uploaded program in the Arduino UNO electrically erasable programmable read-only memory (EEPROM). The respiration rate was displayed on the LCD as shown in Figure 7 and Figure 8. Both figures are examples of normal and abnormal (apnea) respiration rates. Also the “system Error” text is displayed as shown in Figure 9: LCD display of “System Error” when the not logical value appears.

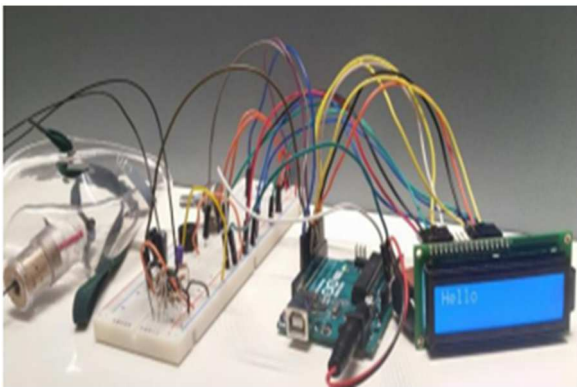


Fig. 6. Project Hardware/Prototype



Fig. 7. LCD display of normal Respiration rate



Fig. 8. LCD display of abnormal Respiration rate (apneic event)



Fig. 9. LCD display of “System Error”

III. RESULT AND DISCUSSION

The proposed design shows satisfying results compared with the previous sleep lab database results. As shown in Figure 10, by counting the numbers of peaks in the circuit’s output signal, the breathing pattern consists of 12 peaks (breaths) per minute which falls in the normal respiration range (12-18 Bpm). One peak represents a one respiration cycle (respiration cycle = inhalation + exhalation). Refer to characteristic of NTC it shows that the thermistor’s voltage decreases as the temperature increases. All of the respiration signals that was taken in this stage were from the pin number 7 of the op-amp which represents the output of the whole designed circuit.

However, Figure 11 shows an apneic event. It shows that the regular respiration is described by the existence of a specific pattern and the existence of an amount of energy in the signal. Apnea is simply detected by the lack of rhythm in addition to a lack of energy [8]. As seen, the subject’s breath stopped for around 10 seconds which represents the sleep apnea breathing pattern.

When doing the trial using Arduino UNO serial monitor, Figure 12 demonstrates the subject’s results. For the first respiration rate = 2.02 Bpm, "Sleep Apnea" warning message was displayed. For the second respiration rate = 17.44 Bpm, no message was displayed because the result falls in the

normal respiration range. The Third respiration rate = 10.32 Bpm, the warning message was displayed. We take Respiration Rate Trials of 3 subjects and plot Time vs. Voltage by using MATLAB.

The three previous graphs show a different breathing rate patterns from a different three subjects. The first, second and third breathing rate = 13.50 Bpm, 13 Bpm and 15.4 Bpm respectively. So, the proposed respiratory monitoring system is capable to measure and display a reliable respiration rate values for different subjects.

A prototype device for respiratory monitoring system for sleep apnea patients has been developed, implemented, and tested. This device is suitable for monitoring of respiration and detection of sleep apnea. The proposed device may be used with patients sleep apnea in their homes, as an alternative to medical supervision in healthcare institutions, with a detection degree of accuracy similar to the commercially available devices.

IV. CONCLUSION

The main advantage of proposed device is this is portable, inexpensive, noninvasive, and faster approach to continuously monitor the respiration rate of the sleep apnea patients. A nasal-oral thermistor is used to monitoring respiration rate. It also has audible and visual alarms to alert the patient or its care giver. Normally for diagnosis of sleep apnea Polysomnography (PSG) device is used. A Polysomnography device converts body electrical signals to graphical representation . While doing this assessment, the patient has to stay all night at a specifically prepared sleep laboratory supervised by a sleep technician. A full set of sensors is used in PSG to monitor physical functions in sleep. They are attached to the patient’s body by wires that join into a central box. A computer system is associated with the central box to record, store and display data. Despite that PSG is considered the gold standard for sleep diagnosis, there are several disadvantages associated with this test It does not provide the exact home sleep environment, and it causes discomfort to the patient by all the wires attached to it. The results acquired from the PSG may be distorted and not accurate. PSG needs considerable complex equipment to record data and extensive expert clinicians to determine the severity of sleep disorders. In general, most of the devices and procedures are extremely time-consuming and costly process since it costs several thousand dollars. This is a big problem to the patients and makes them carry on having undiagnosed sleep apnea. 82% of men and 93% of women with sleep apnea are not diagnosed As a result of these disadvantages, the patient may desire to have a home diagnosis as an alternative Portable monitoring is a satisfactory option in patients supposed to have sleep apnea. Portability-based approaches are the primary component for the new health technology to make it comprehensive, cost-effective and a beneficial solution to the sleep apnea patients. Our proposed device is portable and wireless. Adoption of new and innovative sleep health monitoring ideas or

applications can increase the level of promoting better health for the society. In addition, it increases the availability of wide and fast techniques of how the physiological signal can be picked up from the patient, analyzed and displayed. All these in a carry-out small box beside the patient's bed can offer a comfortable, easy to use and possibility of home sleep monitoring to sleep disorder patients [16-23].

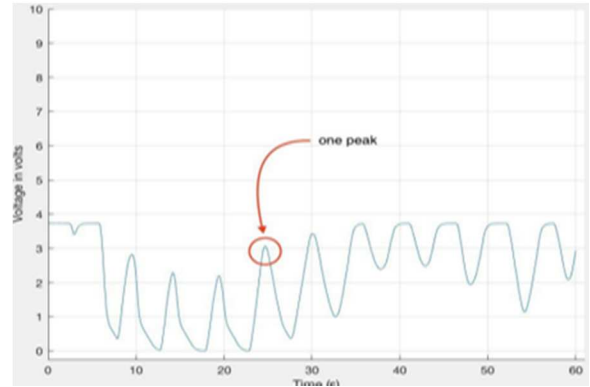


Fig. 10. Voltage vs. Time plotting for normal breathing rate by using MATLAB.

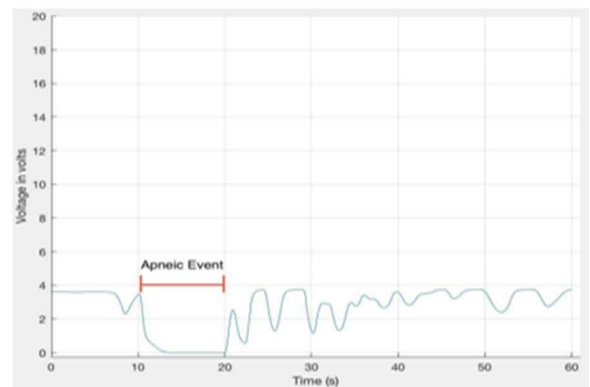


Fig. 11. Time vs. Voltage MATLAB plotting for an apnea that is well determined via the thermistor.

```

Timech started.
Timer sketch started.
startTime: 99
elapsedTime: 29759
1 cycle inhalation + exhalation = 29.76 Sec
respiration rate:2.02 bpm
<<<< SLEEP APNEA >>>>
startTime: 35025
elapsedTime: 3441
1 cycle inhalation + exhalation = 3.44 Sec
respiration rate:17.44 bpm
startTime: 38611
elapsedTime: 5815
1 cycle inhalation + exhalation = 5.82 Sec
respiration rate:10.32 bpm
<<<< SLEEP APNEA >>>>
    
```

Fig. 12. Arduino UNO serial monitor

REFERENCES

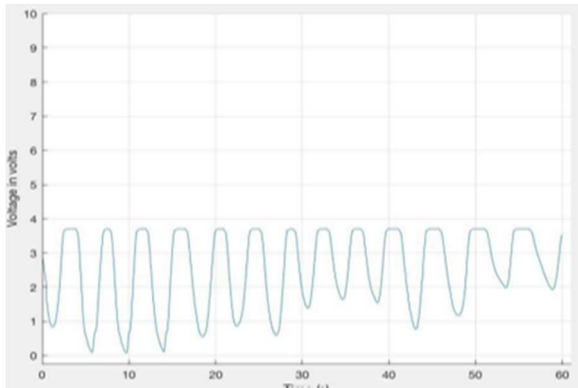


Fig. 13. Time vs. Voltage plotting for the first subject by using MATLAB

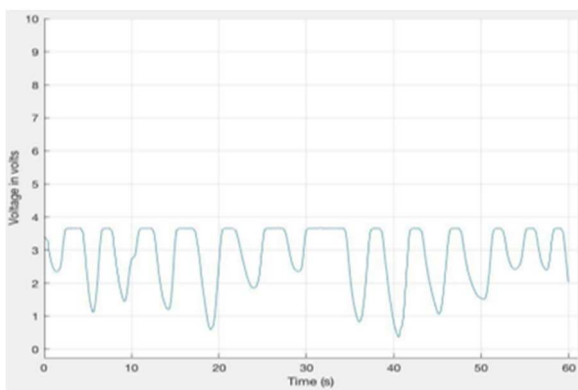


Fig. 14. Time vs. Voltage plotting for the second subject by using MATLAB

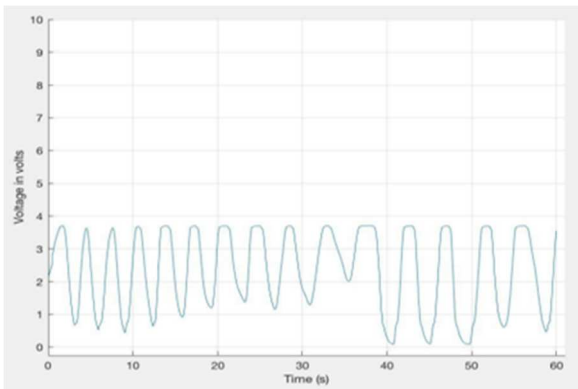


Fig. 15. Time vs. Voltage plotting for the third subject by using MATLAB

[1] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time Sleep Apnea Monitor," *Ieee Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 416–427, 2011.

[2] R. M. Dondelinger, "Apnea Monitors," 2011.

[3] D. J. Mlynek, R. U. S. A. Data, P. Durman, M. Sensors, P. Examiner, and J. A. Hofsass, "United States Patent (19) 11 Patent Number.," no. 19, 2000.

[4] J. D. Gugliotta, "COMBINATION BREATHING MONITOR ALARMAND AUDIO BABY ALARM," 2002.

[5] K. Nepal, E. Biegeleisen, and T. Ning, "Apnea detection and respiration rate estimation through parametric modelling," *Proc. IEEE Annu. Northeast Bioeng. Conf. NEBEC*, vol. 2002–Janua, no. 4, pp. 277–278, 2002.

[6] G. Daan and V. Pattinasarany, "Apnea recognition using neural networks," 2003.

[7] A. C. M. Cross-talk, N. Farré, R. Farré, and D. Gozal, "Sleep Apnea Morbidity," *Chest*, pp. 1–6, 2018.

[8] "DETECTION OF RESPIRATORY INFORMATION USING ELECTROMAGNETIC BIOSENSORS.pdf,"

[9] C. Medical, "Respiration Laboratory 2006," pp. 0–10, 2006.

[10] J. G and Webster, *Medical Instrumentation Application and Design*.

[11] L. Grote et al, "Finger plethysmography—a method for monitoring finger blood flow during sleep disordered breathing," *Respiratory Physiology & Neurobiology*, vol. 136, (2), pp. 141–152, 2003.

[12] D. HOLDITCH-DAVIS, L. J. EDWARDS and M. C. WIGGER, "Pathologic Apnea and Brief Respiratory Pauses in Preterm Infants: Relation to Sleep State," *Nursing Research (New York)*, vol. 43, (5), pp. 293–300, 1994.

[13] F. Q. AL-Khalidi et al, "Respiration rate monitoring methods: A review: Respiration Rate Monitoring Methods," *Pediatric Pulmonology*, vol. 46, (6), pp. 523–529, 2011.

[14] C. Massaroni et al, "Contact-based methods for measuring respiratory rate," *Sensors (Basel, Switzerland)*, vol. 19, (4), pp. 908, 2019.

[15] B. G. Vainer, "A Novel High-Resolution Method for the Respiration Rate and Breathing Waveforms Remote Monitoring," *Annals of Biomedical Engineering*, vol. 46, (7), pp. 960–971, 2018.

[16] I. J. Brekke et al, "The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review," *PloS One*, vol. 14, (1), pp. e0210875–e0210875, 2019.

[17] D. J. Meredith et al, "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology," *Journal of Medical Engineering & Technology*, vol. 36, (1), pp. 1–7, 2012.

[18] P. Marchionni et al, "An optical measurement method for the simultaneous assessment of respiration and heart rates in preterm infants," *Review of Scientific Instruments*, vol. 84, (12), pp. 121705, 2013.

[19] U. Kyriacos, J. Jelsma and S. Jordan, "Record review to explore the adequacy of post-operative vital signs monitoring using a local modified early warning score (mews) chart to evaluate outcomes," *PloS One*, vol. 9, (1), pp. e87320–e87320, 2014.

[20] E. Helfenbein MS *et al*, "Development of three methods for extracting respiration from the surface ECG: A review," *Journal of Electrocardiology*, vol. 47, (6), pp. 819–825, 2014.

[21] L. Maurya et al, "Non-contact breathing rate monitoring in newborns: A review," *Computers in Biology and Medicine*, vol. 132, pp. 104321–104321, 2021.

[22] P. J. Peyton, M. Wallin and M. Hallböck, "New generation continuous cardiac output monitoring from carbon dioxide elimination," *BMC Anesthesiology*, vol. 19, (1), pp. 28–28, 2019.

[23] R. De Fazio et al, "An overview of wearable piezoresistive and inertial sensors for respiration rate monitoring," *Electronics (Basel)*, vol. 10, (17), pp. 2178, 2021.

ENHANCED DV-HOP NODE LOCALIZATION ALGORITHM BASED ON NEAREST NEIGHBOUR DISTANCE AND HOP-COUNT EVALUATION IN WSNs

1st Kanika Sood

Dept. of ECE

NITTTR

Chandigarh, India

kanikasood27.ks@gmail.com

2nd Kanika Sharma

Dept. of ECE

NITTTR

Chandigarh, India

kanikasharma80@yahoo.com

3rd Amod Kumar

Dept. of ECE

NITTTR

Chandigarh, India

csioamod@yahoo.com

Abstract—The simplicity and ease of implementing the Distance Vector Hop (DV-Hop) localization method is mostly used as a range-free localization technique in location-based services. But it has poor positioning accuracy, particularly in a complicated, unequally distributed structure. To overcome this problem, an upgraded DV-Hop localization method based on BN (Beacon Node), hop thresholds, and balanced matrices are suggested, known as the enhanced-DVHLA method. The enhanced-DVHLA approach proposes the improved contribution in the DV-Hop localization by using ALO optimization algorithm. The proposed method is more efficient and reliable as compared to other approaches in WSNs as it has optimized the error rate (LE) to more than 77 percent, and improved the localization accuracy compared with other techniques. The LER dropped by 40.1 percent, and 27.8 percent, respectively, when compared with the TWDV-Hop and TWDV-Hop-AODV methods.

Index Terms—Node Localization, Wireless Sensor Network, enhanced-DVHLA method, Traditional Distance Vector Hop Algorithm, Node Distribution

I. INTRODUCTION

Wireless Sensor System is an essential innovation for the 21st generation of scientists and academicians. Recent advancements in Sensing Applications (Micro Electro Mechanical Systems) and Communication Technologies have enabled the development of low-cost, low-energy, small, and sophisticated sensing devices that can be deployed across a wide area and connected via wireless channels and the Internet for a variety of commercial and domestic uses. To enable sensors, data processing, integrated computation, and communication in a WSN, sensors and actuators are installed in dense populations and often in huge quantities. Recent technological and computer advancements have resulted in many sensor improvements, computing infrastructure, and wireless transmission. These sophisticated sensors have the potential to bridge the gap between the digital and physical worlds. Sensing devices are used in various instruments, businesses,

equipment, and the environment to aid in the prevention of infrastructure failures, disasters, natural resource preservation, increased output, and protection, among several other aspects.

These intelligent sensors can serve as a link between both the digital and physical world. Sensing devices are employed in various instruments, enterprises, equipment, and the surroundings to assist infrastructure breakdowns, disasters, and preservation of natural resources, animal conservation, increased production, and protection, among other things. In MEMS, VLSI, and wireless networking, advancements have facilitated the use of a wireless sensing system or network. With the development of technologically integrated circuit chips, more efficient integrated circuits are significantly smaller and thinner than previous-generation devices. The evolution of computer processing and sensor systems has led to compact, low-power detectors, regulators, and controllers. Fig. 1 depicts WSN's basic architecture.

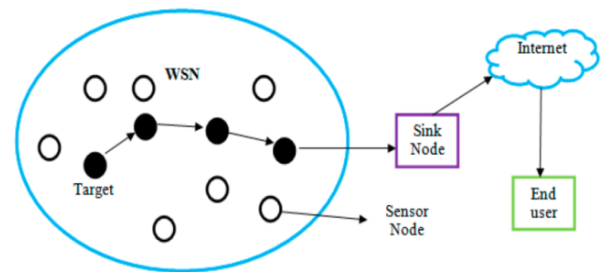


Fig. 1: General Architecture: WSN [2]

The various applications of WSN are explained below: (i) Area Monitoring: Sensor nodes are installed in areas where certain activities must be observed; for example, sensor nodes detect the opponent's location and provide the information to a central for more evaluation. Sensor networks are often used to

track the location of vehicles [3]. (ii) Environment monitoring: WSNs offer a wide range of uses in oceans, forests, and other environments. These systems are used to fire alarms in the woods. WSNs can identify the beginnings of a wildfire and how much it spreads. Animals' movements are also detected by sensor nodes, allowing researchers to study their routines. WSNs are also deployed to monitor the position of plants growing. (iii) Industrial automation: A wireless sensor network is used in various smart applications. For efficient results, it is employed in industrial automation. The wireless sensor network can be used for various applications in industrial automation, including product flow monitoring, protection, and safety, quality control, and factory digitalization. (iv) Smart agriculture system: Traditional irrigation methods increase crop yields by controlling biological impacts. On the other hand, manual farm handling diminishes productivity, consumes a lot of energy, and raises labor costs, making the process less efficient. A smart wireless sensor-based greenhouse with smart sensors is easy to handle, but it also allows us to alter its temperature [4].

WSN network has several advantages; since the WSN (Wireless sensor network) is expandable, new nodes or sensors can be added at any moment. WSN networks are dynamic. Due to battery exhaustion and other issues, adding or transferring some new sensor nodes into the network can meet the task requirements. Because the topology may change, dynamic, reconfiguration and self-adjustment functions are required for the WSN topology [5]

- The sensor node in a WSN can only communicate with its immediate neighbours. However, only a multi-hop path will suffice if a node needs to communicate outside RF (radio frequency) range. The intermediary nodes in this multi-hop path can transfer data.

- Architectural boundaries aren't a problem for WSN because it's flexible by nature.

- All wireless sensor nodes can be accessed through a central monitoring system.

- Because WSN is wireless, it does not require cables or wires.

- WSNs, which refer to the difference between a wired and a wireless connection, can be employed on a large scale in various situations, including industrial, medical, monitoring, and agriculture [6].

Positioning is commonly used to analyze the sensor node location in wireless sensor networks. Installing a Global Positioning System across each sensor is expensive in a WSN with hundreds of sensors. GPS (Global Positioning System) may not provide accurate location results in an inside area. Individually generating location references across each sensor network in a complicated system is not practicable [7]. To solve the shortcomings of prior location tracking, the GPS (Global Positioning System), which is among the most extensively used techniques for location, was established in 1973. Defence, industrial, and, more recently, customer uses have all

employed GPS. GPS delivers three-dimensional geolocation data and needs a direct connection with at least four spacecraft, with only a precision of 0.3 m.

Furthermore, GPS is restricted in its usability in dense forests, hillsides, and even indoor spaces. It does not function over obstructions that hinder LOS (line of sight) transmission between the spacecraft and the Global positioning system. Several of these issues are solved by employing wireless signals for positioning [8], especially with the construction of cellular towers, which offer enough data to locate mobiles. The wireless sensor network can overcome GPS's location limitations. Scientists have developed non-GPS methods for identifying nodes throughout the wireless sensor network. Anchoring detectors are sensors placed in predefined locations and used to locate all other devices in random locations using arrival time and acquired signal intensity approaches. Analyzing location-related factors can infer a component in an unknown area. Techniques like sequential identification, which uses sensors with low interaction bandwidth to achieve location, can be scaled. The main drawback of these methods is that they necessitate exact range and orientation measurements. The embedded methods and methodologies used to obtain location data in indoor spaces include web-access microwaves, structural modelling, multiple antennae, and ZigBee [9].

An improved technique is provided by [13] to minimize localization error in the wireless network. In [13], the DV-Hop algorithm for the localization of normal nodes in WSN was studied. The technique was known as TWDV-hop, where T is the hop threshold and W is the weighted matrix. The algorithm was corrected during the first step in the proposed technique. It resolved the problem of average hop size because with increase in the number of hops between the two beacon nodes, their distance also increases and The average hop size showed much more error. That means the traditional DV-Hop algorithm performs efficiently with the nodes that are much closer to each other and offer a lot of errors if they are farther apart. In [16], a routing protocol based on DV-Hop algorithm is explained known as AODV Algorithm. This protocol integrates the target serial no. and includes routing discovery and routing maintenance. When the source node communicating with other nodes fails to reach the routing of destination node, it requires the grouping of RREQ and it is received by all nodes, It checks whether such information exists or not and then the information should be abandoned accordingly or RREQ is recorded the routing table and broadcasted continuously until some central node or the routing request grouping reaches the routing of destination node.

The existing techniques suffer from high hop value and low accuracy—more accuracy error when the distance between the Beacon nodes is greater. Computational cost is high and requires more energy and, in some cases, incorrect results of position localization. Therefore, there is a need for a methodology where computational cost is low, accuracy is

high, and Beacon nodes are less. Also, there is a need for a methodology where the average hop distance is cut off if the number of hops is more than 3, as it reduces the overall error when finding the location of the unknown nodes. The proposed technique (enhanced-DVHLA) has improved localization accuracy in WSNs. The proposed method created three enhancements depending on the TWDV-Hop, and TWDV-Hop with AODV. (i) It evaluates their estimated distance with the MHN (minimum-hop-number) and signals broadcasting from one node to another for data transmission. (ii) It modifies the average hop-size between UN and BN. After this modification process, the route replies from source to destination node when the signal has been activated. (iii) It calculates the optimized nearest neighbour route based on the minimum distance for data transmission and reduces LE and LER.

The sections of this paper are: The brief Introduction of node localization is explained in section 1. In section 2, various existing methods for node localization are surveyed. The proposed methodology is explained in section 3. The simulation results and discussions are elaborated in section 4. Section 5 defines the conclusion and future scope of the research work.

II. RELATED WORKS

This section provides a survey of several existing methods of node localization in WSNs. Messous S., et al., (2021) [10] presented an enhanced DV-Hop technique to reduce the significant path loss throughout the existing DV-Hop technique. The obtained wireless signal confirmations and the quadratic correction factor were used to measure the location of neural sources and news stations. Furthermore, the suggested technique enhances the effectiveness of position estimation by recursively computing the positioning procedure. The suggested localized approach minimizes path loss and enhances positioning accuracy, according to experimental observations. Kanwar, V., et al., (2021) [11] included a conversion factor in the proposed methodology to change the hop dimensions of the cluster centres. The suggested methodology reduces communication with unidentified and attached endpoints by determining the hop dimensions, including all anchor points at neural sources. The broadcasting abnormality framework demonstrated the usefulness of the proposed automated system in a transversely isotropic system. Simulations were performed in MATLAB, as well as our suggested approach was compared to classic DV-Hop, evolutionary automated process DV-Hop, and metaheuristic optimization-based PSO optimization-based DV-Hop. Compared to existing detection techniques, modelling results indicate that the proposed autoencoder lowers localized failure, error variance, and computation time. Wang, G., et al., (2021) [12] implemented the lowest hop count, as well as a spectrum were set using the RSSI technique so that when the minimal Hop count was similar to the real quantity as well as the location reliability of the DV-Hop method was increased. In a subsequent study, the balanced

refinement for every white node's minimum hop ranging might be done using the balanced parameter to effectively consider the global structure and eliminate the problem generated by the minimum hop distance. Han, F., et al., (2020) [13] studied WND-DV-Hop, an upgraded DV-Hop technique with dynamic load balancing factors and a novel scaled minor-square placement technique. First, the unidentified node's single-hop was adjusted using RSSI (Received Signal Strength Indication) technique. A modified coefficient dependent on the beacon node count was developed to minimize minimum hop range uncertainty. To tackle polynomial equations problems, a novel balanced least-squares technique was incorporated.

Furthermore, extensive experiments were carried out to determine WND-DV-performance, Hop's, and the results compared to effective direction vector-hop, improved DV (direction vector)-Hop, checkout- direction vector hop, and new direction hop shown in the literature. WND-DV-Hop considerably surpassed other localization techniques, according to the results. Xue, D., et al., (2019) [14] presented an improved distance vector hop methodology based on hopping narrowing and range adjustment. The minimum hop distance was reduced by using RSSI (Received Signal Strength Indication) extending technologies. The average path range was reduced using the balanced overall average of hopping range inaccuracy and predicted range inaccuracy. As a result, the Hop-DV localization method's available localization quality had increased, and the positioning error had decreased. The modelling investigation of the revised approach was performed in the MATLAB simulation model. According to the experimental outcomes, the suggested method decreased localization loss and improved localization accuracy.

A N Jadhav et al. 2018 [17] discussed how node location played a vital role in WSNs. It comprised WS (wireless sensor) localization method using EHP (expected-hop-progress) to analyze the area of sensors in the network. The introduced method is based on a careful study of hop-progress in a WSN with random deployed SN and AND (arbitrary node density). The distance between any pair of SNs can be precisely calculated by deriving the EHP from a WSN model in network metrics. The DE (distance estimation) was the central issue in LSs (localization systems) for networks. The research method range-free EHP attains better performance and minimum COH (communication overhead) as compared to some existing techniques such as DV-Hop and AODV protocol. AODV is the RP (routing protocol) based on the distance-vector method that defines the target SNO. (Serial number). This method generally includes RD (routing discovery) and RM (route maintenance). If the start node transmitting with other SNs fails to spread the routing of the sink node, it needs the grouping route request. After the other SNs get this route req, whether such data exists/not. AODV routing protocol, the SN requests for the routing of the sink node through distributed route re message. Like a flooding protocol will unluckily create

an important route req message, outcoming a significant signal error and the COH.

The above discussed methods have various Problems and research gaps which has been discussed below:

- Low accuracy for node localization and low stability of the network.
- More complex to implement in a real-time environment.
- Implementation issues, sparse network issues.
- Computational complexity is high.
- Distance estimation is a main problem in localization systems.
- Node localization error

To eliminate the above research gaps, an enhanced method is proposed in this paper.

III. PROPOSED WORK

The main aim of proposed work is to reduce the LE and improved the LA rate. It will assess the LER in terms of the following metrics: (i) Effect of Total Number of Nodes (ii) Effect of Beacon node ratio, and (iii) Effect of Communication Range. Figure 2 depicts the flow chart of proposed work.

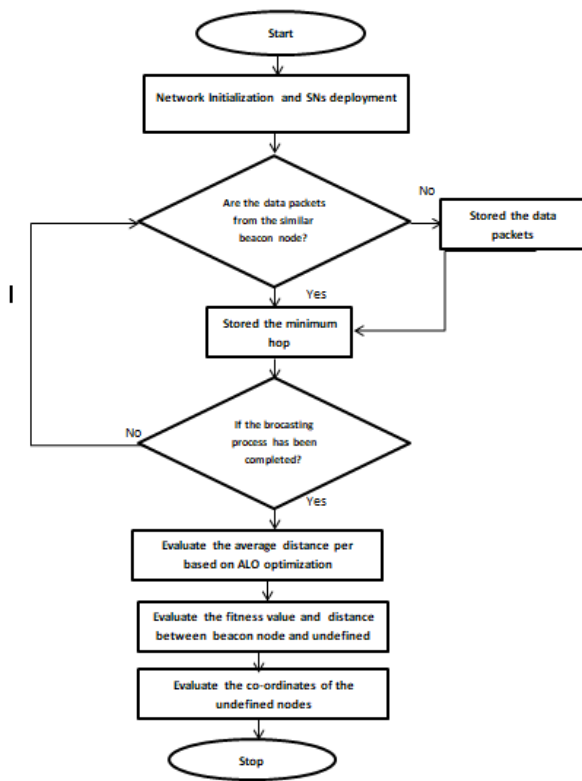


Fig. 2: General Architecture: WSN [2]

The main principle of enhanced DVHLA is to estimate unknown node locations through known node locations and the HC (hop count) between SNs (sensor nodes). WSN networks have two categories of nodes, one is BN whose locations

are known, and the other nodes are UNs (unknown nodes). The main procedure for the enhanced DVHLA method can be described as three steps. (i) Initially, each BN transfers its location throughout the complete network (WSN). By taking the benefit of relay nodes, each node would know the MV (minimum value) of HC to BNs and the corresponding position, if they are associated. (ii) Secondly, the average distance of each hop for BNs is evaluated by the MV of HC and the total distance between BNs. The average distance for a UNs is similar to the BN optimized nearest it. The distances to BNs can be estimated for the UN by multiplying the middle distance by HC's MV to ANs. (iii) Lastly, if there are at least three optimized distances for a UN that have been evaluated in the existing procedure, the location of the UN can be assessed. DV-hop is more straightforward to design than the other LA (localization algorithms). So, it has the demerits of a minimum positioning accuracy rate. This research work has enhanced the DVHLA method by the ALO optimization method HC to optimize distanced EE in this research paper. This proposed method has evaluated the minimum distance and nearest route for data transmission in the network. ALO method helps to optimize the route and distance for high range for data transmission and less error rate as compared with existing methods. This proposed model has reduced the LE and improved the LA rate.

IV. SIMULATION RESULTS AND DISCUSSIONS

A. Arithmetic Formula

LE and LER are used to calculate the achievement of our research based on enhanced-DVHLA algorithm.

- LE (localization error): The difference between real and evaluated coordinates of UNs, its expression is defined in (1)

$$LE = \sqrt{((x1_u - x2_a)^2 + (y1_u - y2_a)^2)} \quad (1)$$

- LER (localization error radius): It is the ratio of average LE to the CR (communication range).

$$LER = \sum_{U,n=1}^N (\sqrt{((x1_u - x2_a)^2 + (y1_u - y2_a)^2)}) / N * r \quad (2)$$

B. Simulation Environment

An illustration of sensor node deployment in two-dimensional space is shown below in fig. 3. The total number of 100 sensor nodes is randomly deployed in the 100*100 meters sq. area, adding twenty BNs (beacon nodes) denoted by the '*' pink star. In training to test the network performance of the Enhanced-DVHLA method, simulations were designed in MATLAB 2018a. The simulation results were compared with TWDV-Hop [13] and TWDV-Hop with AODV [16]. Table 1 defines the experiment metrics.

To better study the research method, all simulations for methods were performed as many as a hundred times for

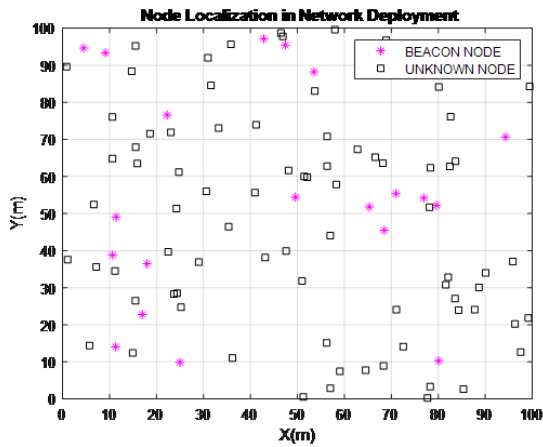


Fig. 3: Node Deployment

TABLE 1. SIMULATION METRICS

Metrics	Values
Size of network	100*100 m
Total number of nodes	100
Beacon Nodes	20
Communication range	25

each outcome since SNs are randomly deployed in the screen field. The proposed model used the average value to calculate the enhanced method. This experiment was defined under the scenario that a hundred SNs were unevenly plotted in the network area of 100*100 meters sq. with 20 per cent BN. The CR (communication range) is 25 meters.

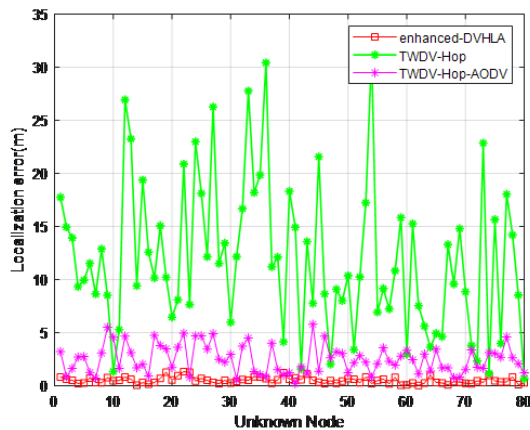


Fig. 4: LR for each UN (unknown node)

Fig. 4 shows the localization error for each unknown SN under three methods in a similar environment. It is observed that the proposed method (enhanced-DVHLA) gave better results. The LE of TWDV-Hop is around 6m and almost

three times larger than the enhanced DVHLA method. All LE of enhanced-DVHLA is between 2m and 3m, with a flat modify trend and almost near direct line, which means the network performance of enhanced-DVHLA is more reliable and efficient. The reduction of LE in the enhanced-DVHLA method is 77 percent. Compared with TWDV-Hop [13] and TWDV-Hop - AODV [16], the error is reduced to around 55 per cent, and 35 percent, respectively.

TABLE 2. COMPARISON ANALYSIS OF LE AND STANDARD DEVIATION

Localization Algorithms	Max-LE (m)	Min-LE (m)	Avg-LE (m)	Std-LE
TWDV-Hop	3.97	2.965	3.425	0.22
TWDV-Hop AODV	3.935	1.881	3.391	0.93
Enhanced DVHLA	2.446	1.47	0.056	0.21

Table 2 shows the comparative analysis of LE and its SD (standard deviation) under three-node localization methods. Compared with the other two ways, the enhanced-DVHLA process has achieved minimum LE in the form of MIN, MAX, and AVG. The proposed method has minimum SD, which indicates the enhanced-DVHLA way had better efficiency and accuracy.

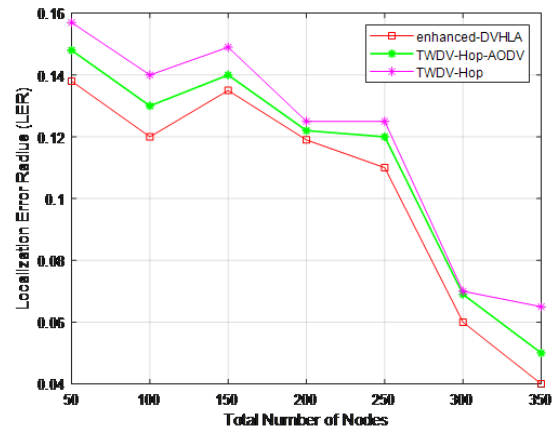


Fig. 5: LER under several total no. of SNs

Fig. 5 defines the localization error radius w.r.t total number of SNs. It defines a downward trend an increase in SNs under three methods. The proposed method always showed the minimum error radius under all conditions, normally when the SNs exceeds 120. The LER of the research method decreased by 70 per cent, and 55 per cent when compared with TWDV-Hop [13] and TWDV-Hop-AODV [16] methods, respectively.

The total no. of SNs was regularly increased from 50 to 350. Though, the CR (communication radius) and the proportion of BNs ratio are fixed at 25 meters and 20 percent, respectively.

TABLE 3. COMPARATIVE ANALYSIS WITH LER VS. TOTAL NO. OF NODES

Localization Algorithms	Max-LER	Min-LER	Avg.-LER
TWDV-Hop	0.1848	0.0426	0.0804
TWDV-Hop AODV	0.148	0.050	0.0791
Enhanced-DVHLA	0.138	0.040	0.068

Table 3 compares LER of proposed method with TWDV-Hop [13] and TWDV-Hop with AODV [16] method.

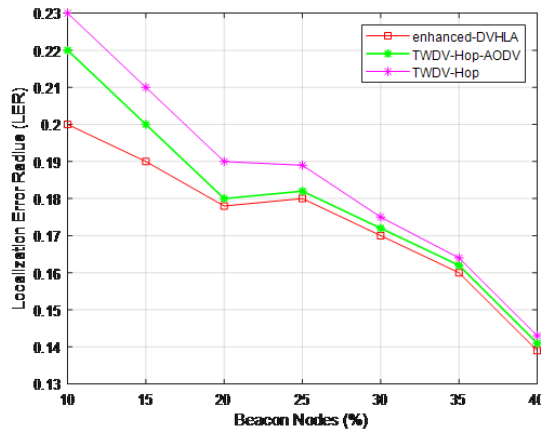


Fig. 6: LER under BNs (beacon nodes)

Fig. 6 defines the LER with variation in BN ratio. The research method consistently scored the minimum LER under all conditions using optimized route selection, commonly when the BN exceeded 30. LER of the research method decreased by 40 per cent, and 30 per cent when compared with different node localization methods such as TWDV-Hop [13] and TWDV-Hop-AODV [16], respectively.

TABLE 4. COMPARATIVE ANALYSIS WITH LER VS. BNS

Localization Algorithms	Max-LER	Min-LER	Avg.-LER
TWDV-Hop	0.2309	0.1594	0.1819
TWDV-Hop-AODV	0.221	0.1551	0.1781
Enhanced-DVHLA	0.20	0.1390	0.1652

Table 4 shows that the research method enhanced-DVHLA outperformed the rest under LER, with the avg. Accuracy rate reaching up to 90 per cent. The localization accuracy is under avg. form of the research (enhanced-DVHLA) method decreased by 42.23 per cent, and 31.4 per cent, compared with TWDV-Hop [13] and TWDV-Hop-AODV [16] methods, respectively.

Fig. 7 defines the LER with variation in CR (communication range); the enhanced-DVHLA always showed the lowest value

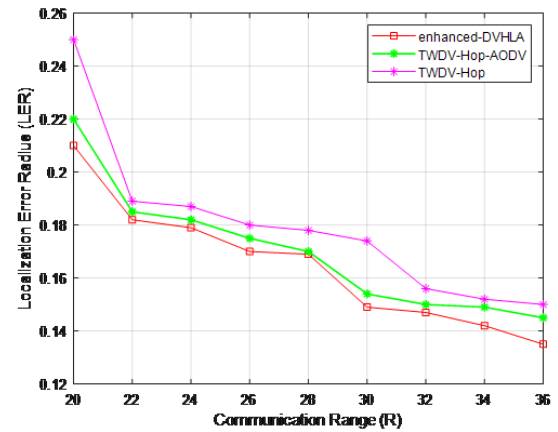


Fig. 7: LER under Communication Range (CR)

under all conditions. The localization error radius of the research method enhanced-DVHLA decreased to 35 per cent, and 37 per cent, compared with TWDV-Hop [13], and TWDV-Hop-AODV [16], resp.

TABLE 5. COMPARATIVE ANALYSIS WITH LER W.R.T. CR

Localization Algorithms	Max-LER	Min-LER	Avg.-LER
TWDV-Hop	0.2154	0.1184	0.1563
TWDV-AODV	0.214	0.1150	0.1521
Enhanced-DVHLA	0.20	0.1023	0.1495

Table 5 shows that the research method has better network performance with reduced LER. Compared with the research model, the LER, under avg form, decreased to 40.1 per cent, and 27.8 per cent compared with TWDV-Hop [13] and TWDV-Hop-AODV method [16].

V. CONCLUSION AND FUTURE SCOPE

This research paper introduces an enhanced DVHLA method based on DV-Hop with the ALO Optimization technique. It optimizes the Localization accuracy of the proposed algorithm for random topology WSNs. The research method enhances DV-hop with an ALO algorithm by localization estimation based on position evaluation and optimized evaluation of the average hop distance of SNs. The evaluation of the research model is well-defined through experiment outcomes compared with TWDV-Hop [13] and TWDV-Hop-AODV [16] methods. The localization accuracy of the research method is better than that of the other two localization methods at several random topologies of WSN. An enhanced-DVHLA (enhanced-distance-vector hop localization) method to overcome the existing problems such as high error rate and less accuracy. Under the localization radius, the research approach provided superior network performance. This proposed model has enhanced the localization accuracy rate while reducing

the localization error. The Localization Error Radius (LER) will be evaluated using the following performance measures: (a) Total Number of Nodes; (b) Beacon Node Ratio; and (c) Communication Range Effect. These recommended parameters are compared to currently used approaches. Compared to the other study models, the LE is optimized by more than 77 per cent. The LER dropped by 40.1 per cent, and 27.8 per cent, respectively, when using the TWDV-Hop and TWDV-Hop-AODV methods.

Further improvements such as (i) the proposed technique will be expanded with 3D WSNs in the future. (ii) Noting that NL (network lifetime) is a significant problem in the wireless sensor network. It plans to explore a trade-off between LA (localization accuracy) and EC (energy consumption) for the research method.

ACKNOWLEDGMENT

Author would like to thank the Director, National Institute of Technical Teachers Training and Research, Chandigarh and also extend my sincere gratitude to Dr. Kanika Sharma, Asst. Professor, NITTTTR Chandigarh and Dr. Amod Kumar, Professor, NITTTTR Chandigarh for their sincere advice that helped me to complete my project.

REFERENCES

- [1] Zheng, J., Jamalipour, A., "Introduction to wireless sensor networks". *Wireless Sensor Networks: A Networking Perspective*, 1, 1-18,2009.
- [2] Manuel, A. J., Devarajan, G. G., Patan, R., Gandomi, A. H. "Optimization of routing-based clustering approaches in wireless sensor network: Review and open research issues". *Electronics*, 9(10), 1630,2020.
- [3] Ali, A., Ming, Y., Chakraborty, S., Iram, S. "A comprehensive survey on real-time applications of WSN", *Future Internet*, 9(4), 77,2017.
- [4] Ramson, S. J., Moni, D. J. "Applications of wireless sensor networks—A survey", In *2017 international conference on innovations in electrical, electronics, instrumentation and media technology (ICEEIMT)* (pp. 325-329). IEEE,2017.
- [5] Zhang, S., Zhang, H." A review of wireless sensor networks and its applications", In *2012 IEEE international conference on automation and logistics* (pp. 386-389). IEEE,2012.
- [6] Gupta, S. K., Sinha, P. "Overview of wireless sensor network: a survey". *Telos*, 3(15 μ W), 38Mw, 2014.
- [7] Cheng, L., Wu, C., Zhang, Y., Wu, H., Li, M., Maple, C. "A survey of localization in wireless sensor network", *International Journal of Distributed Sensor Networks*, 8(12), 962523,2012.
- [8] Kuriakose, J., Amruth, V., Nandhini, N. S. "A survey on localization of wireless sensor nodes", In *International Conference on Information Communication and Embedded Systems (ICICES2014)* (pp. 1-6). IEEE,2014.
- [9] Khalaf, O. I., Sabbar, B. M., "An overview on wireless sensor networks and finding optimal location of nodes", *Periodicals of Engineering and Natural Sciences*, 7(3), 1096-1101,2019.
- [10] Messous, S., Liouane, H., Cheikhrouhou, O., Hamam, H. "Improved Recursive DV-Hop Localization Algorithm with RSSI Measurement for Wireless Sensor Networks", *Sensors*, 21(12), 4152,2021.
- [11] Kanwar, V., Kumar, A. "DV-Hop-based range-free localization algorithm for wireless sensor network using runner-root optimization", *The Journal of Supercomputing*, 77(3), 3044-3061,2021.
- [12] Wang, G., Shi, Y., Liu, J.," Optimization of the DV-hop Localization Algorithm in Wireless Sensor Networks". In *Journal of Physics: Conference Series* (Vol. 2037, No. 1, p. 012088). IOP Publishing,2021.
- [13] Han, F., Abdelaziz, I. I. M., Liu, X., Ghazali, K. H. "An Enhanced Distance Vector-Hop Algorithm using New Weighted Location Method for Wireless Sensor Networks". *International Journal of Advanced Computer Science and Applications*, 11(5),2020.
- [14] Xue, D. "Research of localization algorithm for wireless sensor network based on DV-Hop". *EURASIP Journal on Wireless Communications and Networking*, 2019(1), 1-8,2019.
- [15] Garg, V., Jhamb, M., "A review of wireless sensor network on localization techniques", *Int. J. Eng. Trends Technol*, 4(4), 1049-1053,2013.
- [16] Jadhav, A. N., Patil, P. B," Comparative Analysis of AODV and DV-HOP Wireless Node Localization Techniques", *International Journal of Scientific Development and Research (IJS DR)*,3(6),pp:1-6, 2018.
- [17] Jadhav, A. N., and Prajakta B. Patil. "Comparative Analysis of AODV and DV-HOP Wireless Node Localization Techniques." Volume 3, Issue 6,pp-44-51.

An Exploration of Mis/Disinformation in Audio Format Disseminated in Podcasts: Case Study of Spotify

Kevin Matthe Caramancion

College of Emergency Preparedness, Homeland Security, and Cybersecurity
University at Albany, The State University of New York
Albany, New York, United States
kcaramancion@albany.edu / www.kevincaramancion.com

Abstract—This paper examines audio-based social networking platforms and how their environments can affect the persistence of *fake news* and mis/disinformation in the whole information ecosystem. This is performed through an exploration of their features and how they compare to that of general-purpose multimodal platforms. A case study on Spotify and its recent issue on free speech and misinformation is the application area of this paper. As a supplementary, a demographic analysis of the current statistics of podcast streamers is outlined to give an overview of the target audience of possible deception attacks in the future. As for the conclusion, this paper confers a recommendation to policymakers and experts in preparing for future mis-affordance of the features in social environments that may unintentionally give the agents of mis/disinformation prowess to create and sow discord and deception.

Keywords—*Fake News, Misinformation, Disinformation, Technology Policy, Podcast, Spotify*

I. INTRODUCTION

Information disorders, misinformation and disinformation, are still continuously evolving in digital communities. Colloquially known as *fake news*, their forms vary, but the most prevailing ones are presented through news headlines and fabricated articles. Additionally, visually manipulated media content magnifies the intensity of deception provided by misinformation and disinformation. However, an unusual but classic form of medium, audio communications, and radio broadcasts are recently emerging and exhibiting a very high potential in functioning as falsehood dissemination machines. Mis/Disinformation are still both types of information, and naturally, they can pollute any social platform that is grounded on information sharing, be it of any format, including audio.

The recent discourse on mis/disinformation on audio communication recently came to light when a group of 270 experts, scientists, and medical doctors called out Spotify to establish a clear and concise public policy on misinformation propagated in their platform. Spotify, as a tech company and an audio streaming provider, is bestowed by Section 230 of the U.S Communication Decency Act of 1996 to have the regulatory authority of the content on their platforms. In essence, like any other tech platform, this gives Spotify the

prerogative, *the right but not the responsibility*, to regulate all content in their environments, including the deceptive ones.

The inclusion of audio-formatted cyber deceptions is virtually underrepresented in the existing studies revolving around *fake news*, its properties, and even societal effects. The common misconception that the blanket term *fake news* is exclusive to textual content is paradoxically misleading [1], for the umbrella term *fake news* encompasses and includes non-textual media content such as *deepfakes*, altered images, and even legitimate audio broadcasts contextually spreading lies.

This paper examines the current landscape of audio-anchored information environments, their difference from that of the ones built on general multimodal format, and how they possibly amplify or de-amplify mis/disinformation campaigns. Platforms of this nature, such as Spotify and Apple Music, have more than 400 million active subscribers [2], and their capability to spread information, legitimate or misleading, is, without question, solidly robust. Furthermore, the recent event surrounding Spotify, a music app with no explicit messaging and communication system between its users, left scholars and experts with a sudden interest in these environments as information creation and dissemination machines.

This is an exploratory work grounded on the primal research question, “*What are the similarities and differences of audio-based social networks with the general multimodal platforms with respect to information creation and dissemination processes?*” This will be explored through a benchmarking of information creation and dissemination feature analysis with the latter as the reference. A supplementary demographic vulnerability analysis on the existing statistical data on their users will be further performed to compare the existing podcast audience with that of the prevailing mis/disinformation studies.

The literary contribution of this paper is through the exploratory analysis it confers on the under-investigated particular domain of audio-based information environments. This paper offers insights on the dissemination features of these current environments so that future controls may be projected based on their affordances. As social media

technologies evolve rapidly, unintended consequences beyond their promised features should be mitigated to better promote the well-being of digital citizens depending on these for their information source.

More than its literary significance, the practical implications of this paper rests with the hope that it influences the current and future regulations involving misinformation and disinformation, particularly to audio-based platforms. A coordination in policies across all social networking platforms is highly critical at this point in time. A deceptive content appropriately controlled in one platform but still persists in another will still reverberate in public discourse and can still likely wrongfully shape perception. A carefully shaped policy, particularly for audio-based platforms, will shield these environments from being used as the loophole for *fake news* creation.

The format-wise of this exploratory paper is as follows. The next section, literature review, will provide a quick summative precis of the fundamental background on the topic of *fake news* predictive modeling (vulnerability analysis) used in projecting the cyber risk of individuals, the role of information environments, their design and structure, and how it contributes and prevent to the persistence of *fake news*. The literary section concludes with the seminal gap that this paper will fill in the subsequent chapters. It will then be followed up by the outright comparison of audio-based platforms with multimodal ones. After the establishment of these literary backgrounds and theoretical underpinnings, a case study on the recent issue of Spotify will be the *application area* of this paper, including a demographic analysis on the users (i.e., streamers) of podcast shows. Finally, the conclusion section of this paper bestows its recommendation on this domain and introduces the next stream of empirical research on the area explored by this paper.

II. LITERATURE REVIEW

A. Predictive Modeling of Mis/Disinformation Vulnerability

Mis/Disinformation cyber risk modeling is the mathematical quantification of the projected risk of social media users to massive computational propaganda [3]. It is an application of predictive modeling where the results are interpreted to forecast the likelihood of certain users and segments in the population to fall prey to mis/disinformation attacks [4]. In humans, the metrics (i.e., dependent variables) used to measure the effectual damages are usually sourced from psychological literature such as the perceived doubt and the actual deception ignited by falsehoods campaigns [4]. The purpose of such endeavor is to strengthen the capabilities of the users, especially the revealed vulnerable groups, to such attacks of deception so that proper educational conditioning and interventions can be used to build their resistance to the effects of mis/disinformation.

The independent variables, on the other hand, are spread across multiple domains. The existing studies have exhaustively tested the most optimal configurations on which

are the predictors, moderators, mediators, and control variables. The most visible factors are the characteristics of social media users as humans, such as their demographic and socioeconomic status [3], native language [6], veteran status [7], personal history of being deceived [8], and even psychological and political perceptions [4]. Outside the human conditions, this cyber risk of being deceived by *fake news* may be due to environmental stimuli such as time of the day [5], metadata describing the content [9], and even the way the information flow of an environment based on a platform's structure and design [10]. These are all hypothesized to contribute to the cyber risk of an individual in being mis/disinformed; This stream of research is currently open since the current predictive models have not yet perfected (100% accuracy) the forecasting of such cyber risk/vulnerability.

B. The Information Environment as Cyber Risk Factor

One of the most significant predictors of the vulnerability of social media users to mis/disinformation attacks is the very design of the platforms themselves [10]. In a quick recap, Caramancion (2021) highlighted that this is because the design and policies of an environment dictate how information flows across the medium. The designation on who may create, share, or exclusively receive information (in the form of content) creates a system dynamic on who will be the information (or mis/disinformation) consumers and who has the authority or power to approve or invoke censorship [4]. Caramancion (2021) further raised the difference between the two major types of information environment designs, taxonomic (top-down) and folksonomic (bottoms-up). The importance of this factor is even more pronounced when [10] suggested that all the other remaining factors, human and non-human, are merely secondary to the way an environment is shaped.

Absolute taxonomically designed environments are highly authoritative due to the strict entry and flow of content [11]. Traditional library systems are examples of such design where the classifications used in indexing are usually standardized. The dawn of folksonomy [12], on the other hand, paved the way for the rise and pervasiveness of social networking platforms. Unlike taxonomic systems, the ability to create and disseminate information has been extended even up to the users [13]. For the longest time, this design has been the strength of these networks, but the price that comes with it is the creation and persistence of mis/disinformation [10]. Modern social networking platforms are not dichotomously restricted to either of these but rather made from the combination of both taxonomic and folksonomic influences. For instance, even though individuals may create and share any content as they wish, the privilege of labeling its legitimacy and censoring it still rests with regulators of an environment through the help of independent fact-checkers [14].

C. *The Literary Gap: The Missing Case of Audio-Based Environments*

With the above mentioned as the foundation, the following corollaries are evident:

a. That the current analyses on the information environments and their design are based on the context of mostly textual resources and general-purpose multimodal platforms. Exclusively audio-based social networks are, on their own, operating at a specialized locus, which warrants these networks an inquiry of their own distinct that of the context of general-purpose multimodal platforms.

b. That an analysis on the merits of both taxonomic and folksonomic features on these audio-based environments is virtually *missing* in scientific and commercial literature. As a significant predictor of mis/disinformation vulnerability, an information environment is only grounded on the sum of all its features and the affordances they offer. These include both intended and unintended consequences of such features.

c. That podcast, as the medium of misinformation and disinformation, warrants an inquiry *particularly focused on its context*. Podcasts, by their very nature, are different from textual news headlines and video clips; thereby, the mis/disinformation magnitude and flow that proceeds from them will naturally behave differently. The challenges and opportunities in podcast regulations may or may not resemble that of the typical content format of social platforms.

d. That the demographic and socioeconomic factors as predictors of mis/disinformation vulnerability may or may not be applied on audio-based environments. The existing studies of predictive modeling are mostly (if not all) based on general multimodal platforms, and the demographics of podcast listeners may or may not be in congruence with that of the general multimodal platforms.

III. FEATURE ANALYSIS OF AUDIO-BASED PLATFORMS

A. *Similarities with General Multimodal Platforms*

1. *The Presence of Social Feeds and User Profiles*

One hallmark characteristic of social platforms is the ability of information consumers to have their own social profiles (or user profiles). This, in essence, allows them to be a node in these networked social environments and create their identities. As nodes, the information flow can be traced, and the roles of being a sender or receiver can be technically defined. In the context of mis/disinformation attribution, this is essential, especially in tracing the source of falsehood attacks, the enforcers who spread the deception (i.e., amplifiers), and finally, the intended or unintended targets.

2. *The Presence of Behavioral and Engagement Metrics*

Content-wise, the metrics used to quantify the effects and reach of content, legitimate or not, is an important indicator of its influence on an audience. The number of hits, engagements, and even user views are all examples of such quantifiers. The purpose for such metrics may be used outside the usual projected monetary value they attempt to measure. For instance, in the context of mis/disinformation, the social

reach and virality of deceptive content may indicate how it compares to a truthful one. Information scientists and experts try to investigate the properties of such content for the purpose of making sense of it.

3. *The Presence of Collaborative Tools*

Collaboration features dictate the capability of users in social networks, through collective acts, to work on a common task. Note that these may be in the form of groups productivity tools or group messaging features in multimodal platforms. The focus, however, is not on what they do but what they exhibit—a network in enforcement. In the context of mis/disinformation, this exhibits the abilities of threat agents to perform networked adversarial attacks of deception. Most of the prevailing nation-sponsored massive propaganda (information warfare) rely on network infrastructures to release payloads.

4. *High Interoperability (as Sender) with other Social Apps*

The ability of content to be exported to another social networking platform is a hallmark of intra-environment connectivity and content transfer. With the serialization-and-deserialization construct as inspiration, platform-to-platform interoperability increases the ease of use on the part of users, especially when types of content that are usually shared can be easily migrated. These content—such as videos, music, website links—when shared from another social platform doesn't need a re-instantiation in the new platform. In the context of mis/disinformation, this design exhibits how a polluted platform can echo deceptive content to other platforms with weaker mis/disinformation policies.

5. *The Folksonomic and Taxonomic Influences are Present*

Information organization and its hallmark constructs of folksonomy and taxonomy are present. This can be seen in the way content sorting and categorizations are designed. Audio-based environments and their content can be retrieved and sorted alphabetically based on a resource's title or its associated stakeholders, such as artists and producers. Other related folksonomic metadata such as user's taste of music, preferred artists, and related crowd preference usually triggers recommendations to a user. The biggest implication of this to the vulnerability of users to mis/disinformation attacks is the creation of filter bubbles through hyper-personalization.

B. *Differences with General Purpose Multimodal Platforms*

1. *Lack of Direct User-to-User Communication*

The most visible lacking feature of audio-based platforms is the absence of explicit means to communicate, textual messages and call functions, in them. Unlike multimodal platforms where feedback in the form of comments or even the ability of a user to privately send a message to another user is virtually missing. With this absence, as an applied analysis on the domain of *fake news*, preliminary conception may lead one to believe that this may contribute to the non-promotion of *fake news*.

2. *Approval Required for Content Creation*

Influential and official channel sources of communicative audio broadcasts are usually subject to verification by the moderators of tech platforms. A verified status implies that the legitimacy of the claimed title is sustained. The important note to consider, however, is this verification does not imply or extend to the content disseminated by a verified account. Instances where accounts of such verifications have spread and are currently disseminating harmful, deceptive propaganda remain existing.

3. Peer Activity Monitoring

An audio-based platform, Spotify, allows users to explicitly reveal their listening activity to the people who follow them. This feature, previously included in prevailing multimodal platforms such as Facebook and Instagram, was removed due to the privacy concerns they pose to users. With regards to mis/disinformation attacks, there is a close correlation between privacy and vulnerability to attacks of deception—agents of propaganda create particularized content depending on the applicability of a topic to users and populations.

4. Lack of Personal Broadcasts

The personal broadcast here refers to the ability of a non-authoritative user (e.g., streamer, listener, etc.) to be the source, either producer or creator, of audio information. Unlike the multimodal platforms where users have the prerogative to create virtually any content, be it of any format, users in audio-based platforms at most can create personalized collections such as playlists or saved resources for their personal use, to which in few circumstances can be shared to their connections. As a hypothetical foundation, one may lead to believing that this leads to fewer and weaker instances of *fake news* on these platforms.

5. Exclusively Receiving Node for Interoperability

In the previous section of this chapter, it was noted that the audio-based environment's contents are easily transferrable to other platforms—the reverse, however, is not the same. This is due to the restrictions in the formatting of the audio platforms and the limited capabilities the users are authorized to perform. This characteristic, when superficially analyzed, can lead to the assumption that audio-based platforms are shielded from external injections of mis/disinformation attacks. The case study in the latter chapter examines this *a priori*.

C. Challenges and Opportunities

1. Need for Audio-Based Mis/Disinformation Detection Technologies

Should audio-based environments proven to be a potential breeding ground of mis/disinformation networks, the most visible danger this poses to the scientific and technological community is that most of the deployed AI and ML-based *fake news* technologies are all based on semantic, linguistic, and visual benchmarks. The challenge is translating these technologies to their audio counterpart. This is elusively

challenging due to the current limitations of processing audio resources in real-time monitoring.

2. Consider the Demographics of Audio-Based Environments

The interest in audio-based environments, although emerging, is attracting distinct demographics (as discussed in the next chapter) from general-purpose multimodal platforms. This difference is highly important to consider because the predictive models that forecast the vulnerability of social media users to mis/disinformation attacks are experimented based on multimodal platforms such as Facebook, Twitter, or even YouTube—which is also distinct from audio-based platforms. Such difference will reveal the impact of an environment's format on the way users perceive information.

3. The Role of Labels

Labels on audio content metadata have always been authoritative. However, the label confirming the legitimacy of content, however controversial, is not just ethical but also necessary. Labels are created not for the purpose of necessarily assigning legitimacy judgment on content but rather to be the guide to the users that may interact with these. For instance, simple reminders or warning that content may be disputed is enough for this endeavor. Although this approach sounds conservative compared to its counterpart of outright labeling content as false, this shows the responsibility of the platform in maintaining the integrity of their environments.

4. Simplify the Crowd Reporting

The simplest yet most effective method to tag questioned content is to allow users to report it. This has been, for the most part, present in audio-based environments, but the particular focus on deceptive content, especially in podcasts, should be promoted. This is due to the reason that most algorithmic detections of *fake news* are not deployable to audio content, and as such, the dependency on the role of users will be of high significance. Without such capability, audio-based environments have virtually no remedial or detection mechanism to deceptive content.

5. The Role of Policies Particularly Made for Audio Content

One of the important, if not most, in the non-promotion of misinformation and disinformation in any social platforms, is the establishment of well-crafted policies for their control. Without explicit policies or even acknowledgment of the plausible formulation of deceptive content, a platform implicitly enables falsehood information to persist in its environment. Clear policies that stipulate the categorizations, processes, and remedial response to such occurrences are robust means in building resistance against the agents of deception.

IV. APPLICATION & DISCUSSION

A. The Case of Spotify and COVID-19 Vaccine Mis/Disinformation

In early 2022, Spotify was at the center of debate on free speech and misinformation when one of the most influential

podcast channels hosted by Joe Rogan had given a speaking platform to an infamous medical physician, Dr. Robert Malone, on promoting and spreading COVID-19 debunked conspiracy theories. Such falsehood includes the claim of mass formation psychosis, where the public is hypnotized, similar to the point of compulsion, to act against their will in vaccinating themselves and follow related mandates such as masking and social distancing. Included in the claim that was raised in the podcast was the conspiracy that hospitals and medical providers get incentives when falsely diagnosing patients with positive results. This discourse trended on other social platforms such as Facebook and Twitter, with the medical experts calling out Spotify to act on the brazen and deliberate misinformation coming out on the said podcast channel. Further amplifying the discourse is when other artists pressured Spotify to censor the channel or else they would withdraw their content from the platform.

This case highlights several findings, and they are as follows:

Discussion 1: Audio-Based environments, through Podcasts, can be a Potent Source of Mis/Disinformation

The channel, with hundreds of millions of subscribers and frequent listeners, is currently considered a top podcast on the platform. At the point of the interview, it relayed unrelentless conspiracy theories already debunked on other platforms. This event, as an incident, left unsuspecting advocates surprised that a music application can be (and did) become a dissemination machine of public deception.

Discussion 2: Deceptive Content in a Platform Echoes All Throughout the other Platforms

The interoperability of the whole information ecosystem is highlighted in this case when an agent who is censored (i.e., banned) in other platforms was given an opportunity to create and spread mis/disinformation on Spotify. It yielded the remedial actions of other platforms ineffectual, the topic coming from Spotify, nonetheless trended on Reddit, Twitter, and Facebook. On the side of the spectrum, this interoperability has been used by advocates to call out Spotify using other platforms to act on the deception that transpired on its platform.

Discussion 3: Lack of Clear Policies Against Mis/Disinformation is a Vulnerability

At that point of the incident, Spotify has virtually no remediation or response in place to such occurrence. Thus, no warning or label was present to give cues to the listeners that the podcast content may be misleading or at the very possibility of it currently being disputed. Furthermore, the fact that the medical experts and other artists called out Spotify using other platforms to act on the issue displayed that there is no recourse (or acknowledgment even) that misleading content may exist on Spotify's platform.

B. Demographic Analysis of Audio-Based Platforms

As a supplementary analysis, this paper adds an overview of the demographics of the streamers (or listeners) of podcasts [15] [16]. This is due to the fact that the target audience

considered by the existing predictive models is based on the users of multimodal platforms. The most significant implication of this is that the weights of the factors considered in the current models will not necessarily be accurate when applied to audio-based environments. The vis-à-vis comparison, however, of the demographics between the two types of platforms will not be within the scope of this paper. The focus of this section is to identify potential statistics that may have significant implications to the future cyber risk models of mis/disinformation when audio-based environments are accounted for.

Discussion 1: Most Podcast Listeners are Aged 35-54

Age has always been a considered factor when user data are entered into regression models or ANOVAs to predict the vulnerability of people against mis/disinformation. This is a significant deviation from the usual multimodal platforms where younger populations are the dominant age groups [15][16] in content creation and consumption. The effects of this composition, if there is any, on the system dynamics and interactions with users remain unknown.

Discussion 2: Comedy is the Most Popular Podcast Genre

Unlike multimodal platforms where users have variations in their preference of content, the most dominant choice of genre in podcasts is comedy. This choice may or may not contribute to the cyber risk of populations, especially since satire and humorous posts are often associated with frequent misleading perceptions. Furthermore, the motivations and purpose of a user's information retrieval may dictate the echo chamber one belongs to.

Discussion 3: At least 80% of Podcast Listeners Finish Most of the Episodes

Interestingly, most listeners finish the content of the podcast that they are streaming. This behavior, under the category of information consumption, should be further compared to how information is consumed in multimodal platforms. The commitment of podcast streamers to finish content may be suggestive of the depth of deceptive content that may potentially reach into a person's cognition.

Discussion 4: Over 1/4 Podcast Listeners Have a College Degree

Although the effects of *advanced* academic achievements in *fake news* detection have been widely refuted, the lack of basic information/media literacy training [20] and background in discerning fact from fiction can be a potential cause of vulnerability of a user falling prey to deception. A college degree alone is a weak indicator of vulnerability, but it nonetheless may still be suggestive of basic training in fact-checking and the objective reliability of information sources across academic disciplines.

V. CONCLUSION, RECOMMENDATIONS, & FUTURE WORKS

Like a true villain, the omnipresence of misinformation and disinformation reverberates across all platforms regardless of the scheme of their environments. These cyber

deceptions [17][18][19], and their agents, exploit the weaknesses in the guards in place to protect the platforms from falsehood propaganda. This paper, through an in-depth examination of audio-based environments, unequivocally affirms that podcasts exhibit a high potential to be factories of *fake news*. This is due to the mis-affordance of the features used to disseminate authoritative information and the lack of explicit policies outlining the remediation and response to such misuses, including proper educational training [20] to counter these information warfare.

At this point, total annihilation of mis/disinformation is elusive; however, simple measures such as the establishment of clear policies that lay out the regulations for deliberately misleading content will significantly reduce the opportunities for agents of deception to pollute the audio-based environments. Moreover, timely response from the tech providers, once a feature is misused for purposes outside its intended objectives, can never be overemphasized. The inherent characteristic of technologies to be constantly evolving will be used by threat actors to find newer exploits.

This paper's hope is to call out the attention of the experts to consider unusual but potential surface attacks of mis/disinformation agents. Future iterations of this work will be the analysis of other platforms that may possibly be a new environment for exploitation once the loopholes for audio-based environments are closed. Finally, a new research stream for podcast streamers as the intended targets of deception and as the subjects for mis/disinformation vulnerability and risk modeling [21] will be looked out for.

REFERENCES

- [1] Caramancion, K. M. (2021, October). Textual vs. Visual Fake News: A Deception Showdown. In *2021 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (pp. 31-35). IEEE.
- [2] Spotify. (2022, February 2). Retrieved February 21, 2022, from <https://newsroom.spotify.com/company-info/>
- [3] Caramancion, K. M. (2021, April). The demographic profile most at risk of being disinformation. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-7). IEEE.
- [4] Caramancion, K. M. (2021, May). Understanding the association of personal outlook in free speech regulation and the risk of being mis/disinformed. In *2021 IEEE World AI IoT Congress (AllIoT)* (pp. 0092-0097). IEEE.
- [5] Caramancion, K. M. (2021, September). The Relation Between Time of the Day and Misinformation Vulnerability: A Multivariate Approach. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)* (Vol. 1, pp. 150-153). IEEE.
- [6] Caramancion, K. M. (2022, January). The Role of User's Native Language in Mis/Disinformation Detection: The Case of English. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0260-0265). IEEE.
- [7] Committee on Veterans' Affairs. (2020, December). Hijacking Our Heroes: Exploiting Veterans Through Disinformation on Social Media. In *Congressional Reports*. U.S. Government Publishing Office
- [8] Caramancion, K. M. (2021, October). The Role of Subject Confidence and Historical Deception in Mis/Disinformation Vulnerability. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0541-0546). IEEE.
- [9] Caramancion, K. M. (2020, September). Understanding the impact of contextual clues in misinformation detection. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE.
- [10] Caramancion, K. M. (2021, November). The Role of Information Organization and Knowledge Structuring in Combatting Misinformation: A Literary Analysis. In *International Conference on Computational Data and Social Networks* (pp. 319-329). Springer, Cham.
- [11] Rosenfeld, L., & Morville, P. (2002). *Information architecture for the world wide web*. "O'Reilly Media, Inc."
- [12] Trant, J. (2009). Studying social tagging and folksonomy: A review and framework.
- [13] Luca, M. (2015). User-generated content and social media. In *Handbook of media Economics* (Vol. 1, pp. 563-592). North-Holland.
- [14] Graves, L., & Amazeen, M. A. (2019). Fact-checking as idea and practice in journalism. In *Oxford research encyclopedia of communication*.
- [15] Winn, R. (2021, December 28). *2021 Podcast Stats & Facts (New Research From Apr 2021)*. Podcast Insights®. <https://www.podcastinsights.com/podcast-statistics/> Statista. (2022, February 11). *Share of people who listen to podcasts monthly in the U.S. 2017–2021, by age group*. <https://www.statista.com/statistics/>
- [16] Statista. (2022, February 11). *Share of people who listen to podcasts monthly in the U.S. 2017–2021, by age group*. <https://www.statista.com/statistics/>
- [17] Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. *Data*, 7(4), 49.
- [18] Caramancion, K. M. (2020, March). An exploration of disinformation as a cybersecurity threat. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 440-444). IEEE.
- [19] Caramancion, K. M. (2022, June). Same Form, Different Payloads: A Comparative Vector Assessment of DDoS and Disinformation Attacks. In *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE.
- [20] Caramancion, K. M. (2022). *An Interdisciplinary Assessment of the Prophylactic Educational Treatments to Misinformation and Disinformation* (Doctoral dissertation, State University of New York at Albany).
- [21] Caramancion, K. M. (2022, June). Using Timer Data to Conjunct Self-Reported Measures in Quantifying Deception. In *2022 IEEE World AI IoT Congress (AllIoT)*. IEEE.

Same Form, Different Payloads: A Comparative Vector Assessment of DDoS and Disinformation Attacks

Kevin Matthe Caramancion

College of Emergency Preparedness, Homeland Security, and Cybersecurity
 University at Albany, The State University of New York
 Albany, New York, United States
 kcaramancion@albany.edu / www.kevincaramancion.com

Abstract—This paper offers a comparative vector assessment of DDoS and disinformation attacks. The assessed dimensions are as follows: (1) the threat agent, (2) attack vector, (3) target, (4) impact, and (5) defense. The results revealed that disinformation attacks, anchoring on astroturfs, resemble DDoS's zombie computers in their method of amplification. Although DDoS affects several layers of the OSI model, disinformation attacks exclusively affect the application layer. Furthermore, even though their payloads and objectives are different, their vector paths and network designs are very similar. This paper, as its conclusion, strongly recommends the classification of disinformation as an actual cybersecurity threat to eliminate the inconsistencies in policies in social networking platforms. The intended target audiences of this paper are IT and cybersecurity experts, computer and information scientists, policymakers, legal and judicial scholars, and other professionals seeking references on this matter.

Keywords—Disinformation, DDoS, Cybersecurity, Threat Assessment, Standardization

I. INTRODUCTION

Disinformation attacks, the deliberate method of injecting wrongful information to certain targets, are usually deployed at a massively public scale [1]. Distributed denial of service (DDoS), on the other hand, is the intrusive and forceful disruption of networks and systems by unauthorized hosts through resource exhaustion [2]. The latter, DDoS, is formally classified as an actual cybersecurity attack by the virtue of its very nature to operate at multiple lower layers of the OSI model. While the former, disinformation attacks (or campaigns), are typically considered as mere disorders [3] of social networking platforms and may touch only the highest layer of the OSI, the application (layer 7).

The proposed classification of disinformation to be considered as an actual cyber threat has recently emerged for the purpose of standardization [4]. The world has realized that social networks operate without geographical boundaries. The variations in user policies on the social platforms may indirectly contribute to or even amplify disinformation attacks and their ramifications.

A uniform approach in its prevention and response, is without a doubt, one of the solid prerequisites to winning the battle against this harmful phenomenon.

The prevailing calls for classification of disinformation, although highly compelling, lack any concrete comparison with an actual cyber threat *per se*, a move that will further strengthen the claims for such advocacies. First, if this phenomenon is proven to exhibit the same elements that of an actual cyber threat, it passes the elemental *presence* test. Or suppose should the case has proven to be otherwise; in that case, a new referenced cyber threat can just be comparatively benchmarked until the aggregate results from the plural comparisons suggest and reveal disinformation's rightful nature.

There are several potential actual cyber threats in the computing industry that can be benchmarked from this *presence* test that will fall outside the scope of this paper but will nonetheless be considered for future works and iterations of this research. This paper's in-depth referenced cyber threat focus is distributed denial of service (DDoS), as per the author's *a priori*. As such, for the constitutive research objective, the main question that this paper seeks to give illumination to is: "What are the similarities and differences of disinformation and DDoS?". This will be achieved through an in-depth analysis of the elements (i.e., threat agent, attack vector, target, impact, defense) of each phenomenon. This juxtaposition will be concluded with a recommendation based on the referenced cyber threat of this paper, DDoS.

The literary significance of this work is through its contribution to the existing research on the interdisciplinary reach of fake news (particularly disinformation), including but is not strictly limited to technology and social media policies, cybersecurity, digital forensics, the social sciences, behavioral and predictive modeling. All these involved domains are currently affected by the ramifications of fake news and both of its forms—misinformation and disinformation. An interdisciplinary approach to its phenomenon, drawing from the learnings of each of the involved fields, can be a promising blueprint of a robust policy to control these deceptions effectively. As a novel initiative, may this paper's structure

serve as a model for comparing disinformation attacks with other cyber threats.

There are multiple practical significance and implications of this paper. First, the influence it hopes to confer on social networking platforms their usage policies, specifically on the response to actors and content of disinformation. As it currently stands, there are differences in the responses of the prevailing platforms on how they deal with deliberate deception made by users. This very variation is a weakness and contributes to the growth and reach of disinformation. *A disinformation attack censored by one platform but nonetheless persists in another will still cascade throughout the whole information ecosystem.* The very nature of the information ecosystem to disseminate information, regardless of language, geographical, cultural, and demographic barriers, applies to the stratum of tech companies or the brand of the platform. A coordination in policies gives remedy to this limitation, which can be achieved through technical standardization at the computing industry-wide level.

Additionally, the epistemological provenance of disinformation attacks compared to the existing established cyber threats is relatively new. Thus, this dimension requires a thorough probe to better understand its nature. This paper will attempt to fill that gap by dissecting the anatomy of disinformation attacks, followed by a comprehensive analysis of its parts/elements.

This paper, format-wise, will (1) provide the literary constructs/background in understanding the narrative points explained in this paper, including the fundamentals of cyber threats, the anatomies of (a) disinformation attacks, (b) DDoS, (c) the literary gaps in this stream of research including a recap for the current calls for standardizations. It will be followed by the (2) discussion of the seminal methodology used in this research, the comparative threat analysis, including its controls and limitations. The latter part includes the (3) presentation of findings, (4) discussion and analysis of findings, before finally (5) highlighting the conclusion, prospectus for future research, and bestow recommendations based on the findings discovered.

II. LITERATURE REVIEW

A. Fundamentals of a Cyberthreats and Cyberattacks

Although the key terms *cyberthreat* and *cyberattack* are closely related, they are distinct from each other and are not interchangeable [4]. The former is a vulnerability that has the potential to be exploited; it is distinct from an ongoing attack that may be passive or active [5]. Due to this definition, it can be logically inferred that a cyberthreat is often a *prerequisite* of a cyberattack. This distinction in their definition is essential to keep in mind because the approach to them depends on their form; A cyberthreat is better prevented in escalating to an actual cyberattack where when there is already an ongoing cyberattack, the response is typically more aggressive to mitigate its effects and damages.

Form-wise, the Open Group (2009) defined cyberthreats as anything that is capable of causing an unwanted or harmful

consequence in computer systems [6] or in digital life in general [7]. On the other hand, a cyberattack has the presence of an *agent* [8] with the intent of exploiting the vulnerability through a particularly sophisticated method (or technically called *vector*) [9] aimed at a specific *target(s)* [10]. Due to their frequency, cyberattacks have an established protocol of response through either mitigation techniques or *defense* [10] for their counter. Most importantly, their historical occurrences have allowed the security industry to project their *impact* [10].

B. Anatomy of Disinformation Attacks

Disinformation, as an information disorder, is distinct from misinformation [3]. Disinformation is the deliberate act of creating and spreading untruthful content on social networking platforms, whereas misinformation typically happens by accident due to outdated information, miscaptioned images, and mislabeled headlines [11]. The prong that separates these two phenomena is the *intent* that comes with the act. Although these terms are commonly associated with the colloquial phrase *fake news*, they are more particular in their technical definitions. As a matter of truth, the very phrase *fake news* itself is misleading and may mean different contexts depending on the usage and applications, so it is typically not recommended to be used in more formal settings [12].

Advocates call for the classification of disinformation as an actual cybersecurity threat [4] and consequently, its more mature and dangerous form, *disinformation attacks*, should be classified as an actual cyberattack. A disinformation attack is a massive injection of deceptive propaganda in social platforms, typically state-sponsored [13]. These attacks are carefully engineered to influence and shape public opinion. Unlike mere disinformation, disinformation attacks are typically cascaded by astroturfs (bot accounts) and radical user accounts. When an entity (person or organization) is targeted, consistent untruthful narratives about the target will be bombarded across all social networking platforms.

C. Anatomy of Distributed Denial of Service (DDoS)

Denial of service (DoS) is the phenomenon characterized when a system or resource situated in a network is unable to provide its usual services and functions to other hosts. This is due to the strain induced on the resources by the gibberish requests that appear legitimate [14]. The distinctive attribute of DoS is the originating point of the exploit, as it is usually a single node in a network [15]. Although the current exploits of this kind are usually of malicious nature, it is important to note that at the technical level, DoS is a cyberthreat rather than a cyberattack because it can happen by mistake or unintentionally.

When the disruption, however, is sourced from multiple end devices and networks, it escalates to an even more serious cyberattack, distributed denial of service (DDoS) [15][16]. Compared to mere DoS, DDoS attacks are more complex and challenging to mitigate due to the involvement of botnet, end devices that are usually hijacked connected to the internet to enforce the attack for a controller [16]. The controller of these

hijacked devices is usually situated in a remote and difficultly untraceable network. Interestingly, the dual role of the hijacked device as unknowing perpetrators put them as unsuspecting victims too.

D. Importance of Threat Classification and Role of Standardization

Beyond the technical strategies in detecting and combatting these cyberthreats and cyberattacks, an equally important dimension for their holistic control is the correct applications of regulations. *Standardization* [17] allows uniformity in the approach to the remediation of all threats and attacks. Key stakeholders, when imbued with variation in their perceptions and policies, will result in inconsistencies and even outright polarization on the ethicality of a certain threat or attack. This very inherent resulting discord at the minimum will make the threat persist or may even amplify the effects of any threat or attack. A formal classification and elevation of a phenomenon to a threat is the first move in recognizing, industry-wide, that it is unacceptable to all the involved entities.

More importantly, *threat classification* [18] allows all the fields involved to contribute to solving the complexities of these threats and attacks, not just in a multidisciplinary but also in an interdisciplinary format. Finally, threat classification allows a richer context and more thorough understanding of these through the accounting of existing threat architecture, which may have directly or indirectly contributed to the emergence of the newer ones, which are typically more sophisticated. By tracing its origins, the possibility of eradicating them, not just at the surface but in their roots, can be achievable.

E. The Literary Gap: Comparison with an Actual Cyberattack

With these underpinnings, the seminal corollary is evident: That the existing calls for the classifications of disinformation and disinformation attack as an official cyberthreat and cyberattack, respectively, are not uncalled for. In their forms, they mimic the typical characteristics for their appropriate considerations. However, tangible evidence of such is missing in the existing literary works. There are multiple threat classification models that may be used to meet; however, a simple yet explicitly robust method of assessment is surprisingly virtually missing in these classificatory proposals, *a direct comparison of the proposed phenomenon with an actual cyberthreat/attack*. This paper fills this gap by outright doing so with an actual comparative vector analysis of cyberattack DDoS and disinformation attacks as the phenomenon in question. If they closely resemble each other, the affirmation it confers to the existing proposals will strengthen their claims. If not, a new benchmarked threat can be referenced, or the reasons for rejection will finally come into the picture. Either way, this will make the stream on this research domain move forward, a hope this paper aims to achieve.

III. METHODOLOGY

A. Summary, Overview, and Format

This paper explicitly compares two phenomena, DDoS and disinformation attacks. The former is an actual cyberattack as recognized in authoritative standards and manuals, while the latter is a candidate cyberattack speculated to contain the key elements congruent with the prevailing cyberthreats and cyberattacks.

On one end, DDoS and its substantial elements will function as the benchmark. On the other, disinformation attacks will be carefully analyzed and thoroughly probed if it meets the substantial *presence* (or absence) test per each threat/attack element. The substantial elements are exhaustively and systematically compiled by [19] in their work and will be each individually discussed in the following subsections. The number of elements present is directly proportional to the strength of indication that disinformation attacks should be classified as actual cyberattacks.

Cyberattack Component	Indicator(s) of Presence / Absence
Agent	A cyberthreat/vulnerability triggered by an <i>entity</i> (user, organization, or nation)
Vector	An established <i>pathway</i> of the payload from the agent to the target/s
Target	Documented past and current <i>impairments</i> in at least one layer of the OSI model
Impact	Evidence of the ability to affect systems, users, and organizations
Defense	Documented preventative controls and remediation strategies

Table 1: Summary of the Prong Applied to the Candidate Phenomenon

B. Agent

The agent of an attack is the entity that exploits a particular threat and triggers it to initiate the vector sequence. As it currently stands, an agent is always a human source. Institution and state-sponsored agents also fulfill this requirement. Albeit their sophistication, mechanical enforcers like bots and zombie computers are mere helpers of the agent. To further clarify this distinction, two prongs that are used to examine if an entity can be classified as the agent are (a) motivation and (b) goals or objectives. Enforcers will fail at these two prongs and, as such, are not considered agents but mere tools of amplification. The malice in motivation and goals is insignificant since intrusion tests meet the criteria of attack classification, albeit the objective of testing the cybersecurity posture of a subject. Tracing the agent is one of the most important (if not most) substantial elements of threat/attacks since most of the accountability is attributed to the sourcing agent. Existing laws and regulations, to the best of their prowess, usually prosecutes the agent more than the enforcers, should they happen to be human too.

C. Attack Vector

The vector of an attack, in essence, is the method of delivering the payload to the intended targets. It is the totality of the steps undertaken by the agent alongside the tools and strategies to exploit an attack surface. The vulnerability, at this point, has been used, and the transition of a cyberthreat to a

full-blown cyberattack has taken place. It is important to note, however, that cyberattacks have the distinguishing characteristic of the established pathway (i.e., the *vector*) connecting the sourcing agent to the target(s) discussed in the next subsection. Both active and passive cyberattacks meet this requirement. After an attack vector has been initiated and established by an agent, the payload will then flow to the target(s) internally or externally. The determination of the direction of the payload flow, as directed by the structure of the vector, is of utmost importance to the response of attack as it'll dictate the appropriate structure of mitigation techniques that will be used to minimize the effects of the attack; These are further discussed in the last subsection of this chapter.

D. Target

In the context of computer systems, the target(s) of cyberattacks are technically NOT directly humans or users. They are often directly and indirectly affected, but the attacks technically impair component(s) of the OSI layer model. For instance, malware and ransomware may make data unusable, rendering users unable to make something out of it; a simple denial of service attack may hinder users in accessing a resource but in all technicalities, what is broken is the access and authorization components of an application or the network hosting it.

Unlike a physical hazard that physically puts a human in danger, the commonality of cyberattacks is their natures are all concentrated in the cyberworld. This is not to say that humans will not be affected as they will certainly be. The prong of a candidate attack will be examined if it causes mostly a physical or cyber effect. It has to be primarily concentrated on the latter and affecting the users thereafter directly or indirectly.

E. Impact

The impact is the dimension where the humans or users come in. However, users are not exclusively the only impacted components of cyberattacks; for instance, they can be whole systems or even organizations (nations, vulnerable groups). Impacts are usually measured on several metrics, including but are not strictly limited to: (a) the gravity of their disruption scope and length, (b) novelty or precedence, (c) economic value, and even (d) psychological effects.

The *presence* test applied on this paper is a dichotomous test as it checks only the binary value (absence or presence) of any of these stipulated forms of impact to answer the research question of this paper. It is essential to note, however, that further quantification of these impacts is possible and is performed in many research streams and fields. This sub-area is outside the scope of this paper and one of this paper's limitations.

F. Defense

The defense mechanisms against a cyberattack fall under two categories, mitigation and remediation. Mitigation refers to the procedures employed *prior* to an attack as preventative measures and *during* an actual attack to limit the damages produced after its deployment. Forms of mitigation

procedures are often detection technologies that alert stakeholders of a possible attack, such as intrusions in a perimeter-based on a given set of indicators.

Remediation, on the other hand, is the act of correction *after* an attack has taken for the purpose of closing the loops that allowed it to happen and persist in the first place. Remediation strategies are equally as important as mitigation. It can include policy changes, hardening the skill set of users involved, and more. Types of defense mechanisms vary and are often used in conjunction with each other. For this paper's test, an examination of the current preventative and remediation if it currently exists is the success-and-failure metric for this component.

IV. RESULTS

A. Overview

The table below is the comparative summary assessment of DDoS against the candidate phenomenon, disinformation attacks. The values per element are dichotomous and non-scaled. Further discussions and explanations on the important points are included in the next chapter to interpret and make sense out of the findings presented below.

	DDoS	Disinformation Attacks
Agent	Cybercriminals, State-sponsored Organizations, Vendors (<i>Stresser Services</i>)	Radical Extremists, State-sponsored Organizations, Vendors (Ads)
Vector	Invalid authentication requests amplified by Zombie computers, TCP/IP Protocols as the attack surface (DNS, UDP, NTP, etc.)	Content Duplicators and Spreaders primarily through and amplified by astroturfs and bot accounts, appearing in any legitimate format (text, images, audio, video)
Target	Networks, Websites, Systems (technically OSI Layers 7, 6, 5, 4, 3)	News Headlines Manipulation, Article Fabrication, Content Miscalcaptioning, Deepfakes, Misleading Graphics, and Incomplete Videographic Creation (technically & exclusively OSI Layer 7)
Impact	Operational Disruption, Financial and Productivity Loss, Brand and Reputational Damages	Reputational Damages, Financial Loss (through Stocks and Securities Depreciation), Social Polarization
Defense	Web Application Firewalls, Intrusion Prevention Systems (IPS), Intrusion Detection Systems (IDS), Access Control Lists (ACL), Load Balancers (as backup)	Fake News Detection Technologies (mainly powered by Machine Learning models), Censorship on Agents and Secondary Actors, Content Demotion and Appropriate Labelling

V. ANALYSIS & DISCUSSION

Discussion 1: There is a Thin Line Between Radical Extremists and Cybercriminals

An in-depth look at the comparative table reveals that the usual agents for both disinformation attacks and DDoS are almost the same except for radical extremists. Agents of DDoS are, as per the existing US laws, considered cybercriminals since DDoS is officially sanctioned as an illegal activity. Radical extremists and their acts, on the other hand, are currently protected by the First Amendment (or Free Speech). Nonetheless, the usual motivations and deliberate intention of a threat agent are visibly present in disinformation attacks.

Discussion 2: The Vector of Disinformation Attacks Closely Resembles that of DDoS's

The vector structures, styles, and patterns of both DDoS and disinformation attacks are essentially the same. They both rely on amplification to undertake their respective payloads. To contribute to its acts of disruptions, DDoS actors harness the zombie computers. These computers and their respective users are often unaware of their inclusion. Furthermore, zombie computers are often situated in foreign countries, remote to both the agent and the targeted networks, rendering them elusive to trace. Disinformation attacks, on the other hand, harnesses troll farms and automatically generated *astroturfs* to successively inject social networking platforms with deceptive information. To add to these attacks, uninformed users [20-27] may further unknowingly share these to their feeds, reverberating the falsehood attacks mirroring a zombie computer's behavior.

Discussion 3: Unlike DDoS, Disinformation Attacks Targets Exclusively the Application Layer

The most visible difference between DDoS and disinformation attacks is the various OSI layers affected and reached by the former. The attack surface of DDoS is wider due to the mechanism of networking models and protocols to assure reliability. This may be due to acknowledgment requests and message trace bounce backs. Disinformation attacks, on the other hand, exclusively pollute the application layer through the content at the social networking platforms. Although the formats of deception highly vary, including text, image, audio, and video graphics—there is no attempt at intrusion at the networked elements and protocols. Agents and their enforcers carefully follow the rules and policies set by the social networking platforms.

Discussion 4: The Impacts of Disinformation Attacks are Congruent with that of DDoS's

Additionally, the impacts of both DDoS and disinformation attacks are both operating at a systemic and networked level. DDoS aims to disrupt resources and deny users of access; in the CIA triad of cybersecurity, this is the last abbreviation, **Availability**. While disinformation attacks, on the other hand, pollute and reduce the overall quality of information lurking around social networking platforms, making the users doubt content legitimacy, even the truthful

ones, which exactly represents the I in the CIA triad of cybersecurity, **Integrity**. Although they target different aspects of protocol stacks, they both ultimately bring dangers and inconvenience to society and users alike.

Discussion 5: Defense for Both DDoS and Disinformation Attacks are Primarily Mitigative rather than Preventative

As it currently stands, the prevailing responses and measures against both DDoS and disinformation attacks are mostly deployed after they have occurred or activated only at the time of their discovery. While it makes sense due to their design, advanced predictive technologies that may forecast when phenomena of such nature will most likely arrive may be needed to fully protect users from their effects. While the current detection and prevention systems are not completely impregnable, they are the essential controls that are put in place for basic defenses. Finally, increasing user awareness [28] and modification of policies after an attack is a reasonable practice currently implemented for both DDoS and disinformation attacks.

VI. CONCLUSION AND FUTURE WORKS

This paper examined the phenomenon of disinformation attacks through a comparative analysis with distributed denial-of-service (DDoS) attacks. The analysis revealed that the two are strongly similar in terms of all elemental criteria. Importantly, the vectors, amplification methods, styles, and structures of their forms are indistinguishable. The only trivial difference is the payload each injects, whereas DDoS mainly massively disrupts operations, disinformation attack massively deceives users.

With these compelling findings, this paper concurs and highly recommends that disinformation be classified as an actual cybersecurity threat. Its more advanced form, disinformation attacks, exhibits the presence of the notable elements that of an actual cyberattack. This warranted classification will compel common remediation approaches and instill unified philosophies across all social networking platforms against this phenomenon. Without such classification, the current architecture of social networks will allow the platforms to have their own policies in limiting misinformation and disinformation, yielding inconsistencies. This renders the war against *fake news* difficult to triumph over.

This paper, as a stream of research, introduced a novel approach to assess how a candidate phenomenon can be tested of its eligibility to a formal classification of a cybersecurity threat. By using an existing cyber threat as a benchmark and individually comparing their properties through an absence/presence test, their congruence can be examined. Future iterations of this study will use other cyber threats as benchmarks to further strengthen the claims and validity of the recommendation conferred by this paper. Projected benchmark cyber threats include advanced persistent threats (APTs), phishing, and adware.

REFERENCES

- [1] Fallis, D. (2015). What is disinformation?. *Library trends*, 63(3), 401-426.
- [2] Douligieris, C., & Mitrokotsa, A. (2004). DDoS attacks and defense mechanisms: classification and state-of-the-art. *Computer networks*, 44(5), 643-666.
- [3] Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe, 27.
- [4] Caramancion, K. M. (2020, March). An exploration of disinformation as a cybersecurity threat. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 440-444). IEEE.
- [5] Shirey, R. W. (2000). IETF (Internet Engineering Task Force) RFC 2828. Internet Security Glossary.
- [6] The Open Group, 2009. Technical Standard. Risk Taxonomy. ISBN: 1-931624-77-1. Document Number: C081.
- [7] Taylor, H. 2021. What Are Cyber Threats and What to Do About Them. Introduction to Cybersecurity. Retrieved from <https://preyproject.com/blog/en/what-are-cyber-threats-how-they-affect-you-what-to-do-about-them/>
- [8] Casey, T., Koeberl, P., & Vishik, C. (2010, April). Threat agents: A necessary component of threat analysis. In *Proceedings of the sixth annual workshop on cyber security and information intelligence research* (pp. 1-4).
- [9] Irmak, E., & Erkek, İ. (2018, March). An overview of cyber-attack vectors on SCADA systems. In *2018 6th international symposium on digital forensic and security (ISDFS)* (pp. 1-5). IEEE.
- [10] Poehlmann, N., Caramancion, K. M., Tatar, I., Li, Y., Barati, M., & Merz, T. (2021). The Organizational Cybersecurity Success Factors: An Exhaustive Literature Review. *Advances in Security, Networks, and Internet of Things*, 377-395.
- [11] Caramancion, K. M. (2021, November). The Role of Information Organization and Knowledge Structuring in Combatting Misinformation: A Literary Analysis. In *International Conference on Computational Data and Social Networks* (pp. 319-329). Springer, Cham.
- [12] Habgood-Coote, J. (2019). Stop talking about fake news!. *Inquiry*, 62(9-10), 1033-1065.
- [13] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019, May). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference* (pp. 218-226).
- [14] Gu, Q., & Liu, P. (2007). Denial of service attacks. *Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, 3, 454-468.
- [15] Gasti, P., Tsudik, G., Uzun, E., & Zhang, L. (2013, July). DoS and DDoS in named data networking. In *2013 22nd International Conference on Computer Communication and Networks (ICCCN)* (pp. 1-7). IEEE.
- [16] Prasad, K. M., Reddy, A. R. M., & Rao, K. V. (2014). DoS and DDoS attacks: defense, detection and traceback mechanisms-a survey. *Global Journal of Computer Science and Technology*.
- [17] Xie, Z., Hall, J., McCarthy, I. P., Skitmore, M., & Shen, L. (2016). Standardization efforts: The relationship between knowledge dimensions, search processes and innovation outcomes. *Technovation*, 48, 69-78.
- [18] Jouini, M., Rabai, L. B. A., & Aissa, A. B. (2014). Classification of security threats in information systems. *Procedia Computer Science*, 32, 489-496.
- [19] Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. *Data*, 7(4), 49.
- [20] Caramancion, K. M. (2021, April). The demographic profile most at risk of being disinformation. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-7). IEEE.
- [21] Caramancion, K. M. (2021, May). Understanding the association of personal outlook in free speech regulation and the risk of being mis/disinformed. In *2021 IEEE World AI IoT Congress (AlloT)* (pp. 0092-0097). IEEE.
- [22] Caramancion, K. M. (2020, September). Understanding the impact of contextual clues in misinformation detection. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE.
- [23] Caramancion, K. M. (2022, January). The Role of User's Native Language in Mis/Disinformation Detection: The Case of English. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 260-265). IEEE.
- [24] Caramancion, K. M. (2021, October). Textual vs. Visual Fake News: A Deception Showdown. In *2021 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (pp. 31-35). IEEE.
- [25] Caramancion, K. M. (2021, December). The Relation of Online Behavioral Response to Fake News Exposure and Detection Accuracy. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 0097-0102). IEEE.
- [26] Caramancion, K. M. (2021, October). The Role of Subject Confidence and Historical Deception in Mis/Disinformation Vulnerability. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0541-0546). IEEE.
- [27] Caramancion, K. M. (2021, September). The Relation Between Time of the Day and Misinformation Vulnerability: A Multivariate Approach. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)* (Vol. 1, pp. 150-153). IEEE.
- [28] Caramancion, K. M. (2022). An Interdisciplinary Assessment of the Prophylactic Educational Treatments to Misinformation and Disinformation (Doctoral dissertation, State University of New York at Albany).

MATHEMATICAL MODELING OF CREDIT SCORING SYSTEM BASED ON THE MONGE-KANTOROVICH PROBLEM

1st Gayrat Juraev

dept. of Information Security
National University of Uzbekistan
named after Mirzo Ulugbek
Tashkent, Uzbekistan
gujurayev@gmail.com

2nd Kuvonchbek Rakhimberdiev

dept. of Information Security
National University of Uzbekistan
named after Mirzo Ulugbek
Tashkent, Uzbekistan
qquuvvoonn94@gmail.com

Abstract—In this article the issues of further improvement of the credit system in order to ensure the implementation of the Decree of the President of the Republic of Uzbekistan dated May 12, 2020 No 5992 "On the Strategy of Banking Reform in the Republic of Uzbekistan for 2020-2025" studied. Problems of information security in the credit system and their solutions, issues of secure storage of personal data of borrowers through blockchain technology, as well as effective modeling of lending and borrower relations, lending institutions on the basis of established criteria recalculation model and quantitative analysis of the assessment, the application of the Monge-Kantorovich problem in the development of the credit scoring model, the process of forming a credit scoring model based on the model applied to the Monge-Kantorovich transport problem. The credit scoring model based on the Monge-Kantorovich problem is similar to the Bayesian network-based credit scoring model, and calculations are performed using elements of probability theory. Borrowers were evaluated using optimal evaluation criteria. As a result, the process of deciding whether or not to grant a loan to a borrower has been completed. The results obtained were analyzed and a credit scoring model based on the Monge-Kantorovich problem was proposed.

Index Terms—bank, lending, borrower, blockchain, credit scoring, transport task, rating, minimization, maximization, client, empirical, information security, confidentiality, lending, borrower

I. LITERATURE REVIEW

The banking system is one of the most important sectors of the country's economy. Today, in many developed countries, much attention is paid to the introduction of modern banking services, efficient customer service and the implementation of secure lending. The lending process is one of the main departments of a bank or other lending institution, and the implementation of secure lending increases the efficiency of the organization. Secure lending process includes proper assessment of borrowers, prevention of threats to the customer's customer data, ie protection against such actions as cyber attacks, physical destruction of data, violation of confiden-

tiality. In many developed countries, blockchain technology is widely used to secure modern banking services. These include automatic payments, P2P payment systems, and P2P lending systems based on blockchain technology [4], [5], [6].

The financial security of a lending institution depends on the correct assessment of borrowers and the risk of default. The higher the number of reliable borrowers in the lending process, the lower the risk of non-repayment. Therefore, the issue of effective evaluation of borrowers is a topical issue. Borrower evaluation models are generally referred to as credit scoring. Currently, economists, mathematicians and cybernetic scientists are conducting research on mathematical modeling of the credit scoring process and the development of effective scoring models [7]. Credit scoring models are based on machine learning and statistical analysis methods.

Several credit scoring models have been developed based on machine learning methods. There are currently many publications devoted to the study of machine learning models. In particular, evaluation models based on machine learning appeared in 1941 in the scientific work of David Durand. In David Durand's scientific work, the borrowers are divided into groups such as good or bad borrowers, which are evaluated by statistical methods in the allocation of loans to enterprises, firms, organizations of the banking organization [3]. Borrower assessment models were improved in the second half of the 20th century, and linear probability models and discriminatory methods based on statistical methods began to be used to solve credit scoring problems [2]. Crook, J. N., Edelman, D. B. and Thomas, L.C. In his works, models of assessment of borrowers by the method of logistic regression are presented [8].

Alternatively, B. Baesens et al. In a 2003 article, Benchmarking state-of-the-art classification algorithms for credit, looks at support vector machines, neural network, decision trees, k-nearest neighbor, linear programming, and Bayesian network classifier [9], [10]. This paper aims to mathematically model a credit scoring system based on the problem of machine learning by analyzing the above methods. We propose a system based on the Monge-Kantorovich problem in modeling

a credit scoring system [11], [12].

II. METHODOLOGY

Suppose there is a certain bank engaged in lending to individuals. Customers apply to banks for loans. The decision to issue a loan is made by the bank based on information about the client. The bank receives information about the client from various sources: from the client himself, from the credit bureau and from other sources [28], [29]. We will consider the information provided by the client himself. The bank receives it through a questionnaire filled out by the borrower. In the questionnaire, the borrower indicates the following data: gender, age, marital status, presence of children, monthly income, availability of real estate, etc. Based on these data, we will divide customers into groups in which they are similar in certain respects. [30], [31]. For each client, using the Bayesian method, we will find a rating - the empirical probability that the client will repay the loan, provided that he belongs to this group. Having found the distribution of ratings, we thereby construct a scoring model [13], [14], [15].

To apply the method, the data must satisfy the following conditions:

- independence clients do not collude to repay the loan;
- homogeneity data are taken from the same general population;
- equiprobability clients are equally likely to be distributed into groups.

The Kantorovich problem is linear in contrast to the complex Monge problem: it is required to find the minimum of a linear functional on a convex set. If the Monge problem has a solution, then the Kantorovich problem has a solution. In some special cases, it is possible to construct a solution to the Monge problem by solving the Kantorovich problem, but this is quite rare [16], [17], [18].

However, there is a close relationship between the Monge and Kantorovich problems. In this regard, the term Monge-Kantorovich problem has become generally accepted. Linear programming begins with the Kantorovich problem [19], [20], [21]. The problem of finding the distribution of ratings comes down to finding the joint distribution of the probabilities that the client will repay the loan and the probabilities that he will fall into a specific group [22].

Thus, we have reduced the problem of finding a joint distribution to a transport problem solved by linear programming methods. However, in this case, the problem can be solved more elegantly. In this case, we immediately proceed to finding the ratings. Consider a discrete version of this problem. It is called the transport problem of linear programming. The resulting problem is typical for linear programming. It is solved by the simplex method of linear programming. We will not give an algorithm for solving it here, but note that there is a procedure in the MATLAB environment that finds a solution to the transport problem [37].

III. INTRODUCTION

It is known that the improvement of information systems around the world, the widespread use of digital technologies

and the development and improvement of methods and algorithms for information protection are important. In particular, the consistent penetration of information technology in many areas in the Republic of Uzbekistan contributes to the growth of the country's economy.

The growth of business entities in a market economy requires the improvement of financial relations between them. Therefore, the country is undergoing many reforms to develop the banking sector, which is one of the most important sectors of the economy. To this end, the Government of Uzbekistan has adopted a number of decrees and resolutions. In particular, great attention is paid to the implementation of the Decree of the President of the Republic of Uzbekistan dated May 12, 2020 No 5992 On the strategy of reforming the banking system of the Republic of Uzbekistan for 2020-2025.

The role of credit systems in the development of business entities is undoubtedly of great importance. However, at the same time, the imperfection of the existing lending systems in the Republic of Uzbekistan leads to some conflicts. In this situation, the issues of improving the country's credit system are of great practical importance.

The lending system is one of the most important links of a bank or other lending institution. The implementation of effective lending processes is the basis for the effective operation of lending institutions [32], [33], [34]. Effective lending is the maximum benefit of the lending institution's lending process and ensuring the confidentiality of borrowers' information. By reducing the risk of non-repayment of loans in lending institutions or by providing loans to reliable borrowers, banking institutions can achieve maximum profit. The confidentiality of customers' financial information is ensured using modern information security mechanisms [35], [36].

One of the most effective ways to eliminate defaults in a bank or other lending institution is to assess the creditworthiness of borrowers. There are classic and machine-based methods of evaluating borrowers.

Classic methods credit scoring models

- Models of scoring (rating) evaluation
- Statistical assessment of the borrower
- Altman Model
- Chesser Model

Credit scoring models based on machine learning

- Linear regression
- Logistic regression
- Discriminant Analysis
- Decision Trees
- Support Vector Machine
- Bayesian networks
- Neural networks
- Genetic algorithms
- Combined methods
- Methods based on fuzzy logic

Today, machine-based methods that are more effective than the classic scoring methods listed above are widely used. In this paper, we propose a credit scoring mole based on the Monge-Kantorovich problem. This model is similar to

the Bayesian-based credit scoring model and is based on an intellectual analysis of data from a bank or other lending institution. The Monge-Kantorovich problem is presented as follows.

IV. MONGE-KANTOROVICH PROBLEM

Monge's problem. Given two probability spaces (X, A, μ) and (Y, B, ν) and a non-negative measurable function h on $X \times Y$, called the cost function. It is required to find a mapping $T : X \rightarrow Y$ that is measurable with respect to the pair (A, B) , etc., $T^{-1}(B) \in A$ for all $B \in \mathcal{B}$, that takes the measure μ to ν and delivers a minimum to the expression

$$M_h(\mu, T) := \int_X h(x, T(x))\mu(dx) \tag{1}$$

among all such mappings. The condition that μ goes into ν means that it ν coincides with the image of the measure μ when displayed T , which is given by the formula $\mu \circ T^{-1}(B) := \mu(T^{-1}(B)), B \in \mathcal{B}$. If some $T \in T(\mu, \nu)$ gives a minimum, then this is T called the optimal mapping of the measure μ or ν optimal transportation [1].

In the interpretation of Monge himself, it was about the transfer of soil for construction work, i.e. both measures were the usual volumes, the function was the usual distance, and the work done was minimized. Outwardly, the task looks like an applied one, but it is not. This formulation is not the only economic formulation of the transport task. A discrete version of the Monge and Kantorovich problems arises in many modern fields.

Kantorovich problem. In 1942 major L.V. Kantorovich submits his note for publication, in which a problem is posed that is close to the Monge problem, but with one fundamentally important nuance: in the Kantorovich problem, instead of searching for the mapping T (optimal transportation), it is proposed to find only the optimal transportation plan, i.e. such a probability measure σ on $(X \times Y, A \otimes B)$, that its projection on X and Y are μ and ν , respectively, and σ delivers a minimum to the expression

$$K_h(\sigma) := \int_{X \times Y} h(x, y)\sigma(dxdy) \tag{2}$$

over all probability measures σ from the class of (μ, ν) probability measures on the product $(X \times Y, A \otimes B)$, giving μ and ν when projecting on X and Y .

The Kantorovich problem is linear, in contrast to the complex Monge problem: it is required to find the minimum of a linear functional on a convex set. If the Monge problem has a solution, then the Kantorovich problem has a solution. In some special cases, it is possible to construct a solution to the Monge problem by solving the Kantorovich problem, but this is quite a rarity. [23]

However, there is a close relationship between the Monge and Kantorovich problems. In this regard, the term Monge-Kantorovich problem has become generally accepted. Linear programming begins with the Kantorovich problem. Consider

a discrete version of this problem. It is called the transport problem of linear programming. Let's formulate it [24].

Departure points given A_1, \dots, A_m in each of which there are a_1, \dots, a_m units of cargo, and destinations B_1, \dots, B_n with the needs b_1, \dots, b_n of units of cargo. The cost of transporting a unit of cargo is known $c_{i,j}$ from A_i to B_j . You need to make a transportation plan. In this case, the total need for cargo coincides with the total amount of cargo at the points of departure, etc.

$$\sum_{i=1}^m a_i = \sum_{j=1}^n b_j \tag{3}$$

Let us construct a mathematical model of this problem. Need to find x_{ij} the amount of cargo transported from the point of origin A_i to the destination B_j . Obviously, the following conditions must be met:

$$\sum_{j=1}^n x_{ij} = a_i, i = 1, \dots, m \tag{4}$$

$$\sum_{i=1}^m x_{ij} = b_j, j = 1, \dots, n \tag{5}$$

$$x_{ij} \geq 0 \tag{6}$$

In this case, the total cost of transportation should be minimized.

$$\sum_{j=1}^n \sum_{i=1}^m c_{ij}x_{ij} \tag{7}$$

The resulting problem is typical for linear programming. It is solved by the simplex method of linear programming. We will not give an algorithm for solving it here, but note that there is a procedure in the MATLAB environment that finds a solution to the transport problem.

V. BUILDING A CREDIT SCOURING MODEL BASED ON THE MONGE-KANTOROVICH PROBLEM

The content of the issue is similar to the Bayes problem, and the following conditions must be met:

- independence - clients do not collude to repay or not repay a loan;
- homogeneity - data are taken from one general population;
- equiprobability - customers are equally likely to be distributed into groups.

The task of finding the distribution of ratings comes down to finding the joint distribution of the probabilities that the client will repay the loan and the probabilities that he will fall into a specific group. It boils down as follows: knowing the joint distribution and the probability of getting into each group $P(X_{i,j}, Y_j)$ and $P(Y_j)$, we can find the rating

$$P(X_{i,j}|Y_j) = \frac{P(X_{i,j}, Y_j)}{P(Y_j)} \tag{8}$$

Let us reduce the problem of finding a joint distribution to a transport problem. denote $p_{i,j} = P(X_{i,j}, Y_j)$. At the same time, the marginal probability distribution is $P(Y_j)$ defined in Chapter 2. Let us indicate the marginal distribution $P(X_i)$ calculated as the ratio of the number of customers who repaid (not repaid) the loan to the total number of customers [23], [24].

We will build it as the ratio of the number of clients satisfying a pair of values of random variables (X, Y) to the total number of clients. Fix the number of clients corresponding to each possible pair (X, Y) .

TABLE I
THE NUMBER OF CLIENTS WHO REPAID AND DID NOT REPAY THE LOAN IN EACH GROUP

	X_i	Y_i		X_i	Y_i		X_i	Y_i		X_i	Y_i
No	0	1	No	0	1	No	0	1	No	0	1
1	20	12	16	19	5	31	10	5	46	49	14
2	45	22	17	34	25	32	21	17	47	69	27
3	44	20	18	45	15	33	20	13	48	58	18
4	21	10	19	21	4	34	16	3	49	27	9
5	15	10	20	11	3	35	8	4	50	20	3
6	12	3	21	18	6	36	11	6	51	5	2
7	23	17	22	36	11	37	27	17	52	12	3
8	31	17	23	47	11	38	27	10	53	2	2
9	15	8	24	19	6	39	8	4	54	3	2
10	4	3	25	5	3	40	6	5	55	2	2
11	18	12	26	20	5	41	49	12	56	2	1
12	28	23	27	49	10	42	65	32	57	6	2
13	35	13	28	52	16	43	65	15	58	6	4
14	14	5	29	30	4	44	28	9	59	4	0
15	12	6	30	28	10	45	14	2	60	1	1

$P(X_1) = 0.716742$, $P(X_2) = 0.283257$. X_1 - the number of those who repaid the loan, X_2 - the number of those who did not repay the loan [36], [37].

It is required to find a joint distribution $p_{i,j}$ such that: Minimizes the linear shape,

$$\sum_{i=1}^2 \sum_{j=1}^{60} |X_{i,j} - Y_j| p_{i,j} \tag{9}$$

$$0 \leq p_{i,j} \leq 1 \tag{10}$$

$$\sum_{i=1}^2 \sum_{j=1}^{60} p_{i,j} = 1 \tag{11}$$

$$\sum_{i=1}^2 p_{i,j} = P(Y_j), \quad j = 1, \dots, 60 \tag{12}$$

$$\sum_{j=1}^{60} p_{i,j} = P(X_i), \quad i = 1, 2 \tag{13}$$

As a cost function, $h(x, y)$ we choose the modulus of the difference between $X_{i,j}$ and Y_j : $h(X_{i,j}, Y_j) = |X_{i,j} - Y_j|$.

Thus, we have reduced the problem of finding a joint distribution to a transport problem solved by linear programming

methods. However, in this case, the problem can be solved more elegantly. In this case, we immediately proceed to finding the ratings [27], [28].

Consider the 1st condition for $p_{i,j}$. Since it is necessary to minimize the sum of non-negative elements, we can proceed to minimizing the sum over i , while fixing the specific j :

$$\sum_{i=1}^2 |X_{i,j} - Y_j| p_{i,j} \rightarrow \min \tag{14}$$

Using the fact that $P(X_{i,j}|Y_j) = \frac{p_{i,j}}{P(Y_j)}$, let's move on to

$$\sum_{i=1}^2 |X_{i,j} - Y_j| P(X_{i,j}|Y_j) \rightarrow \min \tag{15}$$

What is equivalent to minimizing

$$\sum_{i=1}^2 \left| 1 - \frac{X_{i,j}}{Y_j} \right| P(X_{i,j}|Y_j) \tag{16}$$

From minimization, we can make an equivalent transition to maximization:

$$\sum_{i=1}^2 \left| 1 - \frac{X_{i,j}}{Y_j} \right| (1 - P(X_{i,j}|Y_j)) \rightarrow \max \tag{17}$$

Denote

$$\tilde{P}(X_{i,j}|Y_j) = 1 - P(X_{i,j}|Y_j) \tag{18}$$

$$\tilde{c}_{i,j} = 1 - \frac{X_{i,j}}{Y_j} \tag{19}$$

$$\tilde{P} = \begin{pmatrix} \tilde{P}(X_{1,j}|j) \\ \tilde{P}(X_{2,j}|j) \end{pmatrix} \tag{20}$$

$$\tilde{c} = \begin{pmatrix} \tilde{c}_{1,j} \\ \tilde{c}_{2,j} \end{pmatrix} \tag{21}$$

Thus, the problem is reduced to maximizing the scalar product $\langle \tilde{P}, \tilde{c} \rangle$.

We use the Cauchy - Bunyakovsky inequality, which says that $\langle \tilde{P}, \tilde{c} \rangle \leq \sqrt{\langle \tilde{P}, \tilde{P} \rangle} \sqrt{\langle \tilde{c}, \tilde{c} \rangle}$. Equality is achieved when the vectors are collinear. Hence we get that $\tilde{P}(X_{i,j}|Y_j) = \tilde{c}_{i,j}$.

From here we obtain the solution of our problem:

$$P(X_{i,j}|Y_j) = \frac{X_{i,j}}{Y_j} \tag{22}$$

Thus, the transport problem was solved clearly. We have received customer ratings.

Table 2 calculates the rating scores for 60 pairs of clients. Based on these rating points, decisions are made based on the criteria in Table 3 for borrowers.

Based on the ratings, we can classify clients according to the empirical probability of their repayment of the loan.

Based on the evaluation criteria outlined in Table 3, the decision to grant a loan to a borrower is made as follows.

TABLE II
CUSTOMER RATINGS

	Rating score		Rating score		Rating score
1	0.625	21	0.75	41	0.803279
2	0.671642	22	0.765957	42	0.670103
3	0.6875	23	0.810345	43	0.8125
4	0.677419	24	0.76	44	0.756757
5	0.6	25	0.625	45	0.875
6	0.8	26	0.8	46	0.777778
7	0.575	27	0.830508	47	0.71875
8	0.645833	28	0.764706	48	0.763158
9	0.652174	29	0.882353	49	0.75
10	0.571429	31	0.736842	50	0.869565
11	0.6	31	0.666667	51	0.714286
12	0.54902	32	0.552632	52	0.8
13	0.729167	33	0.606061	53	0.7
14	0.736842	34	0.842105	54	0.6
15	0.666667	35	0.666667	55	0.4925525
16	0.791667	36	0.647059	56	0.666667
17	0.576271	37	0.613636	57	0.75
18	0.75	38	0.72973	58	0.6
19	0.84	39	0.666667	59	1
20	0.785714	40	0.545455	60	0.4535365



Fig. 1. Ratings graph

- 1) The client's rating lies in the half-interval $(0.7, 1]$. The client is considered reliable. You can give him a loan.
- 2) The client's rating lies in the half-interval $(0.5, 0.7]$. The client is considered medium-risk. If the credit conditions are revised, then this client can be granted a loan.
- 3) Customer Rating The customer $(0, 0.5]$ is considered risky. Such a client should not be given a loan.

Based on the above evaluation criteria, we classify the assessments of borrowers in Table 2 by reliability levels. The indicators of the classified estimates are shown in the following tables.

TABLE III
CRITERIA FOR EVALUATING BORROWERS

Evaluation criteria	The level of confidence of borrowers
High level	$(0.7, 1]$
Medium level	$(0.5, 0.7]$
High level	$(0, 0.5]$

TABLE IV
BORROWER GROUPS WITH LOW RELIABILITY

Groups	Low reliability indicators
55	0.4925525
60	0.4525265

TABLE V
BORROWER GROUPS WITH MEDIUM RELIABILITY

Groups	Medium reliability indicators
1	0.625
2	0.671642
3	0.6875
4	0.677419
5	0.6
7	0.575
8	0.645833
9	0.652174
10	0.571429
11	0.6
12	0.54902
15	0.666667
17	0.576271
25	0.625
31	0.666667
32	0.552632
33	0.606061
35	0.666667
36	0.647059
37	0.613636
39	0.666667
40	0.545455
42	0.670103
54	0.6
56	0.666667
58	0.6

VI. RESULTS

Thus, we analyze the data tables classified above. The tables are divided into 60 groups, and calculations were performed using 1997 customer data.

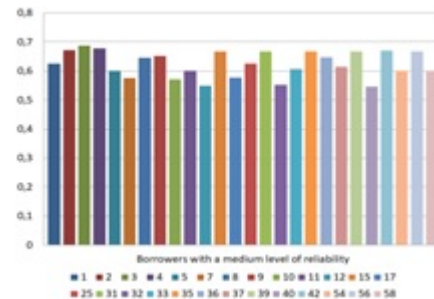


Fig. 2. Borrowers with a medium level of reliability

As a result, in Table 4, groups 55 \rightarrow 0.4925525 and 60 \rightarrow 0.4525265 were assessed, of which 55 and 60 customer groups are **low levels** of reliability borrowers. These borrowers are considered insolvent and will not be given credit. According to the results given in Table 5, 5, 7-12, 15, 17, 25, 31, 32, 33, 35-37, 39, 40, 42, 54,56, 58 groups of borrowers have

TABLE VI
BORROWER GROUPS WITH HIGH RELIABILITY

Groups	High reliability indicators
21	0.75
22	0.765957
23	0.810345
24	0.76
45	0.875
26	0.8
27	0.830508
28	0.764706
29	0.882353
30	0.736842
51	0.714286
52	0.8
13	0.729167
14	0.736842
55	0.4925525
16	0.791667
57	0.75
18	0.75
19	0.84
20	0.785714
43	0.8125
44	0.756757
46	0.777778
47	0.71875
48	0.763158
49	0.75
50	0.869565
53	0.7
34	0.842105
56	0.666667
38	0.72973
59	1

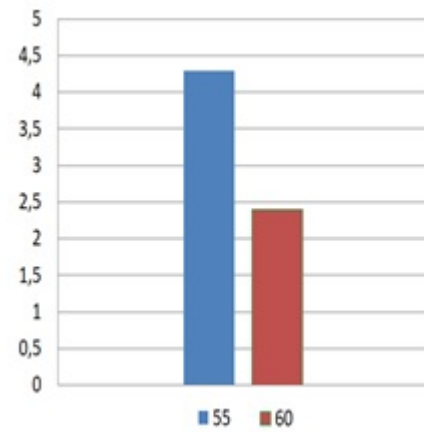


Fig. 4. Borrowers with a low level of reliability

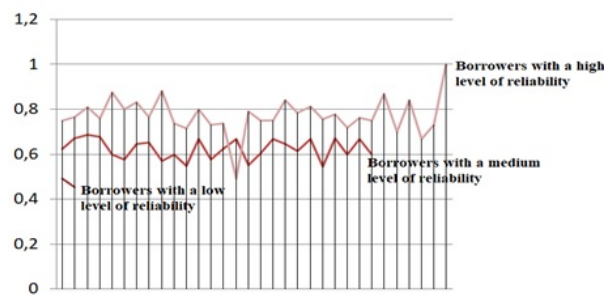


Fig. 5. Graph of score classification of groups of borrowers

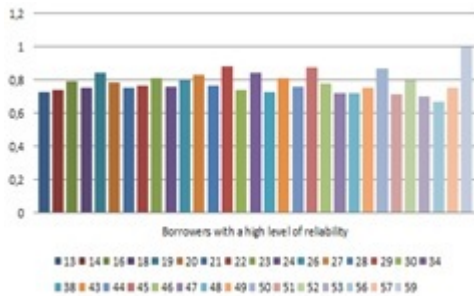


Fig. 3. Borrowers with a high level of reliability

an **medium levels** of reliability can be granted a loan by re-analyzing their data. In addition, Table 6 lists 6, 13, 14, 16, 18-24, 26-30, 34, 38, 41, 43-53, 57, 59 groups of borrowers with **high levels** of reliability. Allocating loans to borrowers with a high level of reliability will lead to effective results. Now, we analyze the market share of customers in terms of reliability levels Figure 2, Figure 3, Figure 4, Figure 5.

From the above graphs, we can see that in the credit market we are analyzing, the share of groups of borrowers with a high risk of default is 3.3%, borrowers with a medium level of reliability 43.3% and trustworthy borrowers 53.4%. The results of this analysis show that the credit institution is effectively lending process.

VII. CONCLUSION

In short, the effectiveness of lending by banks or other lending institutions depends on lending to reliable borrowers. Therefore, many lending institutions face the problem of attracting customers and verifying customer trust.

In solving this problem, a new credit scoring model based on the Monge-Kantorovich problem was proposed. The credit scoring model based on the Monge-Kantorovich problem is distinguished by the fact that it is mathematically based and the conditions for its application are defined. Using this credit scoring model, the process of evaluating borrowers using data from lending institutions is presented, and scientific results are obtained.

Currently, credit scoring systems are widely used by banks or other lending institutions of the Republic of Uzbekistan. This shows that credit scoring models are an effective tool in evaluating borrowers as well as profitable lending. Using this credit scoring model, we can obtain the estimates of borrowers who are most likely to be involved in retail lending and micro lending. This leads to an efficient implementation of the lending process. Therefore, research is being conducted around the world to improve effective credit scoring models and develop new models. Therefore, in order to further develop the credit system in the Republic of Uzbekistan, it is necessary to improve credit scoring mechanisms.

REFERENCES

- [1] V. Bogachev, A. Kolesnikov, "The Monge-Kantorovich Problem: Achievements, Connections and Prospects," *Russian Mathematical Sciences*, 67: 5(407) (2012), pp. 3-110.
- [2] A. Kabulov, I. Saymanov, I. Yarashov, F. Muxammadiev, "Algorithmic method of security of the Internet of Things based on steganographic coding," 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings, pp. 1-5.
- [3] A. Kabulov, E. Urunboev, I. Saymanov, "Object recognition method based on logical correcting functions," 2020 International Conference on Information Science and Communications Technologies, ICISCT 2020.
- [4] A. Kabulov, I. Saymanov, "Application of IoT technology in ecology (on the example of the Aral Sea region)," 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021.
- [5] A. Kabulov, I. Saymanov, M. Berdimurodov, "Minimum logical representation of microcommands of cryptographic algorithms (AES)" 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-6.
- [6] A. Kabulov, I. Normatov, S. Boltaev, I. Saymanov, "Logic method of classification of objects with non-joining classes," *Advances in Mathematics: Scientific Journal* 9 (10), pp. 8635-8646.
- [7] R. Anderson, "The Credit Scoring Toolkit. Theory and Practice for Retail Credit Risk Management and Decision Automation," New York: Oxford university press.
- [8] E. Mudretsova, "Degree work on the topic: Evaluation of customer ratings based on a mathematical model of scoring," M.: MIEM, 2013, pp. 81-89.
- [9] S. Rachev, "The Monge-Kantorovich problem on the displacement of masses and its application in stochastics," *Theory of Probability and its Applications*, 29: 4 (1984), pp. 625-653.
- [10] S. Glasson, "Censored Regression Techniques for Credit Scoring," RMIT University, 2007, pp. 196.
- [11] T. Abdullaev, G. Juraev, "Application three-valued logic in symmetric block encryption algorithms," *Journal of Physics: Conference Series*, 2021, 2131(2), 022082, pp. 1-9.
- [12] T. Abdullaev, G. Juraev, "Selection of the optimal type of the gaming function for symmetric encryption algorithms," *AIP Conference Proceedings*, 2021, 2365, 020004, pp. 1-7.
- [13] T. Abdullaev, G. Juraev, "Development of a method for generating substitution tables for binary and ternary number systems," *AIP Conference Proceedings*, 2021, 2365, 020003, pp. 1-10.
- [14] A. Ikramov, G. Juraev, "The Complexity of Testing Cryptographic Devices on Input Faults," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021, 13041 LNCS, p. 202209.
- [15] G. Juraev, K. Rakhimberdiev, "Modeling the decision-making process of lenders based on blockchain technology," 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-5.
- [16] A. Kabulov, M. Berdimurodov, "Parametric Algorithm for Searching the Minimum Lower Unity of Monotone Boolean Functions in the Process Synthesis of Control Automates" 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-5.
- [17] A. Kabulov, M. Berdimurodov, "Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal)" 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-5.
- [18] A. Kabulov, I. Kalandarov, I. Yarashov, "Problems of Algorithmization of Control of Complex Systems Based on Functioning Tables in Dynamic Control Systems" 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-4.
- [19] A. Kabulov, I. Normatov, I. Kalandarov, I. Yarashov, "Development of An Algorithmic Model and Methods for Managing Production Systems Based on Algebra over Functioning Tables" 2021 International Conference on Information Science and Communications Technologies, ICISCT 2021, pp. 1-5.
- [20] S. Banerjee, R. Biswas, M. Gangopadhyaya, "Design of Grey Wolf Optimizer based Amended Equalizer for Universal Mobile Telecommunications System," 2021 5th International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2021, 2021
- [21] S. Ray, A. Chatterjee, D. Lodh, S. Bishnu, M. Gangopadhyaya, "Resonant Frequency Optimization of L-Shaped Feed Cylindrical Liquid Antenna using Genetic Algorithm," 2021 5th International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2021, 2021.
- [22] H. Khujamatov, I. Siddikov, E. Reypnazarov, D. Khasanov, "Research of Probability-Time Characteristics of the Wireless Sensor Networks for Remote Monitoring Systems," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, 2021.
- [23] I. Siddikov, D. Khasanov, H. Khujamatov, E. Reypnazarov, "Communication Architecture of Solar Energy Monitoring Systems for Telecommunication Objects," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, 2021.
- [24] H. Khujamatov, A. Lazarev, N. Akhmedov, E. Reypnazarov, A. Bek-turdiev, "Methods for Automatic Identification of Vehicles in the its System," International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021, 2021.
- [25] A. Kabulov, A. Babadzhanov and I. Saymanov, Correct models of families of algorithms for calculating estimates, *AIP Conference Proceedings*, 2022 (accepted).
- [26] A. Kabulov, A. Babadzhanov and I. Saymanov, Completeness of the linear closure of the voting model, *AIP Conference Proceedings*, 2022 (accepted).
- [27] A. Kabulov, I. Normatov, E. Urunbaev, F. Muhammadiev, "Invariant continuation of discrete multi-valued functions and their implementation," 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [28] A. Kabulov, A. Ashurov, M. Berdimurodov, "Analytical transformations in minimizing logical functions," 2020 International Conference on Information Science and Communications Technologies, ICISCT 2020.
- [29] A. Kabulov, I. Normatov, A. Seytov, A. Kudaybergenov, "Optimal management of water resources in large main canals with cascade pumping stations," IEMTRONICS 2020 - International IOT, Electronics and Mechatronics Conference, Proceedings.
- [30] A.V. Kabulov, I.H. Normatov, A. Karimov, "Algorithmization control of complex systems based on functioning tables," *Journal of Physics: Conference Series* 1441 (1).
- [31] A. Kabulov, I. Normatov, E. Urunbaev, A. Ashurov, "About the problem of minimal tests searching," *Advances in Mathematics: Scientific Journal* 9 (12), pp. 10419-10430.
- [32] A. Kabulov, I. Normatov, A. Karimov, E. Navruzov, "Algorithm of constructing control models of complex systems in the language of functioning tables," *Advances in Mathematics: Scientific Journal* 9 (12), pp. 10397-10417.
- [33] A. Kabulov, E. Urunbaev, I. Normatov, A. Ashurov, "Synthesis methods of optimal discrete corrective functions," *Advances in Mathematics: Scientific Journal* 9 (9), pp. 6467-6482.
- [34] M. Shaw, N. Mandal, M. Gangopadhyay, "A compact polarization reconfigurable stacked microstrip antenna for WiMAX application," *International Journal of Microwave and Wireless Technologies*, 2021, 13(9), p. 921936
- [35] M. Shaw, S. Mitra, M. Gangopadhyay, "A triple band scalene triangular microstrip patch antenna," 2021 5th International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2021, 2021
- [36] A. Kabulov, I. Kalandarov, I. Saymanov, "Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules," *Transportation Research Procedia*
- [37] A. Kabulov, "Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions," *Cybernetics*, 1979, 15(3), p. 305311.
- [38] A. Kabulov, G. Losev, "Local algorithms simplifying the disjunctive normal forms of Boolean functions," *USSR Computational Mathematics and Mathematical Physics*, 1978, 18(3), p. 201207.
- [39] A. Kabulov, "Local algorithms on yablonskii schemes," *USSR Computational Mathematics and Mathematical Physics*, 1977, 17(1), p. 210220.
- [40] I. Normatov, E. Kamolov, "Development of an algorithm for optimizing the technological process of kaolin enrichment," IEMTRONICS 2020 - International IOT, Electronics and Mechatronics Conference, Proceedings, 2020, 9216371

Forecasting Model Comparison for Soil Moisture to Obtain Optimal Plant Growth

Sachintha Balasooriya*
Faculty of Engineering,
Sri Lanka Technological
Campus
Padukka, Sri Lanka
Sachinthab@sltc.ac.lk

Chuong Nguyen
School of Science and
Technology,
RMIT University,
Ho Chi Minh, Vietnam
s3651570@rmit.edu.vn

Ilya Kavalchuk
Department of Electrical
Engineering,
Alasala University,
Saudi Arabia
ikavalch@gmail.com

Lasith Yasakethu
Faculty of Engineering,
Sri Lanka Technological
Campus
Padukka, Sri Lanka
lasithy@sltc.ac.lk

Abstract: The advent of industry 4.0 has seen a massive increase in the connectivity of electronic devices to the internet. It also results in the implementation of data gathering schemes for environmental factors. One such field is the agricultural sector. In Sri Lanka, where this research was conducted, agriculture accounts for one fifth of the country's gross national production. The introduction of wireless sensor networks in the field of agriculture has shown some of the underlying factors that affect the crops and by extension, the harvest and yields. Recoding of environmental factors such as soil moisture, temperature, humidity, sunlight, etc. has enabled the modeling of the conditions in the plantations and nurseries. Thereby, delivering an understanding of what suboptimized factors can be improved. Also, two models are utilized to forecast the next-step moisture content at Boralanda town in Sri Lanka based on previous read values: Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long-Short Term Memory (LSTM) Neural Network. It is shown that the LSTM model is superior with much lower error when predicting many time steps.

Keywords; Precision agriculture, LSTM, SARIMA, Wireless sensor networks

I. INTRODUCTION

Agriculture still remains on the main sources of income for developing countries like Sri Lanka. Over 25% of the work force in Sri Lanka is engaged in the agricultural sector [1-3]. Developing technologies in the fields of wireless connectivity and machine learning have given rise to new ways of managing agriculture and optimizing the processors, with maximizing the yields and reducing production costs. Precision agriculture delivers the means of observing the environmental features of the area, assessing and predicting the inputs needed for the short-term and long term in the future. In order to improve the crop yields and by extension agricultural production of the country, the gathered information and the prediction reached by the computer simulations can be delivered

to the farmer to adjust the input to the plantations and take better care of the crops. In the long term this could save a large amount of money, since the inputs to the plants are precisely controlled and the wastage and over use can be drastically reduced.

Another aspect of this implementation is for horticultural nurseries. Horticulture uses the knowledge of growing plants to intensively produce plants. The process is done not only to generate food, but mostly to propagate ornamental and flowering plants. In a market like Sri Lanka, the horticultural industry accounts for 20% of the total agricultural output of the country [4, 5]. It can be seen as very lucrative emerging market mainly for exportation.

One important aspect of this paper is introducing a suitable forecasting model for forecasting soil moisture content by several steps based on the captured data of the moisture sensor itself. The experiment data includes moisture content captured within 18 days in October 2019 with 1351 samples (sample rate of about 10 minutes) at Boralanda town in Sri Lanka. The data set is then shown to be non-stationary, where trend and seasonality components exist. A stationary series is a series where its mean and variance don't change over time and covariance function is not a function of time. A trend makes a series become non-stationary by adding an increasing or decreasing trend component, which changes the mean overtime. Also, seasonality is present when there is a repeated pattern in a same time period while the other time periods don't include the pattern, which changes the mean and variance overtime [6]. We will overcome these problems by introducing and comparing two different forecasting model – Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long-short Term Memory (LSTM) Neural Network [7]. SARIMA is a well-known and simple forecasting model that takes cares of both non-seasonal and seasonal components in a series. LTSM

is a prevailing neural network structure that can be used to fit sequential data and a high degree of complexity [8].

The paper presents the observation and modelling of the environmental conditions in a controlled green house. The main focus has been analysis of soil moisture in the plant pots. Section II-precision agricultural systems; is an in-depth explanation on the different stages in a general precision agricultural system. Section III-data gathering electronics; will give a briefing on the electronics behind the data gathering process. Section IV-forecasting models; introduces the machine learning algorithms and software used to create the predictive model. Section IV will be an explanation of the model generated. Section V presents the results of this experiment and the advantages of using such a system. Finally, Section V would be the conclusions and future work.

II. PRECISION AGRICULTURAL SYSTEMS

This application of wireless networks that measure environmental conditions for automated control of resources or predicting and/or estimating the approximate depletion time of resources is known as precision agriculture [9]. Furthermore, this concept expands as, to create an maintain a complex system, typically controlled through electronic computation for adjusting resources and nutrients in the soil to attune to a specific composure suited for a plant variety [10].

The stages of precision agricultural system can be described as follows [11]. The initial and most important step of setting up a precision agricultural infrastructure is locating the wireless sensor network. The sensor nodes will act as the feelers in the system to gauge in sudden changes in environmental conditions. These sensors are used to measure vital information such as [12, 13]; soil moisture, pH level of soil, air temperature (if within a greenhouse), humidity, sunlight levels, etc. the next stage of the infrastructure is the data gather and accumulation. The data gathering includes the gauged parameter as well as the location of the sensing instrument. This data is transferred to a local route to then be recorded on a server. Finally, control decisions are made by humans or artificial intelligence based on the sensor data [14].

III. DATA GATHREING ELECTRONICS

One of the major issues in this application of IoT (Internet of Things) is the maintenance of the sensor network and electronics. Two challenges in managing the electronics are; exposure of physical components

to harsh environmental conditions, and the battery life of nodes.

Since the sensors need to be in physical contact with the elements to obtain readings, their internal components should be shelled for damage. Ultimately, this is a role that is passed on to the manufacturer.

Secondly, the nodes/sensor modules are required to operate with minimum interaction or maintenance by the humans. Typically, the deployment of such networks is vast and covers a large area. Therefore, the number of planted nodes could easily reach up to hundreds. In such an installation, retrieval of nodes for periodic maintenance such as battery recharging would be cumbersome. Therefore, optimization for energy efficiency is of the utmost importance. [15], has conducted their own research as well as presented past research on the energy consumption of wireless sensor networks designed for agricultural purposes. Accordingly, the devices transmit data to the routed periodically. In the meantime, in between transmissions, the electronics are forced into a sleeping mode so the current draw off the battery is low (in microampere range) [16, 17].

For reference, the data gathering instruments used in this research record and transmit data every 10 minutes. During test the electronics system only drew 56 microamperes during forced sleep mode. One single node in this implementation has enough energy to run for approximately one year without recharging. This is not to say that unexpected maintenance issues will not arise.

A. Embedded Hardware design

In this specific application we have adopted a simple embedded system with a microcontroller attached to wireless communication module for data transmission and a capacitive soil moisture sensor for gather information on the moisture level on the soil as shown in the Fig. 1.

The data measured by the soil sensor is converted internally in the MCU to a percentage moisture value. The moisture within any sample will not drop significantly, instead it will gradually decrease under natural conditions. Therefore, multiple reading for one sample is not required.

The NRF24 wireless communications module was selected to transmit the data from the node to the router placed within each green house. The module and their communication protocol were selected for its capability to transmit the data packages securely while also consuming a very low amount of power. The gathered data within each green house is transmitted

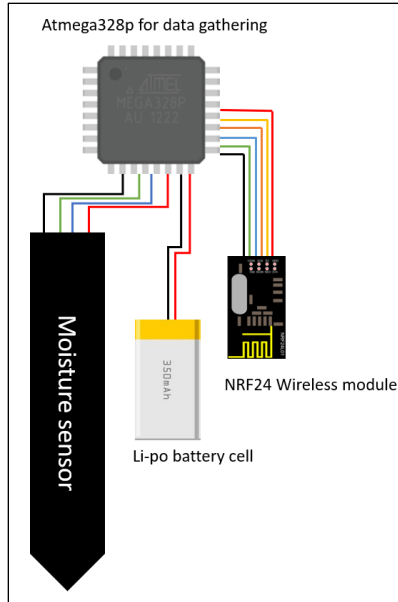


Fig. 1 – Embedded hardware system

by the router to the cloud over Wifi/ethernet, where it was subjected to the following analysis as described on this paper.

IV. FORECASTING MODELS

A. Seasonal Autoregressive Integrated Moving Average

Seasonal Autoregressive Integrated Moving Average (SARIMA), an extended version of ARIMA, is one of the most popular forecasting methods for univariate time series data. ARIMA, while having the capability to deal with trend components of data, lacks the consideration of seasonal components in its own algorithm. SARIMA expands its capability by adding autoregression (AR), differencing (I) and moving average (MA) for seasonal components as part of its equation.

Here, the autoregression (AR) model is the regression of time-series values to itself. In particular, the concurrent values of the series is considered to depend on its previous values, called lags. The maximum lag is denoted as p . The initial value of p is determined by using a partial auto-correlation function (PACF) plot. The Moving Average (MA) model presents the error of time series, with the argument that the current value of error also depends on some of its own lag, referred to as q . The initial value of q can be determined using an auto-correlation function (ACF) plot. Finally, I is the order of integration, which is the number of non-seasonal differences needed to make the time series become stationary [7]. The general

model is denoted as $SARIMA(p, d, q) \cdot (P, D, Q)^S$, where: p, d, q are AR, I and MA orders of trend elements, respectively; P, D, Q are AR, I and MA orders of seasonal elements, respectively. In 1970, Box and Jenkins introduced a multiplicative model of SARIMA, provided in [18].

$$\phi_p(B)\phi_p(B^S)\nabla^d\nabla_s^D X_t = \theta_q(B)\theta_q(B^S)\epsilon_t \quad (1)$$

Here, B is the backward shift operator.

The ordinary AR and MR components are presented as the polynomials $\phi_p(B)$ and $\theta_q(B)$ of the order p and q , respectively.

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 \dots - \phi_p B^p \quad (2)$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 \dots - \theta_q B^q \quad (3)$$

The seasonal AR and MA components are presented as the polynomials $\Phi_P(B^S)$ and $\Theta_Q(B^S)$ of the order P and Q , respectively.

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} \dots - \Phi_P B^{PS} \quad (4)$$

$$\Theta_Q(B^S) = 1 - \theta_1 B^S - \theta_2 B^{2S} \dots - \theta_Q B^{QS} \quad (5)$$

The non-seasonal and seasonal difference components:

$$\nabla^d = (1 - B)^d \quad (6)$$

$$\nabla_s^D = (1 - B^S)^D \quad (7)$$

Finally, X_t is observed value at time t ($t = 1, 2, 3 \dots, n$) and ϵ_t is residual at time t

The approach of Box and Jenkins requires four phases [19]:

Model identification. This phase involves analyzing and verifying the stationarity of the time series using various methods. There methods can be using ACF and PACF plots; Mann-Kendall trend test; unit root test. In addition, the type of the model (non-seasonal or seasonal) and its orders (p, q, P, Q) are determined using ACF and PACF plots, as discussed.

Model Estimation. This phase involves reviewing the model (as well as its orders) defined in previous phase and using a common error metric to determine the best one. Some popular error metrics are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)

$$MAE = \frac{\sum |error|}{n} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum (error)^2}{n}} \quad (9)$$

Model Diagnostic. This phase uses the ACF and PACF plots of residuals to check if the residuals follow white noise or not. If it doesn't, we need to choose another model.

Model Forecasting and evaluation. This last phase involves using the model to forecast the future outcome. Moreover, testing data can be used to verify the performance of the model using suitable error criteria (MAD or MSE).

Four phases are iterated until a satisfactory model is achieved.

B. Long-short Term Memory Neural Network

Recurrent neural network (RNN) is an artificial neural network architecture specialized for sequential data such as handwriting, speech etc. Therefore, it is capable of handling time-series data as well. Here is the general structure of RNN

We can observe that the network feedback to itself at each time step. Hence, it can be unfolded as multi layers feed forward neural network. The formula for hidden state s_t of network and its output y_t are given in [20]

$$s_t = f(Ux_t + Ws_{t-1}) \quad (10)$$

$$y_t = g(Vs_t) \quad (11)$$

Fig. 2 shows x_t as the input to the system at time t , s_t is the hidden state of the cell at time t , which is stands for memory of the network; f is an activation function for the hidden state (\tanh or ReLU); U, V, W are parameters of network, which is shared across all time steps; y_t is the output of the network at time t and g is its activation function (SoftMax function).

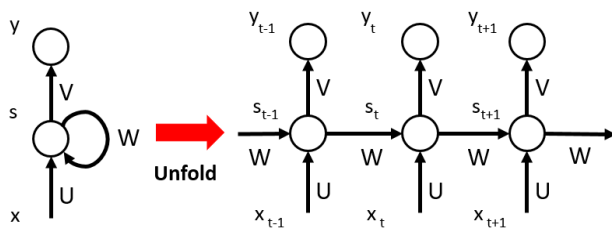


Fig. 2 - Recurrent Neural Network

It can be observed that with such an enormous number of layers as the number of input increases, the gradients propagate all the way back to initial values will either suffer from vanishing gradients or exploding gradient. In particular, the deeper of the layers, the more matrix multiplications have to be applied because of the chain rule. Therefore, if they are small, the

The gradient will shrink exponentially to an extremely small value, such that the model cannot learn more. This phenomenon is called vanishing gradient. On other hand, if they are large, the gradients get exponentially larger and eventually crash the model. This phenomenon is called zn exploding gradient. Long-short Term Memory (LSTM) Neural Network solves these problems by introducing the concept of using different activation function layers called “gates” [8]. The proposed structure of one cell of LSTM is given as [21] (Fig. 3).

In Fig. 3 showing one cell of LTSM, x_t is current input at time t ; h_{t-1} and c_{t-1} are the previous cell output and state, respectively; h_t and c_t are previous cell output and state, respectively. f_t is the forget gate.

This sigmoid layer determines which data is irrelevant and should be forgotten. It takes x_t and h_{t-1} as input and give Boolean value as output for each piece of information (0 is giving up the information while 1 is keeping the information)

$$f_t = \sigma(x_t * U_f + h_{t-1} * W_f) \quad (12)$$

i_t is the input gate. This sigmoid layer determines which new data is to be written to cell state. Similar to the forget gate, its outputs are Boolean values that decide which piece of new information should be written

$$i_t = \sigma(x_t * U_i + h_{t-1} * W_i) \quad (13)$$

\bar{c}_t is the candidate gate. This \tanh layer regulates the vectors created by x_t and h_{t-1} to be in the range from -1 to 1 , which provides faster convergence.

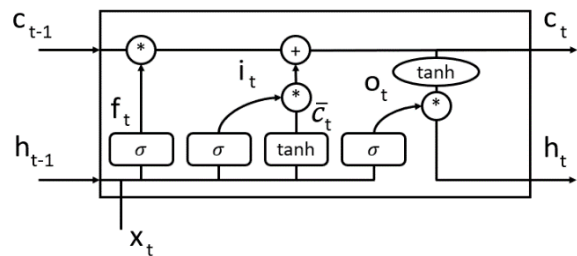


Fig. 3 – One cell of LTSM

$$\bar{c}_t = \tanh(x_t * U_c + h_{t-1} * W_c) \quad (14)$$

The multiplication between i_t and \bar{c}_t provides new regulated chosen values.

Next, the new value of current state c_t is calculated using previous state c_{t-1} along with calculated f_t , i_t and \bar{c}_t

$$c_t = f_t * c_t + i_t * \bar{c}_t \quad (15)$$

Finally, the output h_t of the cell is generated using 2 components: are regulated version of current cell state c_t using \tanh function and filtered values of h_{t+1} and x_t .

$$o_t = \sigma(x_t * U_o + h_{t-1} * W_o) \quad (16)$$

$$h_t = o_t * \tanh(c_t) \quad (17)$$

V. FORECASTING RESULTS

A. Development of SARIMA model

The data set consists of 1351 data points of moisture content (percentage) of the soil within the greenhouses located in Boralanda town in Sri Lanka with a sample rate of 10 minutes. As we have 6 observations per hour and one day (24 hours) is a full cycle, the frequency that we choose to decompose the data into additive components is 144 (6 multiplied with 24) and shown on Fig. 4.

Here, we can observe the presence of trending and seasonality in the data. Hence the data shows strong correlation to each other as in Fig.5 and is non-stationary. This implies that the current data set is inappropriate since a linear model like SARIMA requires the observations to be independent [19].

We also deduced the ACF and PACF plots of the original data in Fig. 5 and Fig. 6.

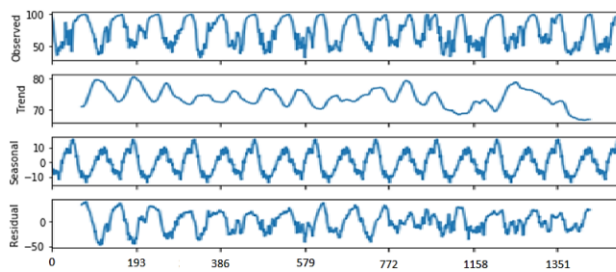


Fig. 4 – Decomposed additive components

In order to make the data stationary, the first seasonal difference was taken, as shown on Fig. 7 and Fig. 8.

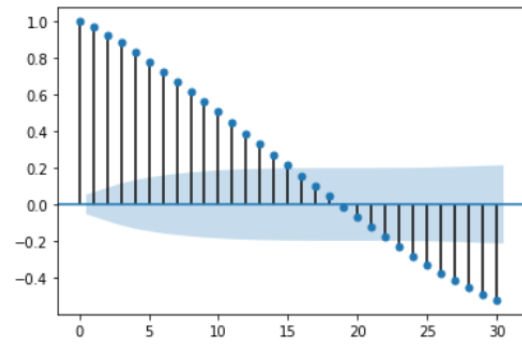


Fig. 5 - ACF of original data

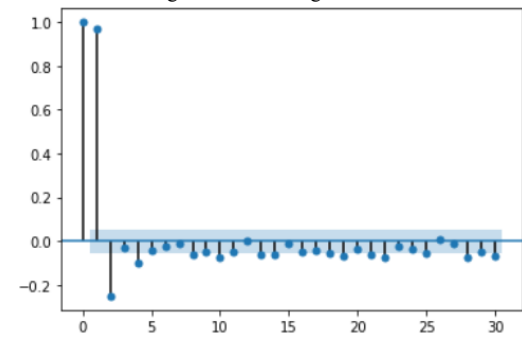


Fig. 6 – PACF of original data

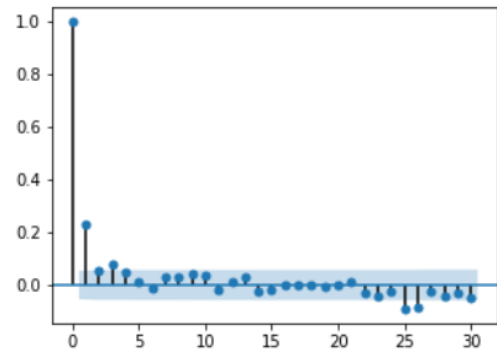


Fig. 7 - ACF of D=1

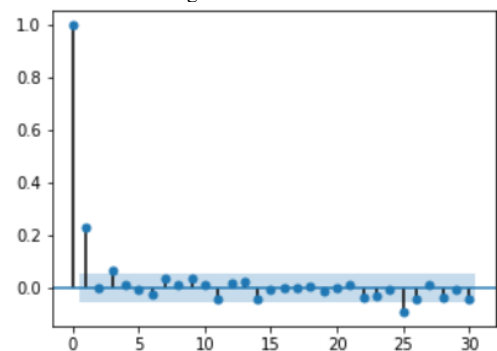


Fig. 8 – PACF of D=1

We can observe that after the first seasonal difference, there is a very little correlation between the data. Thus, the series is considered to be stationary.

In the ACF plot, the positive significant spikes at lag 1 and 3 suggest a non-seasonal MA(2) component, and two negative significant spikes at lag 25 and 26 might suggest a seasonal MA(2) component. Also, in PACF plot, the positive significant spikes at lag 1 and 3 suggest a non-seasonal AR(2) and the negative significant spikes at lag 25 suggest a seasonal AR(1) component. Consequently, our final model for SARIMA is: $(2, 0, 2) \cdot (1, 1, 2)^2$.

After this, the SARIMA model will be trained with the original data.

B. Development of LSTM model

The training of the LSTM model involves the following steps:

- **Difference the series.** As discussed, a stationary series is very easy to predict and fit well in training model. The difference series is performed with one lag, similar to SARIMA. Also, this implementation is used to facilitate overfitting due to the correlation between the data.
- **Scale the data.** The data will be scaled down to the range from -1 to 1 in order to achieve fast convergence.
- **Train the model with training data.** Here, 4 neurons are used for LSTM network with a batch size of 1. The training iteration is chosen as 30.
 - **Predict the next steps with inversed scaling and inversed difference.** After training the model, we use it to predict the next steps where the predicted value will be used to predict the next step. Also, we need to scale back the predicted data to its intended value and apply back the non-stationarity.

C. Forecasting results and discussion

The results of two model with 48-step prediction for a duration of 8 hours are given in Fig. 9. The output prediction of the models compared to the raw test data is discussed with the corresponding inputs to the system. Table 1 gives the RMSE and the MAE values of the two compared networks to the original raw data.

Table 1 – RMSE and MAE of forecasting model

Model	RMSE	MAE
SARIMA	10.3617	6.27421
LSTM Network	1.98741	1.26152

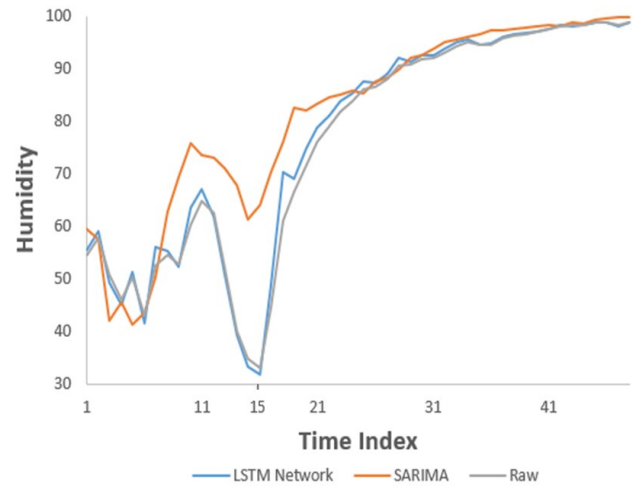


Fig. 9 – Comparison of SARIMA and LSTM network in forecasting moisture data

Although, two models exhibit a fairly small amount of error, compared with SARIMA, LSTM network performs significantly better for our data set of humidity with a much smaller RMSE and MAE. Therefore, we choose to use LSTM in forecasting the content of moisture.

The graph proves how close the LSTM Network follows the raw data compared to the SRIMA Network. Even at sudden changes, for example around 15-time index (which corresponds to the period of excessive evaporation in the morning), the LSTM Network prediction is acceptably accurate. Whereas, the SARIME Network shows a significant deviation from the real case.

VI. CONCLUSION AND FUTURE WORK

In conclusion, we have tested two different networks for forecasting the required data for a selected amount of time. We have successfully defined and compared two forecasting models: SARIMA and LSTM. By observing the results, it is evident that the LSTM network is much more versatile and performs outstandingly better than the controlled network. Due to lower error in predicting many time steps ahead, LSTM is considered to be more suitable for our model.

For future work, the models can and will be expanded for multiple parameter analysis. The correlations of humidity, sunlight, air temperature and weather forecasts with soil moisture can be taken into account in the future. The collection of larger sets of data and different combinations of SAMRIMA would potentially lead to the discovery of better models to predict resource availability and requirements for the plant ahead of time. By precise control of the inputs

needed for the plant growth, the cost due to wastage can be significantly reduced, and by extension, lift the financial burden on the farmers and planters of Sri Lanka.

VII. REFERENCES

- [1] M. Esham, C. J. C. Garforth, and Development, "Climate change and agricultural adaptation in Sri Lanka: a review," vol. 5, no. 1, pp. 66-76, 2013.
- [2] M. A. Wijeratne, "Vulnerability of Sri Lanka tea production to global climate change," vol. 92, no. 1-2, pp. 87-94, 1996.
- [3] S.-N. N. Seo, R. Mendelsohn, M. J. E. Munasinghe, and d. Economics, "Climate change and agriculture in Sri Lanka: a Ricardian valuation," vol. 10, no. 5, pp. 581-596, 2005.
- [4] D. J. D. Dunham and change, "Crop diversification and export growth: dynamics of change in the Sri Lankan peasant sector," vol. 24, no. 4, pp. 787-813, 1993.
- [5] R. P. Mahaliyanaarachchi, H. R. Rosairo, and M. J. F. o. A. s. Esham, Sabaragamuwa, University of Sri Lanka, "Potential High Value Horticultural Crops, Their Financial and Marketing Feasibility," 2004.
- [6] K. Yang and C. Shahabi, "On the stationarity of multivariate time series for correlation-based data analysis," presented at the Fifth IEEE International Conference on Data Mining, Houston, TX, USA, 2005.
- [7] H. Musbah and M. El-Hawary, "SARIMA Model Forecasting of Short-Term Electrical Load Data Augmented by Fast Fourier Transform Seasonality Detection," presented at the 2019 IEEE - Canadian Conference on Electrical and Computer Engineering (CCECE), Edmonton, 2019.
- [8] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM Neural Networks for Language Modeling," presented at the 13th Annual Conference of the International Speech Communication Association, Portland, 2012.
- [9] S. Blackmore, "Precision Farming: An Introduction", Outlook on agriculture, vol. 23, no. 4, pp. 275-280, 1994, doi:10.1177/003072709402300407.
- [10] D. J. Mulla and Y. Miao, "Precision farming". Journal of Water Resource and Protection, Vol.1 No.4, October 16, 2009.
- [11] A. Kalra, R. Chechi, and R. Khanna, "Role of Zigbee Technology in agriculture sector," in *National Conf. on Computational Instrumentation NCCI*, 2010, p. 151.
- [12] V. I. Adamchuk, J. Hummel, M. Morgan, S. J. C. Upadhyaya, and e. i. agriculture, "On-the-go soil sensors for precision agriculture," vol. 44, no. 1, pp. 71-91, 2004.
- [13] V. I. Adamchuk, R. V. Rossel, K. A. Sudduth, P. S. J. S. F.-F. Lammers, and R. Applications. InTech, Croatia, "Sensor fusion for precision agriculture," pp. 27-40, 2011.
- [14] H. Liu, Z. Meng, and S. Cui, "A Wireless Sensor Network Prototype for Environmental Monitoring in Greenhouses," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 2007, pp. 2344-2347.
- [15] J. Song and Y. K. Tan, "Energy consumption analysis of ZigBee-based energy harvesting wireless sensor networks," in *2012 IEEE International Conference on Communication Systems (ICCS)*, 2012, pp. 468-472: IEEE.
- [16] S. Kellner, M. Pink, D. Meier, and E.-O. BlaB, "Towards a realistic energy model for wireless sensor networks," in *2008 Fifth Annual Conference on Demand Network Systems and Services*, 2008, pp. 97-100: IEEE.
- [17] G. Anastasi, M. Conti, M. Di Francesco, and A. J. A. h. n. Passarella, "Energy conservation in wireless sensor networks: A survey," vol. 7, no. 3, pp. 537-568, 2009.
- [18] F. F. Nobre, A. B. S. Monteiro, P. R. Telles, and G. D. Williamson, "Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology," *Statistics in Medicine*, vol. 20, pp. 3051-3069, 2001.
- [19] K. B. Tadesse and M. O. Dinka, "Seasonal Time Series Forecasting using SARIMA and Holt Winter's Exponential Smoothing", *Journal of Water and Land Development*, vol. 35, pp. 229-236, 2017.
- [20] M. Lukoševičius and H. tJaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, pp. 127-149, 2009.
- [21] J. Zhang, P. Wang, R. Yan, and R. X. Gao, "Long short-term memory for machine remaining life prediction," *Journal of Manufacturing Systems*, vol. 48, pp. 78-86, 2018.

Detection and Quantitative Prediction of *Diplocarpon earlianum* Infection Rate in Strawberry Leaves using Population-based Recurrent Neural Network

Oliver John Alajas^{1*}, Ronnie Concepcion II², Argel Bandala¹, Edwin Sybingco¹, Ryan Rhay Vicerra², Elmer P. Dadios², Christian Hail Mendigoria¹, Heinrick Aquino¹, Leonard Ambata¹, Bernardo Duarte³

¹Department of Electronics and Computer Engineering, De La Salle University, Manila, Philippines

²Department of Manufacturing Engineering and Management, De La Salle University, Manila, Philippines

³MARE - Marine and Environmental Sciences Centre, Faculty of Sciences, University of Lisbon Campo Grande, Lisbon, Portugal
{oliver_alajas*, ronnie.concepcion, argel.bandala, edwin.sybingco, ryan.vicerra, elmer.dadios, christian_mendigoria, heinrick_aquino, leonard.ambata}@dlsu.edu.ph, baduarte@fc.ul.pt

Abstract—*Fragaria ananassa*, a member of the rose family's flowering plants, commonly recognized as strawberry, is prone to *Diplocarpon earlianum* infection that causes leaf scorch. Assessment via visual inspection of strawberries by farmers is normally ineffective, destructive, and laborious. To address this challenge, the use of integrated computer vision and machine learning techniques was done to classify a healthy from a scorch-infected strawberry leaf image and to estimate the leaf region infection rate (LRIR). A dataset made up of 204 normally healthy and 161 scorch-infected strawberry leaf images was used. Images were initially preprocessed and segmented via graph-cut segmentation to extract the region of interest for feature extraction and selection. The hybrid combination of neighborhood and principal component analysis (NCA-PCA) was used to select desirable features. Multigene genetic programming (MGPP) was used to formulate the fitness function that will be essential for determining the optimized neuron configurations of the recurrent neural network (RNN) through genetic algorithm (GA), and cuckoo search algorithm (CSA), and artificial bee colony (ABC). Four classification machine learning models were configured in which the classification tree (CTree) bested other detection models with an accuracy of 100% and exhibited the shortest inference time of 14.746 s. The developed ABC-RNN₃ model outperformed GA-RNN₃ and CSA-RNN₃ in performing non-invasive LRIR prediction with an R² value of 0.948. With the use of the NCA-PCA-CTree₃-ABC-RNN₃ hybrid model, for crop disease detection and infection rate prediction, plant disease assessment proved to be more efficient and labor cost-effective than manual disease inspection methods.

Keywords—*bio-inspired optimization, computational intelligence, computer vision, digital agriculture, leaf scorch, plant pathology, strawberry leaf spot*

I. INTRODUCTION

Detection and identification of plant diseases is a commonly experienced problem in the agricultural crop production industry since it directly affects the quality of products being sold to local food markets. Researchers found that while these diseases show symptoms through different

parts of a plant, leaves tend to be the most frequently observed part for the detection of infection [1]. Various studies have been conducted that use digital image processing methods to detect and classify leaf diseases [2, 3]. Using a vision-based approach to detect the disease by utilizing leaf images is one of the best options researchers tend to go to [3].

Fragaria ananassa which is a common type of strawberry species that bears strawberry fruit is a good source of vitamin C for the improved immune system, potassium for lowering blood pressure, antioxidants for eliminating free radicals, and reducing inflammation, flavonoids such as anthocyanins, quercetin, and kaempferol for cancer prevention. It also promotes healthy eyesight because it contains ellagic acid and phenolic phytochemicals. The fruit has a low glycemic index and, it contains dietary fiber that helps regulate blood sugar levels [4]. However, it is prone to *Diplocarpon Earlianum* infection which is also known as leaf scorch. It is characterized by small purplish blemishes that appear on the top side of the leaf. Scorched-infected strawberry farms typically suffer from a decline in their strawberry produce's quality and taste. Late diagnosis of scorch infection is one of the main reasons for this. Moreover, monitoring and visually inspecting these fruits are typically done manually which is quite ineffective due to its dependence on the observer's bias, expertise, and observation skills.

A study conducted by [5] used the hybrid model utilizing linear discriminant analysis and decision tree to detect and quantify the damage of bacteria-infected grape leaves, [6] used a gaussian quantum-behaved particle swarm and recurrent neural network to assess the disease of corn leaves suffering from leaf spots, and [7] classified three rice leaf disease using a hybrid machine learning and deep neural network. Genetic programming, genetic algorithm, and machine learning models were implemented by [8-12] with the utilization of lettuce features to monitor the growth stages of lettuce. *Musa acuminata*'s post-harvest analysis was done by [13] using hybrid machine learning and deep transfer networks. Meanwhile, research by [14] used an adaptive

neuro-fuzzy inference system to classify three different lettuce seed varieties. On the other hand, research by [15] examined the use of lettuce images to indirectly predict the amount of nitrate present in aquaponic water flowing in a crop chamber system of a vertical farm with the use of a recurrent neural network that is optimized by genetic algorithm.

In this research, machine learning (ML) algorithms such as support vector machine (SVM), Naive Bayes (NB), classification tree (CTree), and linear discriminant analysis (LDA), were implemented to detect a scorch-infected leaf. This detected leaf is then examined to quantify the leaf region infection rate (LRIR) in terms of the area of the whole leaf with the use of a three-layered recurrent neural network (RNN) that is optimized by three bio-inspired optimization algorithms: genetic algorithm (GA), cuckoo search algorithm (CSA), and artificial bee colony (ABC). These models are responsible for the infection rate prediction process of the study. This study contributes to the: (1) development of an efficient and non-invasive vision-based detection of strawberry scorch leaf disease that would help the farmer to employ better agricultural practices; (2) development of an innovative and accurate technique in predicting the leaf infection rate that would help plant pathologist on how to treat strawberry leaves with such severity of infection; and (3) determination and analysis of most significant leaf phenotypes in terms of morphological, texture and spectral traits that have high relevance to leaf scorch disease.

II. MATERIALS AND METHODS

For this study, the realization of the model framework design is presented in Fig. 1. Input images of two classes of leaves (healthy and scorched-infected) were enhanced with the use of contrast improvement to make a distinct identity between the two and, to make the segmentation procedure much simpler. Two types of the strawberry leaf are being initialized as an input for image processing: healthy leaf and scorch-infected leaf. These images were segmented to extract the leaf area of both classes (whole and scorched area) for the calculation of the scorch infection rate. Once features are extracted, an optimized recurrent neural network will do the job of infection rate prediction. To accomplish such a challenging task, MATLAB R2020b software was used to perform vision-based leaf inspection, computational processing, model creation, optimization, classification, and estimation.

A. Strawberry Leaf Dataset Information

The dataset is composed of 365 images with a 1:1 aspect ratio and 256-by-256-pixel density. The number of the healthy leaf is 204 images while the scorched-infected leaf is 161 images. The dataset can be found on [16].

B. Image Enhancement (Contrast Improvement)

Completely raw images of the strawberry leaf (healthy and scorched) were enhanced by adjusting the contrast of the images through histogram equalization. By applying the *imadj* built-in command of MATLAB, the images' top 1% and bottom 1% of all pixel values were adjusted from a default value of 0 to 1 for low-in, high-in, and low-out, high-out values. By doing this, the color map of the image histogram

will stretch from a 0 to 255 range. The contrast settings used were as follows: (0.2, 0.9), (0, 1). An image output sample together with its corresponding histogram is presented in Fig. 2. To make the image enhancement task easier, a looping program was designed to preprocess all 365 images simultaneously with the use of an image batch processor, a built-in application inside MATLAB software.

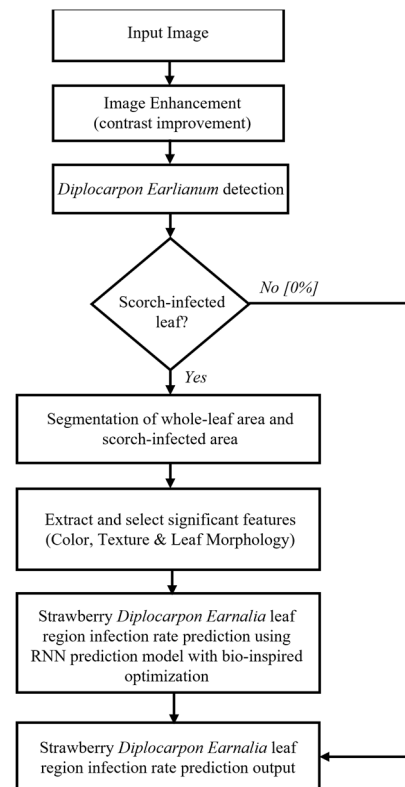


Fig. 1. Methodological framework for the strawberry leaf scorch classification and infection rate prediction using bio-inspired (population-based) optimization of recurrent neural network

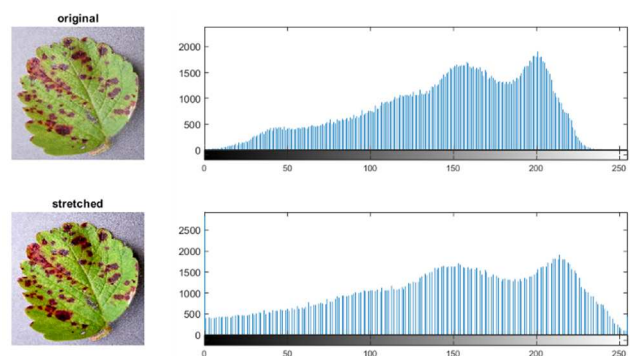


Fig. 2. Image histogram stretching of scorch-infected strawberry leaf for better segmentation results

C. Strawberry Leaf Feature Extraction, Selection, and Dataset Normalization

To establish a model capable of detecting and computing the amount of infection, eighteen features were extracted from the input images. It consists of color features (RGB,

HSV, L a*b*, YCbCr), textural features (correlation, contrast, energy, entropy, homogeneity), and morphological features (leaf vegetative pixel area) (Fig. 3). These features are the primary basis for the prediction model capable of estimating the leaf region infection rate of a strawberry leaf infected by *Diplocarpon Earlianum*. Using the extracted morphological area of the scorch-infected leaf ($A_{scorched}$) and the whole leaf area ($A_{whole\ leaf}$), the leaf region infection rate (LRIR) can be solved using (1).

$$LRIR = (A_{scorched} / A_{whole\ leaf}) \times 100 \quad (1)$$

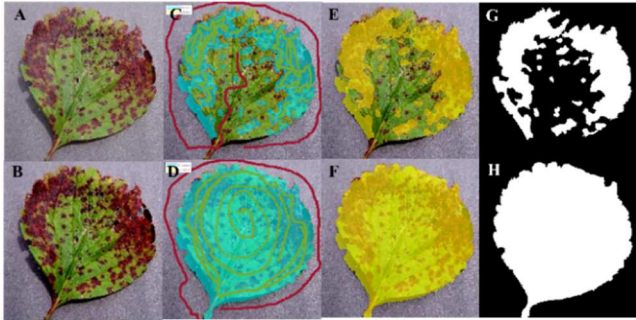


Fig. 3. Segmentation of the scorched-infected strawberry leaf (a) raw leaf image (b) pre-processed leaf image (c) lazy snapping annotation for the scorch-infected area (d) lazy snapping for the whole leaf area (e) scorch-leaf masked region (f) whole-leaf masked region (g) binary masked image of leaf scorched areas (h) binary masked image of leaf whole area

Since there are 18 features, the model might take quite some time to predict the correct output, hence, only significant features should be considered for the model development. The most significant leaf features were selected by using a hybrid approach. It is done by using neighborhood and principal component analysis (NCA-PCA). PCA will give the number of significant features that need to be included. On the other hand, NCA is used to rank the features based on a particular feature's variability and impact on the prediction results. This process gave out a three-feature vector of R, G, and a*. After extracting the important features, they were further improved by applying data normalization. This step is implemented to promote proper scaling of data and, to prevent the unequal scattering of data. The normalization method, Z-score data normalization [17], was used. It is primarily based on getting column vector data's mean and standard deviation (Fig. 4). To do this, the data (x) of a certain feature (e.g., contrast) is subtracted by the mean derived from that column vector (\bar{x}). Lastly, the value is divided to the standard deviation of that column vector (σ).

$$Normalized\ value = \frac{x - \bar{x}}{\sigma} \quad (2)$$

D. Strawberry Leaf Health Status Identification using Machine Learning Model

For the detection task, machine learning algorithms such as support vector machines (SVM), Naïve Bayes (NB), classification tree (CTree), and linear discriminant analysis (LDA) were considered. Optimization parameters for each algorithm are as follows. SVM has a box constraint (1.6545), kernel (3.507), bias (1.0384), and solver (sequential minimal

optimization). NB has a 0.0512 kernel and cosine distance. LDA was configured with a delta of 0.0293 and gamma of 0.6766. Also, Bayesian optimization was used for the hyperparameter tuning for all the machine learning algorithms mentioned. To make the classifier model robust and unbiased, the image dataset was partitioned into three subparts: training data (56%), validation data (24%), and testing data (20%). This data partitioning technique is called stratified sampling in statistical jargon. The purpose of this step is to lower the error rate in the classification task of the model.

The assessment of each of these models will be based on the following criteria: accuracy, inference time, Matthew's correlation coefficient (MCC), precision, recall, fall-out, and hamming loss.

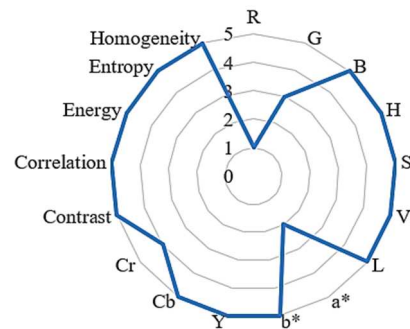


Fig. 4. Characterization of color-texture-morphological features based on NCA-PCA feature selection

E. Objective Function Creation using Genetic Programming

One of the ways to create an objective function for unconstrained optimization problems is using genetic programming (GP). It is a type of computational evolutionary algorithm that combines the use of a regression tree and a genetic algorithm. This union of two powerful algorithms is also known as multiple-gene symbolic regression. This is used to define the mathematical model to be minimized by the optimization algorithms to define the best possible combination of a three-layered neural network which will be used to predict the infection rate that a strawberry leaf contains. GPTIPsv2, a genetic programming tool that has built-in compatibility with MATLAB, is utilized to generate the objective function. The hyperparameters used are as follows: population size of 50, maximum generations of 50, training instances of 25, tournament size of 50, elite fraction of 0.1, Pareto tournament probability of 0.2, maximum genes of 10, maximum tree depth of 5, infinite number of total nodes, crossover probability of 0.84, mutation probability of 0.14 and expressional complexity measure.

The fitness function (3) is based on three input variables, represented by N_1 , N_2 , and N_3 for each layer of a three-layered neural network. The output parameter is a single variable that will provide the root mean square error (RMSE) value which will help the optimization algorithms figure out the right combination of neurons in consideration with the smallest RMSE value.

$$\text{Root Mean Square Error} = f(N_1, N_2, N_3) \quad (3)$$

F. Population-based Optimization of Neurons in Recurrent Neural Network

1) Genetic Algorithm (GA)

Genetic algorithm (GA) is one of the well-known bio-inspired algorithms recognized in the field of artificial intelligence. It is derived from the biological evolution theory called Darwinism, which is primarily based on population growth reproduction that introduces the concept of gene combination and mutation. The hyperparameters are as follows: population size is 50, the number of generations is 50, and the number of runs is 10. The selection parameters are as follows: tournament size of 50, Pareto tournament probability of 0.2, the mutation rate of 0.1, crossover rate of 1, constraint tolerance value of $1e^{-6}$. Moreover, a nonlinear constraint algorithm and selection roulette were implemented. The fitness limit is set to zero with a functional tolerance of $1e^{-6}$. The maximum generation allowed is set to 100.

2) Cuckoo Search Algorithm (CSA)

The cuckoo search algorithm (CSA) is a type of bio-inspired algorithm that mimics the behavior of cuckoo birds in population reproduction [18]. This algorithm utilizes the Levy flight theory to generate a new candidate solution (eggs). The concept is based on the way some particular cuckoo bird species engaged in obligate brood parasitism. The way they do it is by laying their eggs on the nest of a different species of bird (host bird). They simply rely on other host birds to take care of their offspring. The algorithm relies on the fact that there is at least a 10% probability that the host bird will discover that the cuckoo bird's egg is not of its own and thus, has a two-choice: to remove the cuckoo egg from its nest or carry its eggs (excluding the cuckoo egg) and bring it to a newly created nest. The survivability of the cuckoo bird solely depends on their eggs not being discovered by the host bird and, for the host bird to consider the cuckoo egg as its own. The ideal scenario is that the cuckoo egg is not discovered and, the cuckoo egg will hatch and grow together with the host bird's real offspring. Here, the population size for the algorithm is set with a value of 20, the discovery rate probability of alien eggs (or solutions) of 0.25, and the maximum iteration of 100 as the convergence criteria. The population stated is the number of host nests per generation. This means that a nest can contain only one cuckoo egg.

3) Artificial Bee Colony Algorithm (ABC)

The artificial bee colony algorithm (ABC) is a swarm-based metaheuristic algorithm that is primarily based on the honeybees' intelligent foraging behavior [19]. The algorithm can be explained on the principle of how bees coordinate with other bees to find the best possible food source they could find that is close to their beehive. The main components of this algorithm are food sources, employed foragers, and unemployed foragers (scout bees and onlooker bees). In selecting a food source, forager bees evaluate several

properties related to the food source such as nectar taste, energy richness, the difficulty of energy extraction, and closeness to the beehive. The hyperparameters for the implementation of the algorithm are as follows: food source is 100, maximum iteration of 100, the population size of 50, and a limit of 150 for the scouting phase.

G. Strawberry *Diplocarpon earlianum* Leaf Region Infection Prediction using Computational Intelligence

The leaf region infection rate (LRIR) is numerically computed using the developed recurrent neural network (RNN) with three types of optimization methods: GA-RNN, CSA-RNN, and ABC-RNN. The RNN model was built through MATLAB by using the built-in function *newelm*. For the configuration of this network, the training algorithm used is scaled conjugate gradient (SCG), 10,000 iterations (epoch), goal parameter is set to $1e^{-7}$, and the transfer function designation was 'tansig', 'purelin', and 'log'.

To evaluate the computational intelligence models, parameters such as root mean square (RMSE), coefficient of determination (R^2), and mean absolute error (MAE) were recorded. An ideal value of 1 for R^2 is desired while a value of 0 is targeted for MAE and RMSE.

III. RESULTS AND DISCUSSION

A. Strawberry Leaf Health Status Identification using Machine Learning Models

It can be seen in the confusion matrix summary in Fig. 5 that SVM, NB, and CTree were able to classify the scorched strawberry leaf from the healthy ones with 100% accuracy. LDA has the lowest accuracy of 97.37%. In terms of the amount of inference time, CTree outperformed all other three algorithms. CTree is 66.53% quicker compared to SVM, 39.80% quicker compared to NB, and 9.62% faster than LDA. This result is based on the testing phase of the model developed with the use of stratified sampling. It is quite noticeable that the reduction of features used resulted in much less computational time as seen in Table 1. By doing so, the improved performance of the models was able to recognize *Diplocarpon earlianum* leaf infection with much lesser time and computational cost without sacrificing the level of accuracy.

B. Objective Function Based on Genetic Programming

To create a robust model, function settings for the GPIPSv2 were set to times, plus, minus, square, sqrt, sin, cos, add3, mult3, cube, log, abs, and neg. These were utilized to generate a function. The parameters for the programming process were set to 0.1, 0.84, and 0.14 for the ERC probability, crossover probability, and mutation probability respectively.

The derived fitness functions were ranked based on their coefficient of determination (R^2) in descending order. There were 500 models generated in this process. Seventeen models were shown based on highest R^2 values (model 351, 370, 380, 379, 378, 262, 276, 294, 257, 192, 185, 285, 12, 37, 437, 72, and 87). Model 351 fitness function was the one chosen among other models based on its R^2 value of 0.939. The equation presented below in (2) with the use of variable N

represents the number of neurons needed. Additionally, the subscripts (1, 2, and 3) added under the variable N specify what layer is being identified.

$$RMSE = 0.00407N_1 + 0.0223N_2 + 0.0262N_3 + 9.42e-6N_3^2N_2^2 - 0.00112N_1N_2 + 6.34e-4N_1N_3 - 0.00497N_2N_3 + 1.98e-5N_1N_2^2 - 9.96e-6N_2N_3^2 - 8.49e-5N_2^2N_3 - 1.78e-6N_2^3N_3 + 4.24e-5N_1^2 + 0.00466N_2^2 - 9.86e-5N_2^3 - 0.00521N_3^2 + 6.46e-7N_2^4 + 9.42e-6N_3^4 - 1.63e-5N_1N_2N_3 - 0.0209 \quad (4)$$

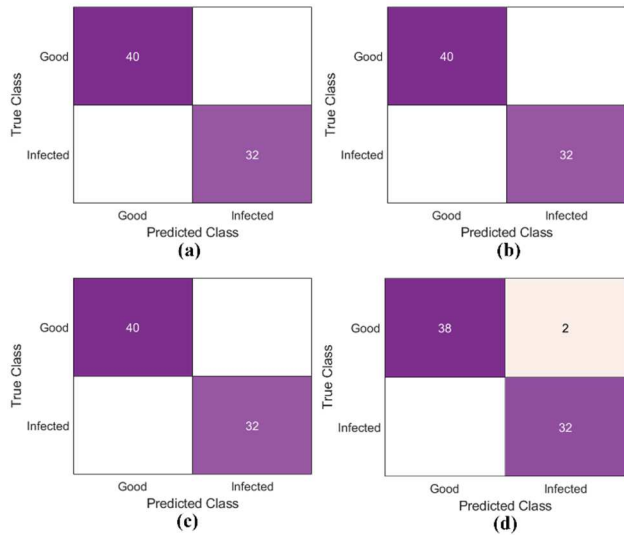


Fig. 5. Confusion matrix data summary for the testing phase of detecting strawberry Diplocarpon Earlianum leaf infection three significant features (a) SVM (b) NB (c) CTree (d) LDA

C. Recurrent Neural Network Configuration based on Optimization Results

Based on the optimization results, GA-RNN has a three-layered-hidden-layer combination of 95.9915, 32.3209, 13.9978 (rounded off to 96, 32, 14), CSA-RNN has a three-layered-hidden-layer combination of 109.9901, 40, 15.2434 (rounded off to 110, 40, 15), and ABC-RNN has a three-layered-hidden-layer combination of 96.0986, 33.3731, 16.4391 (rounded off to 96, 33, 16). The fitness value for the GA-RNN is -0.0017, CSA-RNN has -0.5729, and ABC-RNN has -0.5605. The fitness curve of each algorithm is presented in Fig. 6, Fig. 7, and Fig. 8.

D. Strawberry Diplocarpon Earlianum Leaf Region Infection Prediction using Computational Intelligence

A three-layered recurrent neural network (RNN₃) is responsible for the quantitative prediction of the amount of prediction. The subscript of three indicates the number of layers used inside this type of neural network. Based on Table 2, it can be noted that the ABC-RNN₃ has the lowest MAE and RMSE for the testing phase. It also has the highest R² value which is closer to 1. It bested all other machine learning models used in this study. To compare ABC-RNN₃ has the highest R²: 1.328% higher than GA-RNN₃, and 3.834% higher than CSA-RNN₃. Additionally, ABC-RNN₃ has the lowest RSME: 11.238% lower than GA-RNN₃, and 19.490%

lower than CSA-RNN₃. Moreover, ABC-RNN₃ has the lowest MAE: 0.584% lower than GA-RNN₃, and 9.415% lower than CSA-RNN₃. Among the developed models, CSA-RNN₃ performed the least while GA-RNN₃ is the second-best among the three algorithms in predicting the leaf region infection rate (LRIR) of a *Diplocarpon earlianum* infected strawberry leaf.

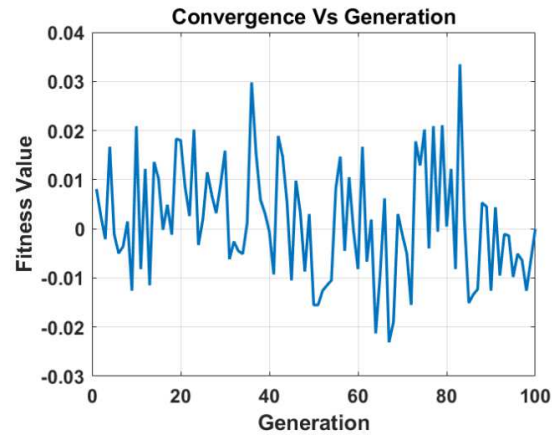


Fig. 6. Fitness curve for GA-RNN optimization for the optimal three-layered neural network combination

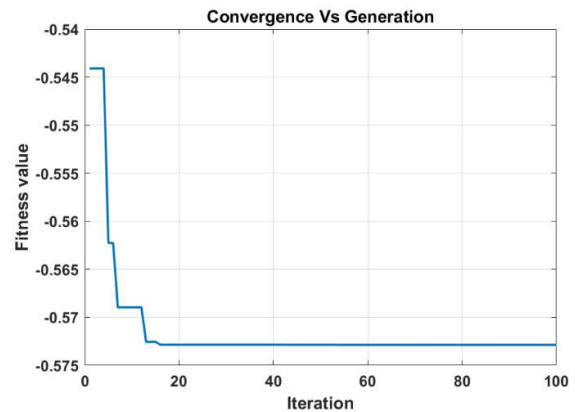


Fig. 7. Fitness curve for CSA-RNN optimization for the optimal three-layered neural network combination

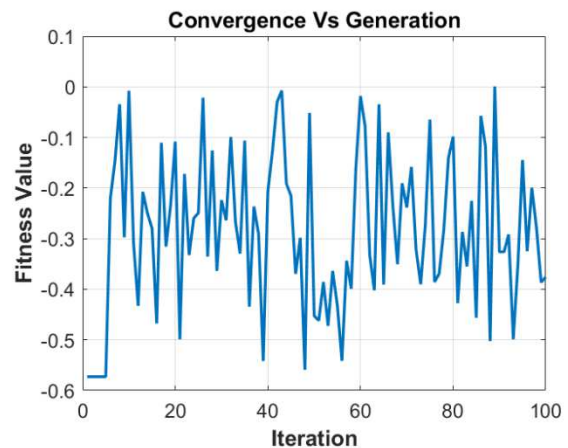


Fig. 8. Fitness curve for ABC-RNN optimization for the optimal three-layered neural network combination

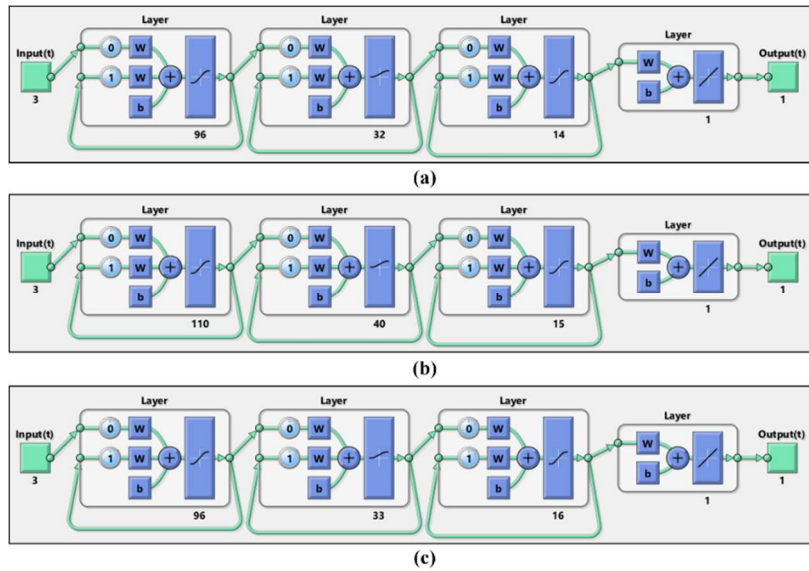


Fig. 9. Recurrent neural network model (a) GA-optimized RNN model (b) CSA-optimized RNN model (c) ABC-optimized RNN model

TABLE I. EVALUATION SUMMARY FOR STRAWBERRY SCORCHED LEAF CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Model	Training	Validation	Testing								
	Accuracy	Accuracy	Accuracy	Fall-out	Precision	Specificity	Recall	F1-score	MCC	Hamming Loss	Inference Time (s)
SVM	100.000	100.000	100.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	44.058
NB	0.985	100.000	100.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	24.495
CTree	100.000	100.000	100.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	14.746
LDA	98.100	99.010	97.370	0.028	0.972	0.974	0.976	0.974	0.947	0.027	16.316

TABLE II. EVALUATION SUMMARY FOR PREDICTING THE STRAWBERRY LEAF REGION INFECTION RATE USING INTELLIGENT MODELS

Model	Training			Validation			Testing		
	RMSE	R ²	MAE	RMSE	R ²	MAE	RSME	R ²	MAE
GA-RNN ₃	1.990	1.000	3.728	0.037	1.000	0.005	2.274	0.936	0.919
CAS-RNN ₃	0.033	1.000	0.010	0.026	1.000	0.005	2.562	0.912	0.914
ABC-RNN ₃	0.000	1.000	0.000	0.029	1.000	0.005	2.063	0.948	0.828

Leaf scorch detection and infection rate prediction were explored to develop a vision-based approach to identifying a strawberry leaf's health condition and quantifying the amount of severity the *Diplocarpon earlianum* fungus had caused. Strawberry leaf health status was classified as good or infected with the help of machine learning models such as SVM, NB, CTree, and LDA which are all found to be highly accurate. These machine learning models were also examined by [5] to classify the leaf health status of grapes infected with black rot; [6] on the other hand, used these models to identify a corn leaf that is exposed to *Maize Cercospora* leaf spot; and [7] to classify three different rice leaf diseases. Among the four classification algorithms, CTree showed a 100% accuracy for this study. Using population-based optimizers such as GA, CSA, and ABC, the recurrent neural network was calibrated

to the ideal combination of three-layered hidden neurons for which ABC-RNN₃ gave out the highest R² value and lowest RMSE and MAE values. Given the reports from [9] that explored bio-inspired optimization with the firefly algorithm, exploring these population-based optimizers opens more options for innovation of computationally inexpensive intelligent models that plant pathologists can consider when it comes to quantitative assessment of leaf infection rate in strawberry leaves. Extracting and selecting the top three features which are R, G, and a* proves that among morphological, textural, and spectral traits, the most significant leaf phenotype that is directly relevant in determining a scorch-infected strawberry leaf is the spectral feature. Using the combination of NCA-PCA, which was also explored by [7, 13], for a contrast-improved strawberry leaf

image, the accuracy of the classification results had drastically improved while the computational cost was lessened, which is beneficial to farmers looking to improve their farming practices without spending too much money. Reducing the features needed for the model development significantly decreases the computational time and cost as observed by the results comparing an eighteen-feature vector to a three-feature vector [8], an eighteen-feature vector to a two-feature vector [13], and a twenty-two-feature vector to a seven-feature vector [14] respectively. Thus, the NCA-PCA-CTree₃-ABC-RNN₃ hybrid algorithm configuration provides a cost-effective approach for strawberry leaf scorch detection and infection rate quantitative assessment which can aid farmers in detecting the crop disease at an earlier stage which will significantly reduce widespread bacterial infection across the farm. Additionally, it will also give plant pathologists an idea of how the infection should be treated given that the amount of infection severity is numerically determined.

IV. CONCLUSION

This study proposed a new technique for detecting and predicting the *Diplocarpon earliarum* infection rate of strawberry leaves using recurrent neural network (RNN) integrated with population-based optimizers, namely genetic algorithm (GA), and cuckoo search algorithm (CSA), and artificial bee colony (ABC). The raw images containing scorched strawberry leaves were segmented using graph-cut segmentation through lazy snapping algorithm. Hybrid neighborhood component analysis (NCA) and principal component analysis (PCA) was done to select the 3 most significant morpho-spectro-textural features of leaves resulting in red, green, and a* components. A genetic programming-based RMSE model was constructed as a function of neuron density on each of the three RNN layers. After a series of explorations in the hyperparameters of GA-RNN, CSA-RNN, and ABC-RNN, the ABC-RNN model with inputs of the 3 most significant features outperformed other models in performing non-invasive leaf region rate infection with an accuracy of 94.8%. Four classification machine learning models were configured in which the classification tree (CTree) bested other detection models with an accuracy of 100%. Hence, this study was able to introduce a technique of integrating NCA, PCA, CTree, ABC and RNN for crop disease detection and infection rate prediction. For plant disease assessment, it was proven to be more efficient and labor cost-effective than manual disease inspection methods. For future studies, the inclusion of other strawberry leaf disease variants caused by other fungal bacteria is highly recommended to expand the range of diseases that can be determined by the model. Moreover, applying the developed model to an actual strawberry farm for leaf health analysis as extended research to further improve the model's efficiency is also suggested.

ACKNOWLEDGMENT

The authors would like to express their sincerest gratitude for the support bestowed by Engineering Research and Development for Technology (ERDT) of the Department of Science of Technology (DOST) of the Philippines as well as

the De La Salle University - Intelligent Systems Laboratory (DLSU-ISL), Manila, Philippines.

REFERENCES

- [1] S. Kaur, S. Pandey, and S. Goel, "Plants Disease Identification and Classification Through Leaf Images: A Survey", Archives of Computational Methods in Engineering, 2018, <https://doi.org/10.1007/s11831-018-9255-6>
- [2] G. Dhingra, V. Kumar and H. D. Joshi, "Study of digital image processing techniques for leaf disease detection and classification," Multimedia Tools and Applications, 2017, <https://doi.org/10.1007/s11042-017-5445-8>
- [3] A. Cruz et. al., "Vision-Based Plant Disease Detection System Using Transfer and Deep Learning," American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Spokane, Washington, July 2017, <https://doi.org/10.13031/aim.201700241>
- [4] C. Sass, "The health benefits of Strawberries", January 8, 2020, Available: Health, <https://www.health.com/nutrition/health-benefits-of-strawberries> [Accessed: March 1, 2022]
- [5] O. J. Alajas, R. Concepcion, E. Dadios, E. Sybingco, C. H. Mendigoria, and H. Aquino, "Prediction of Grape Leaf Black Rot Damaged Surface Percentage Using Hybrid Linear Discriminant Analysis and Decision Tree," 2021 International Conference on Intelligent Technologies (CONIT), 2021.
- [6] R. Concepcion, E. Dadios, J. Alejandrino, C. H. Mendigoria, H. Aquino, and O. J. Alajas, "Diseased Surface Assessment of Maize Cercospora Leaf Spot Using Hybrid Gaussian Quantum-Behaved Particle Swarm and Recurrent Neural Network," 2021 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2021.
- [7] C. H. Mendigoria, R. Concepcion, A. Bandala, O. J. Alajas, H. Aquino and E. Dadios, "OryzaNet: Leaf Quality Assessment of Oryza sativa Using Hybrid Machine Learning and Deep Neural Network," 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2021, pp. 1-6, doi: 10.1109/HNICEM54116.2021.9731957.
- [8] R. Concepcion, E. Dadios, J. Cuello, A. Bandala, E. Sybingco, and R. R. Vicerra, "Determination of Aquaponic Water Macronutrient Concentrations Based on Lactuca Sativa Leaf Photosynthetic Signatures using Hybrid Gravitational Search and Recurrent Neural Network," Walailak Journal of Science and Technology (WJST), vol. 18, no. 10, 2021.
- [9] R. Concepcion and E. Dadios, "Bioinspired Optimization of Germination Nutrients Based on Lactuca sativa Seedling Root Traits as Influenced by Seed Stratification, Fortification and Light Spectrums," AGRIVITA Journal of Agricultural Science, vol. 43, no. 1, pp. 222-232, 2021.
- [10] R. Concepcion, S. Lauguico, J. Alejandrino, J. de Guia, E. Dadios, and A. Bandala, "Aquaphotomics determination of total organic carbon and hydrogen biomarkers on aquaponic pond water and concentration prediction using genetic programming," 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), 2020.
- [11] R. Concepcion, S. Lauguico, R. R. Tobias, E. Dadios, A. Bandala, and E. Sybingco, "Genetic algorithm-based visible band tetrahedron greenness index modeling for lettuce biophysical signature estimation," 2020 IEEE REGION 10 CONFERENCE (TENCON), 2020.
- [12] J. Alejandrino, R. Concepcion, S. Lauguico, R. R. Tobias, V. J. Almero, J. C. Puno, A. Bandala, E. Dadios, and R. Flores, "Visual classification of lettuce growth stage based on morphological attributes using unsupervised machine learning models," 2020 IEEE REGION 10 CONFERENCE (TENCON), 2020.
- [13] C.H. Mendigoria, R. Concepcion II, E. Dadios, H. Aquino, O.J. Alajas and E. Sybingco, "Vision-based Postharvest Analysis of Musa Acuminata Using Feature-based Machine Learning and Deep Transfer Networks," 9th IEEE Region 10 Humanitarian Technology Conference (R10 HTC 2021), in press

- [14] C. H. Mendigoria, H. Aquino, O. J. Alajas, R. Concepcion, E. Dadios, E. Sybingco, A. Bandala, and R. R. Vicerra, "Varietal classification of Lactuca Sativa seeds using an adaptive neuro-fuzzy inference system based on morphological phenes," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 25, Issue 5, pp. 618-624, September 2021.
- [15] O. J. Alajas, R. Concepcion, R. R. Vicerra, A. Bandala, E. Sybingco, E. Dadios, J. Cuello and V. Fonseca, "Indirect Prediction of Aquaponic Water Nitrate Concentration Using Hybrid Genetic Algorithm and Recurrent Neural Network" 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication, and Control, Environment, and Management (HNICEM) 2021.
- [16] J. Arun Pandian and G. Geetharamani, "Data for Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network", *Mendeley Data*, V1, 2019.
- [17] S. H. Javaheri, M. M. Sepehri, and B. Teimourpour, "Chapter 6 - Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection," in *Data Mining Applications with R*, Academic Press, Pages 153-180, 2014, ISBN 9780124115118, <https://doi.org/10.1016/B978-0-12-411511-8.00006-2>.
- [18] X. -S. Yang and Suash Deb, "Cuckoo Search via Lévy flights," 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), 2009, pp. 210-214, DOI: 10.1109/NABIC.2009.5393690.
- [19] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization; artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, 39(3), pp. 459-471 (2007). <https://doi.org/10.1007/s10898-007-9149-x>.

Lower-Limb Exoskeleton Systems for Rehabilitation and/or Assistance: A Review

Dana Terrazas-Rodas
 Facultad de Ingeniería
 Universidad Tecnológica del
 Perú
 Lima, Perú
dterrazas@ieecc.org

Lisbeth Rocca-Huaman
 Facultad de Ingeniería
 Universidad Tecnológica del
 Perú
 Lima, Perú
lisbeth.rocca.huaman@ieecc.org

César Ramírez-Amaya
 Facultad de Ingeniería
 Universidad Tecnológica del
 Perú
 Lima, Perú
ramirezamaya@ieecc.org

Angel E. Alvarez-Rodriguez
 Facultad de Ingeniería
 Universidad Tecnológica del
 Perú
 Lima, Perú
angel_alro2025@ieecc.org

Abstract— Disability is a condition that directly affects the natural performance of human abilities such as movement due to diseases such as cerebrovascular accident (CVA), spinal cord injury (SCI), among others. According to the World Health Organization (WHO), more than a billion people, about 15% of the world's population, experience at least one form of disability caused by aging and the onset of multiple diseases. Lower limb exoskeletons are a set of links connected by joints that are used for multiple applications, including rehabilitation and/or assistance, especially for the elderly. In this recent review, several lower limb exoskeletons are analyzed in categories that are common requirements to propose conceptual designs of new exoskeletons such as the commercial name, the degrees of freedom (DoF), the mechanism that covers the part of the lower limb, the biosignal that controls the functions of the exoskeleton of the lower limbs, the characteristics of the test subjects that participated in the clinical evaluations, the pathology of the end users or potential clients, the application for which they are being designed and the Technological Readiness Level (TRL) which is a NASA measurement scale.

Keywords—*Lower-limb, exoskeletons, rehabilitation, assistance, hip, ankle, knee, biomechatronic, biomechanics*

I. INTRODUCTION

Disability is a condition that directly affects the natural performance of human capacities such as vision, movement, thinking, memory, learning, hearing, mental health and social relationships due to diseases such as paralysis brain, Down syndrome, among others. According to the World Health Organization (WHO), more than one billion people, about 15% of the world's population, experience at least one form of disability caused by aging and the onset of multiple diseases [1]. In 2014, the Economic Commission for Latin America and the Caribbean (ECLAC) determined that more than 70 million people live in conditions of disability [2]. In 2012, the National Specialized Survey on Disability (ENEDIS) and National Institute of Statistics and Informatics (INEI) identified that 5.2% manifested some type of disability in Peru [3]. Due to the COVID-19 pandemic, people with disabilities have become the most affected population due to their economic situation, age, gender, others. Also, although there is no evidence that shows the profile of people who died during the COVID-19 pandemic, it is estimated that people with disabilities were affected [4].

Physical disability is caused by multiple factors which are presented below. The dysfunction of the physical capacity of the lower limbs can be caused by genetic, accidental or acquired causes over time. The most frequent causes are pathologies such as stroke, spinal cord injury (SCI), traumatic brain injury (TBI), cerebrovascular accident (CVA), cancer, amputation, musculoskeletal injury or neuromuscular disease. These conditions are the main cause of the deficiencies recorded in patients who have reduced or no mobility in their lower limbs. [5]. People with this type of disability in the lower extremities show the consequences of their condition when they try to carry out their Activities of Daily Living (ADL). Besides, people with disabilities are economically affected, the majority being the economically inactive population. For this reason, it is people with disabilities who have the highest unemployment rates. Unfortunately, those who manage to get a job seek minimum remuneration or are informal jobs [6]. In addition, it has been recorded that households that have a person with a disability tend to have the highest expenses because basic needs include services such as rehabilitation therapies, personalized education, medications, maintenance of medical equipment used for transportation or medical assistance, care service and among others [3]. Given the above, the technical-scientific community has developed multiple devices as possible solution mechanisms such as mechatronic systems called exoskeletons. Exoskeletons are systems that consist of an electronic, mechanical and control part. These devices are what are commonly known as robots. The exoskeletons are a set of links connected by joints that are capable of executing different actions assigned by the actuators that compose it. José L. Pons [7] in the book *Wearable robots: biomechatronic exoskeletons* originally defines exoskeletons as extensors because they are considered a type of robot whose function is to provide force as support to the human body to carry out activities where the natural capacities of the body are not enough to do them. Therefore, José L. Pons. [7] defines the exoskeleton as a structure that corresponds to the anatomy of the human body to supply external force [7]. These devices are designed for areas such as the military, work and even health. In this last aspect, exoskeletons have positioned themselves in rehabilitation and assistance related to the improvement of motor capacity, being the lower limb exoskeletons most studied by researchers.

II. BIOMECHANICS OF LOWER-LIMB

By evaluating the relationship between each movement of a living organism, and the energy expenditure involved in the forces to generate that movement, we can say that we are fulfilling the objective of the study of biomechanics. This analysis includes the study of the characteristics of human movement from the point of view of kinematics, using parameters such as velocity and direction, and kinetics, observing how the forces applied inside and outside the body generate the movement [8]. The lower limbs are sequentially connected to the trunk of the body through the pelvic girdle, maintaining the weight load of the trunk and upper limbs, as well as to a constant generation of forces by the contact between the foot and the ground. The lower limb can be subdivided into three main parts, the thigh with its hip joint, the leg with its knee joint and the foot with its ankle joint; each with different degrees of freedom and types of movements as seen in Fig. 1.

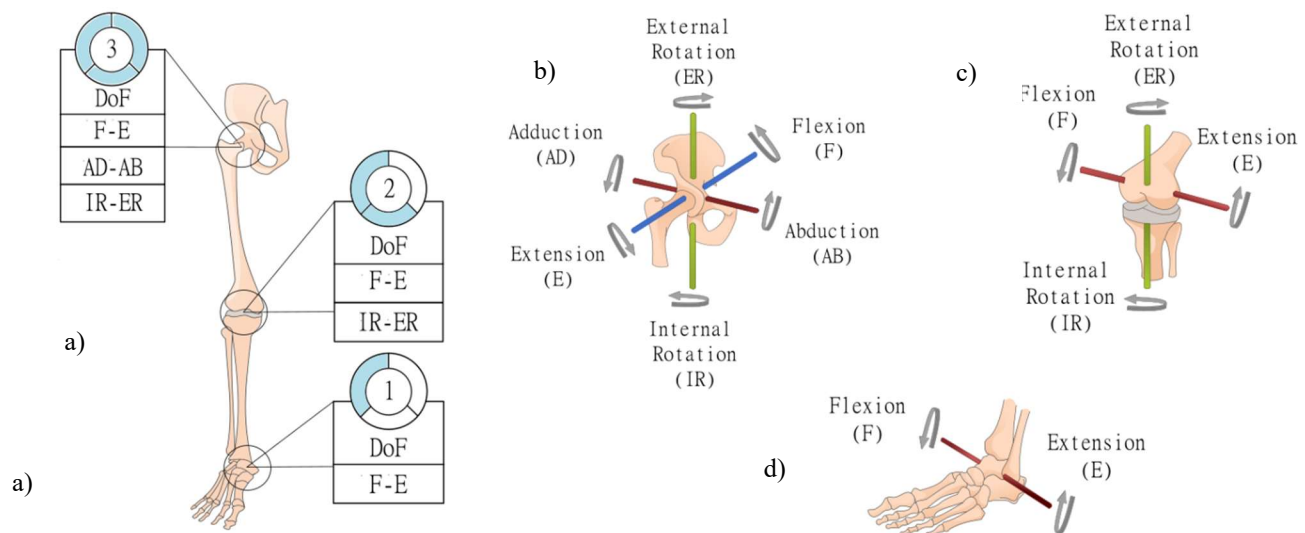


Fig.1. Biomechanics of Lower-Limb. a) Lower Extremity [10] b) Hip [10] c) Knee [10] d) Ankle [10]

A. Biomechanics of the Hip

The hip joint exhibits triaxial motion caused by its coxofemoral joint and acetabulum of the coxal. This type of joint is commonly called enarthrosis. The femur and shaft join the hip at an angle called the cervicodiaphyseal. Its common value varies between 125° and 135° in a healthy adult, increasing to 145° in the first years of life and decreasing to approximately 120° in old age. On the other hand, the hip has 3 DoF, as seen in Fig. 1. First, flexion and extension allow the thigh to move in the sagittal plane. Hip flexion with the knee extended can reach 90° while with the knee flexed it varies between 140° and 160°. In contrast, hip extension with the knee flexed forms an angle of 10° and with the knee extended it reaches 20°. The second one is called the abduction-adduction movement which has an amplitude of 15° in the normal state. Abduction can be increased by 90° by changing the position of other joint movements. The last one is the internal-external rotation where the amplitude is 90° distributed in 60° for external rotation and 30° for internal rotation [9].

B. Biomechanics of the Knee

The knee is a joint that connects the leg with the thigh, and allows cushioning loads between the upward forces of the ground and the weight of the body. This joint fulfills the function of giving stability and mobility to the body; internally it has three joints that involve the femur, the tibia, the fibula and the patella at the bony level, among them it is found the patellofemoral, the tibioperoneal and the most representative, the tibiofemoral, since the movement that it allows is the most notorious in the knee. When aligning the lower extremity, a flat angle between the leg and thigh is not achieved; it generally ranges between 170° and 175°. The knee has 2 DoF as seen in Fig.1. It allows flexion and extension movements on a lateromedial axis where flexion is 135° and extension is 0°. The other movement is internal and external rotation which is only performed when the knee is in semi-flexion which is 45° to 50° for external rotation and 30° to 35° for internal rotation [10].

C. Biomechanics of the Ankle

The ankle joint is located within the structure that connects the foot to the leg. This joint allows, like the others, to perform various movements that enable the gait cycle of the human lower extremity, which is recorded by a motion capture system to analyze the biomechanics of the ankle. Although at first glance it appears that it allows a great variety of movements, the truth is that it only has 1 DoF. As seen in Fig.1, this joint is commonly called the tibiotarsal. It involves contact of the tibia and fibula with the tarsus and provides stability during specific movements. To understand the movement of the ankle, it is important to correctly locate its axis of rotation. This part of the ankle joint crosses the malleoli obliquely to the tibia and forms an angle of 8° with the horizontal plane and 6° with the frontal plane. On the one hand, the pair of movements that the ankle allows are flexion and extension, although for the ankle it is usually called flat flexion and reaches an amplitude of 50°. On the other hand, the ankle joint is capable of dorsiflexion reaching 20° [10].

III. LOWER-LIMB EXOSKELETONS

Currently, lower limb exoskeletons are mechatronic systems that usually work with parts of the human body such as the hip, knee and/or ankle, which are frequently used for assistance and/or rehabilitation. This type of exoskeletons were classified in Table I. in the following categories: commercial name or references, the Degrees of Freedom (DoF) that allow identifying the Range of Motion (RoM), the mechanism that covers the selected part of the lower limb, also considered the biosignal used to control the actions of the exoskeletons,

characteristics of the test subjects who participated in the clinical evaluations to assess the effectiveness and feasibility of the exoskeleton, the pathology of the users for which each of the exoskeletons is intended, application and Technology Readiness Level (TRL) according to NASA [11]. The last one indicates the level of maturity of the lower exoskeleton system, which are levels from 1 to 9 where the first means that it is starting as research and the last means that they have already approved as commercial devices through organizations such as the Food and Drug Administration (FDA) [12] and others.

TABLE I. LOWER-LIMB EXOSKELETONS

Name / Ref.	DoF	Mechanism	Biosignal	Characteristics of test subjects	Pathology	Application	TRL
Shepherd <i>et al.</i> [13]	-	K	EMG	3 WD	CVA	STS (As)	5/6
EICOSI [14]	1	K	EMG	1 M, WD, 35 years, 94 kg, 1.82 m	F/E (K)	As, R	3
WAKE-Up [15]	-	K, A	-	4 TDC, 3 CP	CP, He	Gt (R)	5
HUMA [16]	12	H, K, A	EMG	-	F/E	As	6/7
Kim <i>et al.</i> [17]	3	H, K	EMG	3 M, WD	RM	As	2/3
Lerner <i>et al.</i> [18]	-	K	EMG	7 CP, 5-19 years	CP	Gt (As)	5
Monaco <i>et al.</i> [19]	42	H, K, A	-	-	LoB	As	1
XoSoft [20]	-	H, K, A	-	-	CVA, SCI	As	2
Lerner <i>et al.</i> [21]	30	K, A, Fo	EMG	1 M, CP, 6 years	CP	F/E(K) (As)	5
Torrealba <i>et al.</i> [22]	-	K	-	-	CVA, SCI	Gt (R)	2
ExiO-A [23]	2	H, K	-	1 macaque	-	Gt (R)	5
ANKUR-LL II [24]	3	H, K, A	-	-	-	R	4
E-ROWA [25]	10	H, K, A	EMG	3 M, 1 F	-	Gt (As)	5/6
Khamar <i>et al.</i> [26]	1	K	-	2M, 1F	MD	F/E(K) (As)	2
ARKE [27]	6	LE	-	30 WD, 5 SCI	SCI	Gt (As)	2/3
Kang <i>et al.</i> [28]	-	H	-	7 M, 3 F, WD	F/E (H)	Gt (As)	5/6
Khazoom <i>et al.</i> [29]	-	A	-	8 M, 1 F	PF (A)	Gt (As)	6
Yandell <i>et al.</i> [30]	19	A	EMG	2 M	PF (A)	As	5/6
Gasparri <i>et al.</i> [31]	10	A	EMG	1 WD, 2 CP	PF (A)	Gt (As)	6
Zhang <i>et al.</i> [32]	-	H	EMG	1 M, 32 years, 78kg, 1.78m	Pa, RM	Gt (As)	1
Indego [33]	-	H, K	EMG	5 WD	LoB	Gt (As)	1
Yang <i>et al.</i> [34]	-	LE	-	-	CVA	R	1
KAD [35]	-	K	EMG	3 subjects	-	Gt (As)	3
BioKEX-II [36]	-	K	EMG	12 WD	RM	STS (As)	4
WAXO [37]	-	A	EMG	1 M, 60kg, 1.60m	-	Gt (As)	5
Sado <i>et al.</i> [38]	12	LE	EMG	5 subjects	MD	Gt (As)	5/6
Martini <i>et al.</i> [39]	-	H	-	8 M, 2 F, WD	CVA, SCI	Gt (R)	3
Amiri <i>et al.</i> [40]	4	LE	-	-	CVA	R	1
Vinoj <i>et al.</i> [41]	6	LE	EEG	4 WD, 2 CP	CVA	Gt (R)	6
LOPES III [42]	-	LE	EMG	3 M, 1 F	LoB	Gt, STS (As)	5
Amiri <i>et al.</i> [43]	2	LE	-	-	CVA	R	1

IEMTRONICS 2022 (International IOT, Electronics and Mechatronics Conference)

Chen <i>et al.</i> [44]	7	LE	-	4 subjects, 23-29 years	-	Gt (As)	3
Chairless Chair [45]	-	LE	EMG	45 M, WD	MD	As, R	7/8
Al-Ayyad <i>et al.</i> [46]	1	L	-	1 M, WD	F/E (K)	Gt (R)	5
Ding <i>et al.</i> [47]	-	LE	-	9 M, WD	RM	Gt (As)	3
Li <i>et al.</i> [48]	-	LE	-	1 WD, 25 years, 63 kg, 173 cm	Pa, SCI	Gt (R)	5
Campbell <i>et al.</i> [49]	6	H, K, A	-	-	CVA	Gt (R)	3
Orekhov <i>et al.</i> [50]	-	A, Fo	EMG	6 WD	CP	Gt (As)	4
Wang <i>et al.</i> [51]	-	LE	-	-	-	Gt (As)	4
Chen <i>et al.</i> [52]	3	H, K, A	EMG	-	CVA	Gt (R)	3/4
LLE-RePA [53]	12	H, K, A	-	-	CVA	Gt (R)	3/4
EICOSI [54]	5	H, K	EMG	-	CVA, SCI	Gt (R)	6
Zhou <i>et al.</i> [55]	2	H, K	EMG	1 M, 22 years, 67kg, 1.78m	CVA, SCI	Gt (As)	3/4
Pan <i>et al.</i> [56]	-	LE	-	Amputated	RM	Gt (R)	4/5
Llorente-Vidrio <i>et al.</i> [57]	6	H	EMG	-	RM	Gt (R)	2
Yang <i>et al.</i> [58]	2	H, K, A	EMG	-	CVA	Gt (R)	1
ADAMS [59]	-	LE	-	-	SCI	Gt (R)	1/2
Pérez-San Lázaro <i>et al.</i> [60]	9	H, K, A	EMG	-	SCI	Gt (R)	4
MAK [61]	1	A, Fo	-	5 WDS, 18-75 years	CVA	Gt (R, As)	7
Atalante [62]	18	LE	-	-	Pa	STS (As)	6
Aguirre-Ollinger <i>et al.</i> [63]	-	K, A, Fo	-	8M, WD	CVA	Gt (As)	5/6
Wearable-Walker [64]	8	H, K, A	-	11M	CVA	As	6
LOPES [65]	3	H, K, A	-	2WD, 3WDS	He, CVA	R	5/6
Molazadeh <i>et al.</i> [66]	-	H, K	-	4M (WD), 1M (WDS)	Pa, SCI	R	5/6
GEMS-H [67]	-	H	-	8M, 4F, WD	CVA	As, R	8
Long <i>et al.</i> [68]	14	H, K, A	-	3 subjects	CVA, SCI	Gt (R)	6
Tan <i>et al.</i> [69]	-	LE	-	8M, WS	-	Gt (As)	3/4
Utah ExoKnee [70]	4	K	-	11M, 3F, WS	-	STS (R)	5/6
Wang <i>et al.</i> [71]	-	H, K	-	1 subject, 58kg, 1.72m	SCI, CVA, P	R	2/3
EHTe [72]	6	H, K, A	EMG	1M, 23 years, 65.23kg, 1.8m	CVA, He	Gt (As)	4
Zheng <i>et al.</i> [73]	12	H, K, A	-	-	-	As	4
Sharma <i>et al.</i> [74]	4	H, K, A	-	1 M, WD, 24 years, 55kg, 1.72m	RM	Gt (R)	3/4
Zhao <i>et al.</i> [75]	2	H, K	-	-	Pa	Gt (R)	4
Symbitron [76]	8	H, K, A	-	1 subject, 85kg, 1.8m	SCI	Gt (As)	6
Lin <i>et al.</i> [77]	5	K, A	EMG	1 M, WD, 78kg, 1.78m	Pa	Gt (As)	3
Sun <i>et al.</i> [78]	2	H, K	-	-	CVA	Gt (R)	4/5
Tian <i>et al.</i> [79]	10	H, K, A	-	-	-	Gt (As)	2
Ezhilarasi <i>et al.</i> [80]	6	H, K, A	-	-	-	As	2
Shi <i>et al.</i> [81]	3	H	-	1 M, 65kg, 1.68m	CVA	Gt (R)	5
Chen <i>et al.</i> [82]	2	H, K	EMG	40 subjects	He	Gt (R)	6
AutoLEE-II [83]	12	H, K, A	-	1 M, 1.5 - 1.85m	SCI	Gt (As)	3/4

Note: Abbreviations: MAK: Marsi Active Knee, K: Knee, A: Ankle, H: Hip, Fo: Foot, LE: Lower Extremity, L: Leg, EMG: Electromyography, EEG: Electroencephalogram, WD: Without Disability, M: Male, F: Female, TDC: Typical Development, CP: Cerebral Palsy, SCI: Spinal Cord Injury, CVA: Cerebrovascular Accident, F/E: Flexion/Extension, He: Hemiplegia, Pa: Paraplegia, RM: Reduced Movement, LoB: Lack of Balance, WDS: With Disabilities, P: Poliomyelitis, MD: Musculoskeletal Disorder, PF: Plantar Flexion, STS: Sit To Stand, As: Assistance, R: Rehabilitation, Gt: Gait.

A. Description

Lower limb exoskeletons are a set of links connected by joints that work with different parts of the human body, depending on the application and pathologies for which they were designed. Fig.2 shows examples of several lower limb exoskeleton systems, such as EICOSI [14], Wake-Up [15], HUMA [16], among others. According to Table I., 63.38% do not have a commercial name. So, the majority identify themselves with the name of the main author. Therefore, 36.62% of them have commercial names such as EICOSI [54], Wake-Up [15], HUMA [16], XoSoft [84], EXiO-A [23], ANKUR-LL II [24], E-ROWA [25], ARKE [27], LOPES III [42], BioKEX-II [36], Symbitron [76], among others. Regarding the degrees of freedom of the exoskeletons, the exoskeleton Monaco *et al.* [19] stands out for having the highest number of DoF. This exoskeleton has 42 DoF covering the hip, knee and ankle joints, although others have only 1 DoF were also found, such as EICOSI [14], Khamar *et al.* [26], Al-Ayyad *et al.* [46] and Marsi Active Knee (MAK) [61], which encompass mechanisms such as the knee, leg, ankle, and foot. These latter exoskeletons are smaller compared to Monaco *et al.* [19] which is larger and encompasses the entire lower limb. It is worth mentioning that the number of DoF most used in the design of exoskeletons are 2 and 6, found in 7 and 6 articles, respectively. On the one hand, within the exoskeletons that have 2 DoFS, 5 of them cover only the hip and the knee, among them we have EXiO-A [23], Zhou *et al.* [55], Zhao *et al.* [75], Sun *et al.* [78], and Chen *et al.* [82]. In addition to the two remaining exoskeletons, Amiri *et al.* [43] and Yang *et al.* [58] cover the entire lower limb. On the other hand, within the exoskeletons that have 6 DoF, 5 of them cover the entire lower limb such as ARKE [27], EHTE [72], Vinoj *et al.* [41], Campbell *et al.* [49], and Ezhilarasi *et al.* [80]. Finally, there are the lower limb exoskeletons that cover only the hip as the application of robotics developed by Llorente-Vidrio *et al.* [57].

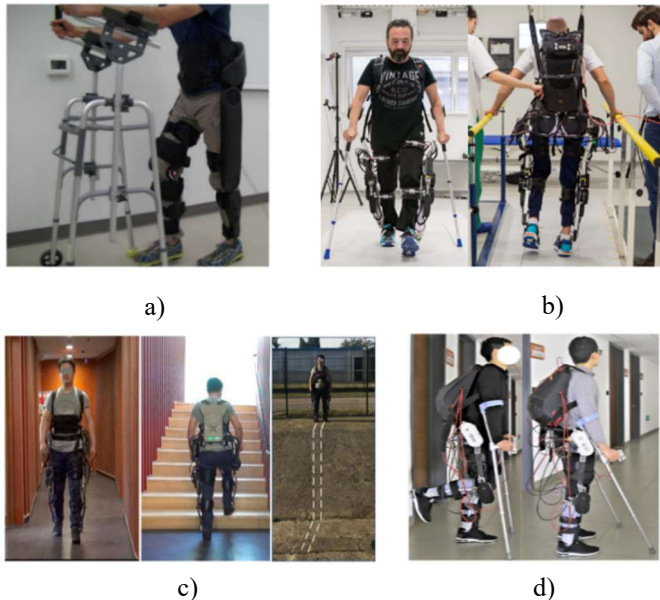


Fig.2. Lower-Limb Exoskeletons. a) Indego [33]. b) Symbitron [76]. c) E-ROWA [25]. d) Li *et al.* [48].

B. Pathologies and Applications

This review identified articles employing 1 to 45 test subjects, generally healthy males (age: 6-40 years, weight: 55-95 kg, height: 1.60-1.82 m) who assisted in the design, revision and fabrication of lower limb exoskeletons. These subjects simulated the physical structure and/or motor function of people who have suffered CVA or SCI that totally or partially affect the lower limbs, generating in them pathologies such as hemiplegia, paraplegia and/or paralysis. In this sense, it was found that 24 of the 71 articles reviewed seek to resolve the pathologies generated by cerebrovascular accidents, and 14, SCI. In addition, pathologies associated with the aging of the population were found, such as reduced movement, problems in maintaining balance, Sit-To-Stand (STS) limitation, among others where 20 articles dealt with the aforementioned pathologies. Of the exoskeletons reviewed, 37 were used for gait assistance, flexion-extension and STS, 30 were used for gait rehabilitation and STS, and 4 of the exoskeletons reviewed were used for both assistance and rehabilitation. The most frequently repeated application is gait rehabilitation and/or assistance, which is presented in 47 articles.

C. Biosignals or Bioelectrical Signals

Biomedical signals are identified by means of transducers that transform the carriers of ionic charges present in the cell membrane, which is found in the skin of the body, into electrical currents that we know as biosignals or bioelectric signal. These are ionic signals, when they are transformed into electrical signals, they are generally positioned on very small voltage scales; that is, they are between microvolts (uV) or millivolts (mV). For application purposes, biomedical signals are conditioned and amplified. In this way, it is possible to use it for different purposes such as the control of a lower limb exoskeleton. According to Table I., it is shown that the most used control signals are the electromyographic (EMG) while only one of them uses the electroencephalogram (EEG) signal. This means that 38.03% of the analyzed exoskeletons prefer to use the EMG signal to control their functions, while 60.53% still do not use biosignals in the proposed devices or choose not to use them yet. Of the 71 exoskeletons analyzed, it was found that only one chose to use the EEG signal.

D. Technology Readiness Levels (TRLs)

According to NASA, the Technology Readiness Level (TRL) is a measurement scale that researchers and engineers use to classify technology at different levels of development [11]. It means that it indicates the level of maturity of the lower exoskeleton system, which are levels from 1 to 9 where the first means that it is starting as research and the last means that they have already been approved as commercial devices through bodies such as the Food and Drug Administration (FDA) [12] and others. According to Table I., most lower extremity exoskeletons are found at TRL 5, TRL 6, or TRL 5/6. This means that around 12.7% of them are in the prototype stage. Furthermore, it was identified that TRL 7 and TRL 8 were the least frequent of the 71 exoskeletons analyzed. That means that ultimately few devices are being tested as prototypes in case studies with test subjects.

IV. CONCLUSION

In conclusion, this research article presents relevant information regarding the characteristics that are taken into consideration when proposing a new design for mechatronic systems such as lower limb exoskeletons. In this review, 71 research articles about lower limb exoskeletons were classified in the categories such as the commercial name, the degrees of freedom (DoF), the mechanism that covers the selected part of the lower limb, the biosignal that controls the functions of the exoskeleton of lower limbs, the characteristics of the test subjects who participated in the clinical evaluations, the pathology of the end users or potential clients, the application for which they are being designed and the Readiness Level (TRL) established by NASA [11]. It was determined that 63.38% do not have a commercial name assigned by their inventors. On the contrary, the minority of the exoskeletons found receive a commercial name, with 36.62% of them specifically as EICOSI [54],[14], Symbitron [76], among others. Regarding the degrees of freedom of the exoskeletons, the exoskeleton Monaco *et al.* It stands out for having the highest number of DoFs, which is 42 DoF, covering the hip, knee and ankle joints. That is, the entire lower limb. Exoskeletons with 1 DoF were also found, such as EICOSI [14], Khamar *et al.* [26], Al-Ayyad *et al.* [46] and Marsi Active Knee (MAK) [61], which encompass mechanisms such as the knee, leg, ankle, and foot. That is, only a specific part of the lower limb. It is important to mention that the number of DoF most used for the design of lower limb exoskeletons are usually in the range of 2 and 6 DoF. Lower limb exoskeletons with 2 DoF such as EXiO-A [23], Sun *et al.* [78], and Chen *et al.* [82] They prefer mechanisms that cover only specific parts of the lower extremity, such as the hip and knee. Regarding lower limb exoskeletons with 6 DoF such as ARKE [27], EHTE [72], Vinoj *et al.* [41], Campbell *et al.* [49] and Ezhilarasi *et al.* [80] they opt for mechanisms that cover almost the entire lower limb. This review identified articles employing 1 to 45 test subjects, generally healthy men (age: 6-40 years, weight: 55-95 kg, height: 1.60-1.82 m). These subjects simulated the physical structure and/or motor function of people who have suffered cerebrovascular accidents in their majority and SCI in their minority that totally or partially affect the lower limbs, generating in them pathologies such as hemiplegia, paraplegia and/or paralysis where these pathologies were related to the aging of the population, such as decreased movement, problems maintaining balance, for which it was determined that the most repeated application is rehabilitation and/or walking assistance. In addition, it was found that 38.03% of the exoskeletons analyzed use the EMG signal to control their functions, while 60.53% still do not use biosignals in the proposed devices or choose not to use them yet. On the other hand, only one of them use Electroencephalography (EEG) signals to control this devices. Finally, most lower extremity exoskeletons are found at TRL 5, TRL 6, or TRL 5/6. This means that around 12.7% of them are in the prototype stage. Furthermore, it was identified that TRL 7 and TRL 8 were the least frequent of the 71 exoskeletons analyzed. That means that few devices are ultimately being tested as prototypes in case studies.

Acknowledgment

This work is supported by IEEE Robotics and Automation Society (RAS) of the Universidad Tecnológica del Perú.

REFERENCES

- [1] "Disability." https://www.who.int/health-topics/disability#tab=tab_1 (accessed Apr. 09, 2022).
- [2] "Regional report on measuring disability: Overview of the disability measurement procedures in Latin America and the Caribbean. Task Force on Disability Measurement Statistical Conference of the Americas (SCA) | Publication | Economic Commission for Latin America and the Caribbean." <https://www.cepal.org/en/publications/36945-regional-report-measuring-disability-overview-disability-measurement-procedures> (accessed Apr. 09, 2022).
- [3] "Persons with disabilities and coronavirus disease (COVID-19) in Latin America and the Caribbean: status and guidelines | Publication | Economic Commission for Latin America and the Caribbean." <https://www.cepal.org/en/publications/45492-persons-disabilities-and-coronavirus-disease-covid-19-latin-america-and-caribbean> (accessed Apr. 09, 2022).
- [4] "Encuesta Nacional Especializada Sobre Discapacidad (ENEDIS) 2012 - [Instituto Nacional de Estadística e Informática - INEI] | Plataforma Nacional de Datos Abiertos." <https://www.datosabiertos.gob.pe/dataset/encuesta-nacional-especializada-sobre-discapacidad-enedis-2012-instituto-nacional-de> (accessed Apr. 09, 2022).
- [5] P. F. Pasquina, C. G. Emba, M. Corcoran, M. E. Miller, and R. A. Cooper, "Lower Limb Disability: Present Military and Civilian Needs," *Full Stride Adv. State Art Low. Extrem. Gait Syst.*, pp. 17–33, Sep. 2017, doi: 10.1007/978-1-4939-7247-0_2.
- [6] "Panorama Social de América Latina", Accessed: Apr. 09, 2022. [Online]. Available: www.cepal.org/apps
- [7] "Wearable Robots: Biomechanical Exoskeletons - José L. Pons - Google Libros." <https://books.google.es/books?hl=es&lr=&id=ovCKTEKEmk&Coi=fnd&pg=PR7&dq=biomecatr+onic+book&ots=NAAePTVMqB&sig=bE3Xx3GRGpv9SiMBHOJD68SSqXoFv=onepage&q=biomecatronic+book&f=false> (accessed Apr. 09, 2022).
- [8] J. Hamill, K. Knutzen, and T. Derrick, "Biomecánica básica Bases Del Movimiento Humano," 2017.
- [9] "Biomecánica de la extremidad inferior. 4. Exploración de la articulación del tobillo | Angulo Carrere | REDUCA (Enfermería, Fisioterapia y Podología)." <http://www.revistareduca.es/index.php/reduca-enfermeria/article/view/113/134> (accessed Apr. 09, 2022).
- [10] "Biomecánica de la extremidad inferior. 4. Exploración de la articulación del tobillo | Angulo Carrere | REDUCA (Enfermería, Fisioterapia y Podología)."
- [11] "Technology Readiness Level | NASA." https://www.nasa.gov/directorates/heo/scan/engineering/technology/technology_readiness_level (accessed Apr. 10, 2022).
- [12] "U.S. Food and Drug Administration." <https://www.fda.gov/> (accessed Apr. 10, 2022).
- [13] M. K. Shepherd and E. J. Rouse, "Design and Validation of a Torque-Controllable Knee Exoskeleton for Sit-to-Stand Assistance," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1695–1704, Aug. 2017, doi: 10.1109/TMECH.2017.2704521.
- [14] H. Rifai, S. Mohammed, K. Djouani, and Y. Amirat, "Toward Lower Limbs Functional Rehabilitation Through a Knee-Joint Exoskeleton," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 2, pp. 712–719, Mar. 2017, doi: 10.1109/TCST.2016.2565385.
- [15] F. Patané, S. Rossi, F. Del Sette, J. Taborri, and P. Cappa, "WAKE-Up Exoskeleton to Assist Children With Cerebral Palsy: Design and Preliminary Evaluation in Level Walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 7, pp. 906–916, Jul. 2017, doi: 10.1109/TNSRE.2017.2651404.
- [16] D. J. Hyun, H. Park, T. Ha, S. Park, and K. Jung, "Biomechanical design of an agile, electricity-powered lower-limb exoskeleton for weight-bearing assistance," *Rob. Auton. Syst.*, vol. 95, pp. 181–195, Sep. 2017, doi: 10.1016/j.robot.2017.06.010.
- [17] H. Kim, Y. June Shin, and J. Kim, "Design and locomotion control of a hydraulic lower extremity exoskeleton for mobility augmentation," *Mechatronics*, vol. 46, pp. 32–45, Oct. 2017, doi: 10.1016/j.mechatronics.2017.06.009.
- [18] Z. F. Lerner, D. L. Damiano, and T. C. Bulea, "The Effects of Exoskeleton Assisted Knee Extension on Lower-Extremity Gait Kinematics, Kinetics, and Muscle Activity in Children with Cerebral Palsy," *Sci. Reports 2017 71*, vol. 7, no. 1, pp. 1–12, Oct. 2017, doi: 10.1038/s41598-017-13554-2.
- [19] V. Monaco *et al.*, "An ecologically-controlled exoskeleton can improve balance recovery after slippage," *Sci. Reports 2017 71*, vol. 7, no. 1, pp. 1–10, May 2017, doi: 10.1038/srep46721.
- [20] J. Buurke *et al.*, "XoSoft – Development of a Soft Modular Lower Limb Exoskeleton," *Gait Posture*, vol. 57, p. 274, Sep. 2017, doi: 10.1016/j.gaitpost.2017.06.413.
- [21] Z. F. Lerner, D. L. Damiano, H. S. Park, A. J. Gravander, and T. C. Bulea, "A Robotic Exoskeleton for Treatment of Crouch Gait in Children With Cerebral Palsy: Design and Initial Application," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 650–659, Jun. 2017, doi: 10.1109/TNSRE.2016.2595501.
- [22] R. R. Torrealba, S. B. Udelman, and E. D. Fonseca-Rojas, "Design of variable impedance actuator for knee joint of a portable human gait rehabilitation exoskeleton," *Mech. Mach. Theory*, vol. 116, pp. 248–261, Oct. 2017, doi: 10.1016/j.mechmachtheory.2017.05.024.
- [23] T. Vouga *et al.*, "EXiO-A Brain-Controlled Lower Limb Exoskeleton for Rhesus Macaques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 2, pp. 131–141, Feb. 2017, doi: 10.1109/TNSRE.2017.2659654.
- [24] S. Mohan, J. K. Mohanta, S. Kurtenbach, J. Paris, B. Corves, and M. Huesing, "Design, development and control of a 2PRP-2PPR planar parallel manipulator for lower limb rehabilitation therapies," *Mech. Mach. Theory*, vol. 112, pp. 272–294, Jun. 2017, doi: 10.1016/j.mechmachtheory.2017.03.001.
- [25] W. Huo, S. Mohammed, Y. Amirat, and K. Kong, "Fast Gait Mode Detection and Assistive Torque Control of an Exoskeletal Robotic Orthosis for Walking Assistance," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1035–1052, Aug. 2018, doi: 10.1109/TRO.2018.2830367.
- [26] M. Khamar and M. Edrisi, "Designing a backstepping sliding mode controller for an assistant human knee exoskeleton based on nonlinear disturbance observer," *Mechatronics*, vol. 54, pp. 121–132, Oct. 2018, doi: 10.1016/j.mechatronics.2018.07.010.
- [27] B. N. Fournier, E. D. Lemaire, A. J. J. Smith, and M. Doumit, "Modeling and Simulation of a Lower Extremity Powered Exoskeleton," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1596–1603, Aug. 2018, doi: 10.1109/TNSRE.2018.2854605.
- [28] I. Kang, H. Hsu, and A. Young, "The Effect of Hip Assistance Levels on Human Energetic Cost Using Robotic Hip Exoskeletons," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 430–437, Jan. 2019, doi: 10.1109/LRA.2019.2890896.
- [29] C. Khazoum, C. Veronneau, J. P. L. Bigue, J. Grenier, A. Girard, and J. S. Plante, "Design and control of a multifunctional ankle exoskeleton powered by magnetorheological actuators to assist

- walking, jumping, and landing," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 3083–3090, Jul. 2019, doi: 10.1109/LRA.2019.2924852.
- [30] M. B. Yandell, J. R. Tacca, and K. E. Zelik, "Design of a Low Profile, Unpowered Ankle Exoskeleton That Fits Under Clothes: Overcoming Practical Barriers to Widespread Societal Adoption," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 712–723, Apr. 2019, doi: 10.1109/TNSRE.2019.2904924.
- [31] G. M. Gasparri, J. Luque, and Z. F. Lerner, "Proportional Joint-Moment Control for Instantaneously Adaptive Ankle Exoskeleton Assistance," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 751–759, Apr. 2019, doi: 10.1109/TNSRE.2019.2905979.
- [32] T. Zhang and H. Huang, "Design and Control of a Series Elastic Actuator with Clutch for Hip Exoskeleton for Precise Assistive Magnitude and Timing Control and Improved Mechanical Safety," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 5, pp. 2215–2226, Oct. 2019, doi: 10.1109/TMECH.2019.2932312.
- [33] A. Martinez, B. Lawson, C. Dorough, and M. Goldfarb, "A Velocity-Field-Based Controller for Assisting Leg Movement during Walking with a Bilateral Hip and Knee Lower Limb Exoskeleton," *IEEE Trans. Robot.*, vol. 35, no. 2, pp. 307–316, Apr. 2019, doi: 10.1109/TRO.2018.2883819.
- [34] Y. Yang, D. Huang, and X. Dong, "Enhanced neural network control of lower limb rehabilitation exoskeleton by add-on repetitive learning," *Neurocomputing*, vol. 323, pp. 256–264, Jan. 2019, doi: 10.1016/J.NEUCOM.2018.09.085.
- [35] P. T. Chinihill, Z. Qiao, S. M. Rezayat Sorkhabadi, V. Jhawar, I. H. Fong, and W. Zhang, "Automatic virtual impedance adaptation of a knee exoskeleton for personalized walking assistance," *Rob. Auton. Syst.*, vol. 114, pp. 66–76, Apr. 2019, doi: 10.1016/J.ROBOT.2019.01.013.
- [36] X. Liu, Z. Zhou, J. Mai, and Q. Wang, "Real-time mode recognition based assistive torque control of bionic knee exoskeleton for sit-to-stand and stand-to-sit transitions," *Rob. Auton. Syst.*, vol. 119, pp. 209–220, Sep. 2019, doi: 10.1016/J.ROBOT.2019.06.008.
- [37] Y. Bougrinat, S. Achiche, and M. Raison, "Design and development of a lightweight ankle exoskeleton for human walking augmentation," *Mechatronics*, vol. 64, p. 102297, Dec. 2019, doi: 10.1016/J.MECHATRONICS.2019.102297.
- [38] F. Sado, H. J. Yap, R. A. R. Ghazilla, and N. Ahmad, "Design and control of a wearable lower-body exoskeleton for squatting and walking assistance in manual handling works," *Mechatronics*, vol. 63, p. 102272, Nov. 2019, doi: 10.1016/J.MECHATRONICS.2019.102272.
- [39] E. Martini *et al.*, "Gait training using a robotic hip exoskeleton improves metabolic gait efficiency in the elderly," *Sci. Reports 2019 91*, vol. 9, no. 1, pp. 1–12, May 2019, doi: 10.1038/s41598-019-43628-2.
- [40] M. S. Amiri, R. Ramli, and M. F. Ibrahim, "Hybrid design of PID controller for four DoF lower limb exoskeleton," *Appl. Math. Model.*, vol. 72, pp. 17–27, Aug. 2019, doi: 10.1016/J.APM.2019.03.002.
- [41] P. G. Vinoy, S. Jacob, V. G. Menon, S. Rajesh, and M. R. Khosravi, "Brain-controlled adaptive lower limb exoskeleton for rehabilitation of post-stroke paralyzed," *IEEE Access*, vol. 7, pp. 132628–132648, 2019, doi: 10.1109/ACCESS.2019.2921375.
- [42] I. Farkhatdinov, J. Ebert, G. van Oort, M. Vlutters, E. van Asseldonk, and E. Burdet, "Assisting Human Balance in Standing With a Robotic Exoskeleton," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 414–421, Jan. 2019, doi: 10.1109/LRA.2018.2890671.
- [43] M. S. Amiri, R. Ramli, and M. F. Ibrahim, "Initialized Model Reference Adaptive Control for Lower Limb Exoskeleton," *IEEE Access*, vol. 7, pp. 167210–167220, 2019, doi: 10.1109/ACCESS.2019.2954110.
- [44] C. F. Chen, Z. J. Du, L. He, J. Q. Wang, D. M. Wu, and W. Dong, "Active Disturbance Rejection with Fast Terminal Sliding Mode Control for a Lower Limb Exoskeleton in Swing Phase," *IEEE Access*, vol. 7, pp. 72343–72357, 2019, doi: 10.1109/ACCESS.2019.2918721.
- [45] T. Luger, R. Seibt, T. J. Cobb, M. A. Rieger, and B. Steinhilber, "Influence of a passive lower-limb exoskeleton during simulated industrial work tasks on physical load, upper body posture, postural control and discomfort," *Appl. Ergon.*, vol. 80, pp. 152–160, Oct. 2019, doi: 10.1016/J.APERGO.2019.05.018.
- [46] M. Al-Ayyad, B. A. haj Moh'd, N. Qasem, and M. Al-Takrori, "Controlling a Lower-Leg Exoskeleton Using Voltage and Current Variation Signals of a DC Motor Mounted at the Knee Joint," *J. Med. Syst.*, vol. 43, no. 7, Jul. 2019, doi: 10.1007/S10916-019-1333-2.
- [47] M. Ding, M. Nagashima, S. G. Cho, J. Takamatsu, and T. Ogasawara, "Control of Walking Assist Exoskeleton with Time-delay Based on the Prediction of Plantar Force," *IEEE Access*, vol. 8, pp. 138642–138651, 2020, doi: 10.1109/ACCESS.2020.3010644.
- [48] Y. Li *et al.*, "Design and preliminary validation of a lower limb exoskeleton with compact and modular actuation," *IEEE Access*, vol. 8, pp. 66338–66352, 2020, doi: 10.1109/ACCESS.2020.2985910.
- [49] S. M. Campbell, C. P. Dلدuch, and J. W. Sensinger, "Autonomous Assistance-as-Needed Control of a Lower Limb Exoskeleton with Guaranteed Stability," *IEEE Access*, vol. 8, pp. 51168–51178, 2020, doi: 10.1109/ACCESS.2020.2973373.
- [50] G. Orehkov, Y. Fang, J. Luque, and Z. F. Lerner, "Ankle Exoskeleton Assistance Can Improve Over-Ground Walking Economy in Individuals with Cerebral Palsy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 461–467, Feb. 2020, doi: 10.1109/TNSRE.2020.2965029.
- [51] J. Wang, Y. Fei, and W. Chen, "Integration, Sensing, and Control of a Modular Soft-Rigid Pneumatic Lower Limb Exoskeleton," <https://home.lierbertpub.com/soro>, vol. 7, no. 2, pp. 140–154, Apr. 2020, doi: 10.1089/SORO.2019.0023.
- [52] J. Chen, Y. Huang, X. Guo, S. Zhou, and L. Jia, "Parameter identification and adaptive compliant control of rehabilitation exoskeleton based on multiple sensors," *Measurement*, vol. 159, p. 107765, Jul. 2020, doi: 10.1016/J.MEASUREMENT.2020.107765.
- [53] S. Han, H. Wang, and Y. Tian, "A linear discrete-time extended state observer-based intelligent PD controller for a 12 DOFs lower limb exoskeleton LLE-RePA," *Mech. Syst. Signal Process.*, vol. 138, p. 106547, Apr. 2020, doi: 10.1016/J.YMSSP.2019.106547.
- [54] K. I. K. Sherwani, N. Kumar, A. Chemori, M. Khan, and S. Mohammed, "RISE-based adaptive control for EICoSI exoskeleton to assist knee joint mobility," *Rob. Auton. Syst.*, vol. 124, p. 103354, Feb. 2020, doi: 10.1016/J.ROBOT.2019.103354.
- [55] L. Zhou, W. Chen, W. Chen, S. Bai, J. Zhang, and J. Wang, "Design of a passive lower limb exoskeleton for walking assistance with gravity compensation," *Mech. Mach. Theory*, vol. 150, p. 103840, Aug. 2020, doi: 10.1016/J.MECHMACHTHEORY.2020.103840.
- [56] C. T. Pan *et al.*, "Development a multi-loop modulation method on the servo drives for lower limb rehabilitation exoskeleton," *Mechatronics*, vol. 68, p. 102360, Jun. 2020, doi: 10.1016/J.MECHATRONICS.2020.102360.
- [57] D. Llorente-Vidrio, R. Pérez-San Lázaro, M. Ballesteros, I. Salgado, D. Cruz-Ortiz, and I. Chairez, "Event driven sliding mode control of a lower limb exoskeleton based on a continuous neural network electromyographic signal classifier," *Mechatronics*, vol. 72, p. 102451, Dec. 2020, doi: 10.1016/J.MECHATRONICS.2020.102451.
- [58] S. Yang, J. Han, L. Xia, and Y. H. Chen, "An optimal fuzzy-theoretic setting of adaptive robust control design for a lower limb exoskeleton robot system," *Mech. Syst. Signal Process.*, vol. 141, p. 106706, Jul. 2020, doi: 10.1016/J.YMSSP.2020.106706.
- [59] A. S. Nair and D. Ezhilarasi, "Performance Analysis of Super Twisting Sliding Mode Controller by ADAMS–MATLAB Co-simulation in Lower Extremity Exoskeleton," *Int. J. Precis. Eng. Manuf. Technol.* 2020 73, vol. 7, no. 3, pp. 743–754, Mar. 2020, doi: 10.1007/S40684-020-00202-W.
- [60] R. Pérez-San Lázaro, I. Salgado, and I. Chairez, "Adaptive sliding-mode controller of a lower limb mobile exoskeleton for active rehabilitation," *ISA Trans.*, vol. 109, pp. 218–228, Mar. 2021, doi: 10.1016/J.ISATRA.2020.10.008.
- [61] G. Puyuelo-Quintana *et al.*, "A new lower limb portable exoskeleton for gait assistance in neurological patients: a proof of concept study," *J. Neuroeng. Rehabil.*, vol. 17, no. 1, May 2020, doi: 10.1186/S12984-020-00690-6.
- [62] M. E. Mungai and J. W. Grizzle, "Feedback Control Design for Robust Comfortable Sit-to-Stand Motions of 3D Lower-Limb Exoskeletons," *IEEE Access*, vol. 9, pp. 122–161, 2021, doi: 10.1109/ACCESS.2020.3046446.
- [63] G. Aguirre-Ollinger and H. Yu, "Lower-Limb Exoskeleton with Variable-Structure Series Elastic Actuators: Phase-Synchronized Force Control for Gait Asymmetry Correction," *IEEE Trans. Robot.*, vol. 37, no. 3, pp. 763–779, Jun. 2021, doi: 10.1109/TRO.2020.3034017.
- [64] C. Camardella, F. Porcini, A. Filippeschi, S. Marcheschi, M. Solazzi, and A. Frisoli, "Gait Phases Blended Control for Enhancing Transparency on Lower-Limb Exoskeletons," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5453–5460, Jul. 2021, doi: 10.1109/LRA.2021.3075368.
- [65] D. Wei *et al.*, "Human-in-the-Loop Control Strategy of Unilateral Exoskeleton Robots for Gait Rehabilitation," *IEEE Trans. Cogn. Dev. Syst.*, vol. 13, no. 1, pp. 57–66, Mar. 2021, doi: 10.1109/TCDS.2019.2954289.
- [66] V. Molazadeh, Q. Zhang, X. Bao, and N. Sharma, "An Iterative Learning Controller for a Switched Cooperative Allocation Strategy During Sit-to-Stand Tasks with a Hybrid Exoskeleton," *IEEE Trans. Control Syst. Technol.*, pp. 1–16, Jul. 2021, doi: 10.1109/TCST.2021.3089885.
- [67] J. Lee, M. E. Huber, and N. Hogan, "Applying Hip Stiffness with an Exoskeleton to Compensate Gait Kinematics," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2645–2654, 2021, doi: 10.1109/TNSRE.2021.3132621.
- [68] Y. Long and Y. Peng, "Extended State Observer-Based Nonlinear Terminal Sliding Mode Control With Feedforward Compensation for Lower Extremity Exoskeleton," *IEEE Access*, vol. 10, pp. 8643–8652, 2022, doi: 10.1109/ACCESS.2021.3049879.
- [69] X. Tan, B. Zhang, G. Liu, X. Zhao, and Y. Zhao, "Cadence-Insensitive Soft Exoskeleton Design With Adaptive Gait State Detection and Iterative Force Control," *IEEE Trans. Autom. Sci. Eng.*, 2021, doi: 10.1109/TASE.2021.3066403.
- [70] S. V. Sarkisian, M. K. Ishmael, and T. Lenzi, "Self-Aligning Mechanism Improves Comfort and Performance with a Powered Knee Exoskeleton," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 629–640, 2021, doi: 10.1109/TNSRE.2021.3064463.
- [71] Y. Wang, H. Wang, and Y. Tian, "Adaptive interaction torque-based AAN control for lower limb rehabilitation exoskeleton," *ISA Trans.*, Oct. 2021, doi: 10.1016/J.ISATRA.2021.10.009.
- [72] X. Zhou, G. Liu, B. Han, L. Wu, and H. Li, "Design of a Human Lower Limbs Exoskeleton for Biomechanical Energy Harvesting and Assist Walking," *Energy Technol.*, vol. 9, no. 1, p. 2000726, Jan. 2021, doi: 10.1002/ENTE.202000726.
- [73] Y. Zheng, Y. Wang, and J. Liu, "Analysis and experimental research on stability characteristics of squatting posture of wearable lower limb exoskeleton robot," *Futur. Gener. Comput. Syst.*, vol. 125, pp. 352–363, Dec. 2021, doi: 10.1016/J.FUTURE.2021.06.053.
- [74] R. Sharma, P. Gaur, S. Bhatt, and D. Joshi, "Optimal fuzzy logic-based control strategy for lower limb rehabilitation exoskeleton," *Appl. Soft Comput.*, vol. 105, p. 107226, Jul. 2021, doi: 10.1016/J.ASOC.2021.107226.
- [75] J. Zhao, T. Yang, Z. Ma, C. Yang, Z. Wang, and J. Xu, "Design of M-G modal space sliding mode control for lower limb exoskeleton robot driven by electrical actuators," *Mechatronics*, vol. 78, p. 102610, Oct. 2021, doi: 10.1016/J.MECHATRONICS.2021.102610.
- [76] C. Mejneke *et al.*, "Symbion Exoskeleton: Design, Control, and Evaluation of a Modular Exoskeleton for Incomplete and Complete Spinal Cord Injured Individuals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 330–339, 2021, doi: 10.1109/TNSRE.2021.3049960.
- [77] J. Lin, N. V. Divekar, G. Lv, and R. D. Gregg, "Optimal Task-Invariant Energetic Control for a Knee-Ankle Exoskeleton," *IEEE Control Syst. Lett.*, vol. 5, no. 5, pp. 1711–1716, Nov. 2021, doi: 10.1109/LCSYS.2020.3043838.
- [78] W. Sun, J. W. Lin, S. F. Su, N. Wang, and M. J. Er, "Reduced Adaptive Fuzzy Decoupling Control for Lower Limb Exoskeleton," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1099–1109, Mar. 2021, doi: 10.1109/TCYB.2020.2972582.
- [79] J. Tian, L. Yuan, W. Xiao, T. Ran, and L. He, "Trajectory following control of lower limb exoskeleton robot based on Udwardia-Kalaba theory," <https://doi.org/10.1177/10775463211031701>, Jul. 2021, doi: 10.1177/10775463211031701.
- [80] D. Ezhilarasi and A. S. Nair, "Modeling and Evaluation of Adaptive Super Twisting Sliding Mode Control in Lower Extremity Exoskeleton," *Int. J. Precis. Eng. Manuf. Technol.* 2021 83, vol. 8, no. 3, pp. 901–915, Apr. 2021, doi: 10.1007/S40684-021-00335-6.
- [81] D. Shi, W. Zhang, W. Zhang, L. Ju, and X. Ding, "Human-centric adaptive control of lower limb rehabilitation robot based on human-robot interaction dynamic model," *Mech. Mach. Theory*, vol. 162, p. 104340, Aug. 2021, doi: 10.1016/J.MECHMACHTHEORY.2021.104340.
- [82] Z. Chen, Q. Guo, Y. Yan, and Y. Shi, "Model identification and adaptive control of lower limb exoskeleton based on neighborhood field optimization," *Mechatronics*, vol. 81, p. 102699, Feb. 2022, doi: 10.1016/J.MECHATRONICS.2021.102699.
- [83] J. Liu, Y. He, J. Yang, W. Cao, and X. Wu, "Design and analysis of a novel 12-DOF self-balancing lower extremity exoskeleton for walking assistance," *Mech. Mach. Theory*, vol. 167, p. 104519, Jan. 2022, doi: 10.1016/j.mechmachtheory.2021.104519.
- [84] J. Buurke *et al.*, "XoSoft – Development of a Soft Modular Lower Limb Exoskeleton," *Gait Posture*, vol. 57, p. 274, Sep. 2017, doi: 10.1016/J.GAITPOST.2017.06.413.

Deep Learning: An Empirical Study on Kimia Path24

Shaikh Shiam Rahman
Department of Information Technology
Georgia Southern University
Statesboro, GA USA
sr16276@georgiasouthern.edu

Hayden Wimmer
Department of Information Technology
Georgia Southern University
Statesboro, GA USA
hwimmer@georgiasouthern.edu

Loreen Powell
Department of Information Technology,
Analytics, and Business Education
Bloomsburg University
Bloomsburg, PA USA
lpowell@bloomu.edu

Abstract— Deep learning has a large interest in medical image analysis as studies have shown several machine learning algorithms were successful in predicting disease. However, more work is needed to better understand the batch size, epoch, and learning rates. An empirical study of image processing with deep learning was conducted on the KIMIA path24 dataset. The rotation, width shifting, height shifting shear range, horizontal flip, and fill mode was used. The network was trained and validated by a total of 22,591 images from the KIMIA path24 dataset. ReLU was used for the convolution layer and softmax for the fully connected layer. Results found the batch size is inversely proportional to the network accuracy, the accuracy of a deep learning network is directly proportional to the number of epochs it passes through, and the learning rate does not bring any change to the network. The network performs best within a preferred learning rate.

Keywords—deep learning, machine learning, deep convolutional networks, image classification

I. INTRODUCTION

For several decades, pathology has been described as the archiving of microscopic information of specimens. This microscopic information is organized by storing specimens on glass slides. The problem with glass slides is that they are fragile and require a very large specially prepared storage room to store the specimens in. This kind of storage requires a lot of logistical infrastructures.

In 1999, whole slide imaging (WSI) was introduced by Wetzel and Gilbertson [1]. WSI can provide high image quality that doesn't decay over time, along with a range of other benefits. WSI can be used by multiple researchers to investigate multiple slides at the same time, and this kind of data is more useful in order to retrieve information and maintain quality control.

Pathology bounded by the WSI system is emerging into an era of digital specialty. WSI is providing solutions for centralizing diagnostic by improving the quality of diagnosis, patient safety, and economic concerns. [2].

The diagnosis of WSI is still difficult. The gigapixel nature of WSI scans makes it difficult to store, transfer, and process samples in real-time. One also needs tremendous digital

storage to archive them [3]. A diagnostic system aided by digital pathology scanned data would allow for a more objective approach and increase our ability to predict an individual's pathological diagnosis and treatment response.

The most common form of machine learning is supervised learning. In computer vision, deep convolutional networks have now become the technique of choice [4]. Deep learning is making major advances in solving problems in the artificial intelligence community. Deep learning deals with the problem of data representation by introducing simpler intermediate representations that allow them to combine in order to build complex concepts [5]. Deep learning methods are multiple levels of representation. These representations are obtained by composing simple but non-linear modules that transform the representation at one level into a representation at a higher abstract level. Recent studies have shown that machine learning algorithms were able to predict disease more accurately than experienced clinicians [6]. Deep learning has a large interest in medical image analysis. It is expected that deep learning will hold \$300 million for the medical imaging market by 2021 [7]. It is of great interest to develop and improve such prediction methods. The goal of this paper is to conduct an empirical study of image processing with deep learning using the KIMIA path24 dataset.

II. RELATED WORKS

A. Deep Learning

Deep learning-based image super-resolution (SR) models have been actively explored and often achieve the state-of-the-art performance on various benchmarks of SR. A variety of deep learning methods have been applied to tackle SR tasks ranging from the early Convolutional Neural Networks based method to recent promising SR approaches using Generative Adversarial Nets [8]. Dong, et al. [9] proposed a deep learning method for a single SR. In this study, the authors contributed three aspects: 1) presented a convolutional neural network for image super-resolution, 2) established a relationship between the deep learning-based SR method and the traditional sparse coding-based SR methods, and 3) demonstrated the usefulness of deep learning in the classical computer vision problem with super-resolution. The method of this study directly established an end-to-end mapping between the low/high-resolution

images. The mapping is represented as a deep convolutional neural network (CNN) that takes the low-resolution image as the input and outputs the high-resolution. With a lightweight structure, the Super-Resolution Convolutional Neural Network (SRCNN) has achieved a superior performance above other state-of-the-art methods. The authors conjectured that an additional performance could be further gained by exploring more hidden layers/filters in the network using different training strategies [9].

Chen, et al. [10] introduced the concept of deep learning into hyperspectral data classification. They introduced the deep learning-based feature extraction for hyperspectral data classification. Their method focused on applying an autoencoder (AE). At first, the authors verified the eligibility of stacked autoencoders by following classical spectral information-based classification. After that, a new way of classifying with spatial-dominated information was presented. Finally, they proposed a deep learning framework by merging the two features. Chen, et al. [10] exploited a single layer autoencoder (AE) and a multi-layer stacked AE (SAE) to learn shallow and deep features of hyperspectral data. AE-extracted features are useful for classification. AE and SAE deep feature extraction models increased the accuracy of SVM and logistic regression while obtaining the highest accuracy when compared with other feature extraction methods.

Chetlur, et al. [11] created a library with optimized routines for deep learning workloads with a similar intent to Basic Linear Algebra Subroutines (BLAS) [12]. They presented a novel implementation of convolutions that provided a reliable performance across a wide range of input sizes, and they took advantage of the highly optimized matrix multiplication routines to provide a high performance without requiring any auxiliary memory. Integrating Nvidia CUDA Deep Neural Network's (cuDNN) library into Caffe improved the performance by 36% on a standard model with a reduced memory consumption. NVIDIA cuDNN's performance is 86% of the maximum performance, with a small mini-batch size of 16. This implementation performed well across the convolution parameter space. NVIDIA cuDNN's library ranged from 23-35% of peak performance on the Tesla K40 and from 30-51% of peak performance on the GTX 980. NVIDIA cuDNN's library provided a performance portability across GPU architectures with no need for users to retune their code as GPU architectures evolve [11].

A large body of the work in deep learning can be classified into: (1) generative, (2) discriminative, and (3) hybrid categories. A deep autoencoder is used for learning efficient encoding or dimensionality reduction for a set of data. It is a non-linear feature extraction method classified as generative. Deep architecture, consisting of both pretraining and fine-tuning stages in its parameter learning, is classified as a hybrid. The concept of stacking, where simple modules of functions are composed first and then they are "stacked" on top of each other in order to learn complex functions, is classified as discriminative [13].

Algorithm-level noise tolerance can be leveraged to simplify underlying hardware requirements. Noise tolerance can lead to a co-optimized system that achieves significant improvements in computational performance and energy efficiency. Deep networks can be trained using only 16-bit wide fixed-point number representation using stochastic rounding with little degradation in classification accuracy [14].

Deep learning methods have dramatically improved the state-of-the-art speech recognition, visual object recognition, object detection, and many other domains, such as drug discovery and genomics. Deep convolutional nets have brought about breakthroughs in processing images, video, speech, and audio; whereas, recurrent nets have shined the light on sequential data, such as text and speech. Deep learning has beaten other machine-learning techniques by predicting the activity of potential drug molecules, analyzing particle accelerator data reconstructing brain circuits, and determining the effects of mutations in non-coding DNA on gene expression and disease. The 2012 ImageNet competition success has brought about a revolution in computer vision. ConvNets are now the dominant approach for almost all recognition and detection tasks. ConvNets also enhanced human performance on some tasks. The combination of ConvNets and the recurrent net modules generate stunning demonstrative image captions [15].

Cross modality feature learning can achieve a better feature for one modality (e.g., video) if multiple modalities (e.g., audio and video) are present at the feature learning time [16]. Ngiam, et al. [16] presented a series of tasks for multimodal learning showing how to train deep networks to learn features that address these tasks. These models were validated on the CUAVE and AVLetters datasets on audio-visual speech classification, demonstrating the best published visual speech classification on AVLetters and effective shared representation learning. Learning a canonical correlation analysis (CCA) with a shared representation of raw data results in a good performance. Learning the CCA representation on the first layer results in a significantly better performance compared to the original modalities for supervised classification.

Papernot, et al. [17] formalized the space of adversaries against deep neural networks (DNNs) and introduced a novel class of algorithms to craft adversarial samples based on a precise understanding of the mapping between inputs and outputs of DNNs. This experiment formally described a class of algorithms for crafting adversarial samples misclassified by DNNs using three tools: the forward derivative, adversarial saliency maps, and the crafting algorithm. These tools were applied to a DNN and used for a computer vision classification task: handwritten digit recognition. The crafting algorithm can reliably produce samples correctly classified by human subjects but misclassified in specific targets by a DNN with a 97% adversarial success rate while only modifying on average 4.02% of the input features per sample.

B. Deep Learning in Medical Images

Brosch and Tam [18] described a novel method called multi-scale structured convolutional neural networks (MS-CNN) for learning the manifold of 3D brain images. The method does not require the manifold space to be locally linear, and it does not require a predefined similarity measure or a prebuilt proximity graph. This manifold learning method was based on deep learning, a machine learning approach that uses layered networks (called deep belief networks, or DBNs). The authors proposed a computationally efficient training method for DBN. The proposed method performed manifold learning by reducing the dimensionality of the input images using a DBN. This method used deep learning to discover patterns of similarity and variability within a group of images. The learned manifold coordinates captured shape variations of the brain that correlated with demographic and disease parameters. The MS-CNN algorithm was much more efficient than traditional, convolution-based methods,

Plis, et al. [19] used deep learning to analyze the effect of parameter choices on data transformations. The authors demonstrated their results in the application of deep learning methods to structural and functional brain imaging data. They also described a novel constraint-based approach to visualize high dimensional data. These methods included deep belief networks and the building block of the restricted Boltzmann machine. The main goal was to validate the feasibility of this application by: (1) investigating if a building block of deep generative models a restricted Boltzmann machine, (2) examining the effect of the depth in deep learning analysis of structural magnetic resonance imaging (MRI) data, and (3) determining the value of the methods for discovering the latent structure of a large-scale. Deep learning has a high potential in neuroimaging applications. The depth of the DBN helped classification and increased group separation. DBNs have a high potential for exploratory analysis.

Payan and Montana [20] used deep learning methods, sparse autoencoders, and 3D convolutional neural networks to build an algorithm that can predict the disease status of a patient based on an MRI scan of the brain. This study demonstrated that the 3D convolutional neural networks outperformed several other classifiers. The authors compared the performance of 2D and 3D convolutional networks. This study took a two-stage approach. The authors used a sparse autoencoder to learn filters for convolution operations, and then built a convolutional neural network in which the first layer uses the filters with the autoencoder. The 3D approach had a superior performance for the 3-way comparison. The 3D approach had the potential to capture local 3D patterns, which may boost the classification performance, albeit only by a small margin.

Deep learning has been applied to medical images and demonstrates promising results in multiple instances including computerized prognosis for Alzheimer's disease [21], tumor segmentation [22], and histopathological diagnosis [23, 24]. Neuroimaging analysis is oftentimes employed to improve diagnostic abilities; furthermore, deep learning models and methods are being increasingly utilized to discover patterns and

rules within patient data as seen in [25-28]. The implementation was done for this classifier using a deep neural network initialized by a DBN (DBN-DNN).

Pinaya, et al. [29] trained a deep learning model, known as DBN, to extract features from brain morphometry data. The deep learning models excel at neuroimaging-based prediction methods and can be useful for demonstrating complex and subtle associations, as well as enabling more accurate individual-level clinical assessments. The strength of deep architecture came from multiple levels of non-linear processing that are well-suited to capture highly varying functions with a compact set of parameters. The deep architecture provided superior performance in classification tasks. The DBN highlighted differences between classes, especially in the frontal, temporal, parietal, and insular cortices and in some subcortical regions, including the corpus callosum, putamen, and cerebellum.

Bao and Chung [30] stated that specific architecture is designed to capture discriminative features for each sub-cortical structure. To improve the performance of CNN, two intuitive ways are generally utilized. Apart from the straightforward enlargement of network architecture, some elegant micro-structures have also been designed to enhance the capability recently. To evaluate the performance of the proposed method, the comparison with other methods has been carried out. As several different voxel resolutions exist in each dataset, affine transformations between atlases and the target image were first conducted as pre-processing. The initial structural surface used in label consistency (Blue Curve at Iteration 0) was generated by fusing the warped label maps with majority voting (MV). The results of MV were also employed as a baseline for comparison. The results indicated that, with a multi-scale strategy, more discriminative features can be captured, and the labeling result can be improved. Due to the lack of constraints among testing patches, embracing learning method alone often led to a rough boundary and desultory segmentation result. Experimental results demonstrated that the proposed method obtained a better performance as compared to other state-of-the-art methods.

de Brebisson and Montana [31] proposed a methodology based on a deep artificial neural network that assigned each voxel in an MR image of the brain to its corresponding anatomical region. The inputs of the network captured information at different scales around the voxel of interest: 3D and orthogonal 2D intensity patches captured a local spatial context while downscaling large 2D orthogonal patches and distances to the regional centroids enforce global spatial consistency. The combined use of three 2D orthogonal patches dramatically improved the segmentation performance compared to 2D or 3D patches. The distances to centroids, in addition to their invariance qualities, significantly outperformed the coordinates. For already manually segmented brains, using estimated centroids yield equivalent results as using the true centroids. It has a mean dice coefficient of 0.725 and an error rate of 0.163 when evaluated on the 20 testing MRIs of the MICCAI challenge. Good validation results were

obtained by training huge networks, sometimes composed of tens of millions of parameters, with a relatively small amount of data. The trained networks overfit the training data; however, they still generalize fairly well to unseen MRIs.

Chen, et al. [32] explored the deep residual learning on the task of volumetric brain segmentation. First, they proposed a deep voxelwise residual network, referred to as VoxResNet. Second, an auto-context version of VoxResNet is proposed by seamlessly integrating the low-level image appearance features, implicit shape information, and high-level context together for improving the volumetric segmentation performance. The results of combining multi-modality and auto-context information give more accurate results visually than only multi-modality information. The proposed algorithm has an application beyond brain segmentation, and it can be applied in other volumetric image segmentation problems.

Choi and Jin [33] developed a fast and accurate method for the striatum segmentation using deep convolutional neural networks (CNN). The delicate segmentation process performed by Local CNN used small portions of the image rather than the whole image. This method suggested that FreeSurfer segmentation included slightly more true-positive voxels and much more false-positive voxels than the CNN-based approach, which resulted in lower Dice Similarity Coefficient (DSC). This approach depended on training using manual segmentation, another recently developed automatic segmentation tool based on multimodal images (MIST), did not require a manually labeled training set, which could be flexible for various types of data.

III. BACKGROUND

A. Single Neural Network

The basic unit of every single network is the neuron. This basic unit is also known as a node. The main computational principle of a neural network is that it receives inputs from any external sources and generates an output.

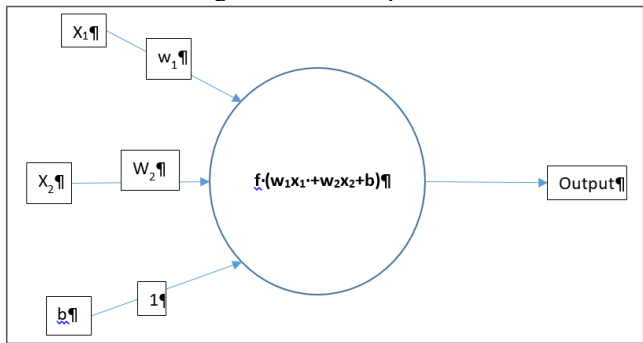


Fig. 1: Single Neuron

Each input has an associated weight which is assigned based on its source, as shown in Fig 1. The node applies a function to the weighted sum of the input. The function that is used for calculating output is called the activation function. The above network, shown in Fig. 1, contains 2 input nodes X_1 and X_2 . The associated weights are W_1 and W_2 . There is another input with weight 1 and value b , which is known as bias.

The output of this neuron is calculated as shown in Fig. 1. The function used to calculate the value of the neural network is known as the activation function. This function is a non-linear function, which is used for introducing the non-linearity to the output of the neural network. According to most of the real-world dataset, the inputs are discrete and non-linear, and our main target is to train our network with this non-linear representation.

Every activation function takes a single number and performs a specific fixed mathematical operation [34]. There are several activation functions that can be used based on the input of the neural network.

B. Feed Forward Neural Network

The feed forward neural network was the first and simplest type of artificial neural network devised [35]. This network model contains multiple basic units referred to as nodes arranged in layers, followed by the same type of adjacent layer. Nodes from adjacent layers have a connection between them, but nodes from the same layer have no connection among them. These connections are called edges. Each adjacent edge has an associated weight with it.

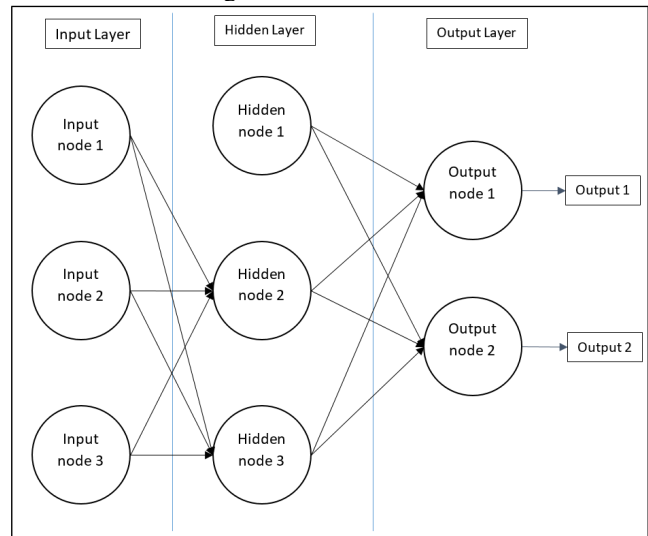


Fig. 2: Feed Forward Neural Network

There are three types of node layers in a feed forward neural network. These are: 1) input nodes, 2) hidden layers, and 3) output node. Three types of nodes in a network are shown in Fig. 2.

- 1) **Input layer:** The input nodes gather the information from the outer world. The combination of input nodes is known as the input layer. These nodes are primarily responsible for receiving inputs from the dataset. No computations are performed in this layer.
- 2) **Hidden Layer:** Hidden layers are those that take inputs from the previous input layer. For each input layer, first, the input and weight of the connections are multiplied, and then all the input weight multiplication products are summed up. After that, the result of the summation is put through an activation function and the output is forwarded

to the next layer. Equation (1.1) shows the output calculation for each node [4]. Here, the function in equation (1) is called an activation function.

$$f(\text{summation}) = f(w_0 * 1 + W_1 * X_1 + W_2 * X_2) \quad (1)$$

- 3) **Output Layer:** The output nodes also works the same way as the hidden nodes. These nodes take inputs from the previous hidden layer. For each output node, they first take the multiplication of the output of the previous node and the weight of the connection. Then, they sum up the multiplication results of all connections and put the summation on an activation function. The final step is to publish the outcome of the activation function as the output of the network.

C. Activation Function

Sigmoid: Equation (2) shows the sigmoid activation function. This takes a real value input and returns a value between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

Tanh: Equation (3) shows the Tanh activation function. It takes a real value input and returns a value between -1 and 1.

$$\tanh(x) = 2\sigma(2x) - 1 \quad (3)$$

ReLU: ReLU stands for Rectified Linear Unit. It takes a real value input and thresholds it between the max and cuts the negative values to 0. Equation (5.4) shows the ReLU activation function.

$$f(x) = \max(0, x) \quad (4)$$

Softmax: The softmax activation function turns the numbers into possibilities that sum to one. Softmax activation function changes the outcomes to a vector, which represents the probability distribution of a list of potential outcomes. Equation (5) shows the sigmoid activation function [4].

$$S(x_i) = \frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}} \quad (5)$$

D. Back Propagation

1) Calculating the Total Error

Once the result for each output node is obtained, the next goal is to calculate the error for each node. It can calculate the error for each output node using the squared error function and sum them to get the total error. Equation (6) shows how to calculate the error for back propagation. Here, target is the expected value from the output node, and actual is the calculated result after forward propagation using the activation function.

The goal with back propagation is to update each weight of the network connections, so they cause the actual

output to be closer to the target output. This helps minimize the error of each node and the network as a whole.

$$Error_{total} = \sum \frac{1}{2} (\text{target} - \text{actual})^2 \quad (6)$$

E. Reducing Loss

Gradient Descent: Calculating the loss function for every single input is not a very efficient approach to find the convergence point. The convex problems have only one minimum data place where the slope is exactly 0. That is the minimum function where the loss function converges. The best approach to obtain the convergence point is called the gradient descent. The first step is to pick a random number between 0 and 1. The gradient descent algorithm then calculates the gradient of the loss curve at the starting point. Something that should be noted is that our gradient needs to have the property of moving both forward and backward. So, we pick our gradient as a vector, which has both direction and magnitude. The gradient always points towards the increase of the loss function. The gradient descent algorithm takes a step in the direction of the negative gradient in order to reduce loss as quickly as possible.

Stochastic Gradient Descent: In gradient descent, the whole dataset on a single iteration is not used. Only a certain amount of data at a single time, making it flow through the network and calculating the gradient is used. This a single iteration. When the dataset is large, then a single portion of data may take a long time to complete computation. A large dataset with randomly sampled examples probably contains redundant data. In this case, a large amount of data is certainly not carrying much productive value. The stochastic gradient descent uses only a single example per iteration from the dataset to calculate the gradient descent. The term stochastic indicates that one data example data is picked randomly per iteration.

An artificial neural network is a machine learning computational model. This model is inspired by the biological neural network in the human brain. This model processes the information the same way the human brain processes the information. Artificial neural networks brought some breakthrough in machine learning research. Some of them are image recognition, computer vision and text processing.

IV. METHODOLOGY

The deep learning network depends on multiple parameters that is useful for tuning our network. These parameters help to increase the performance of the model, reduce the memory allocation and training time. These parameters are: 1) batch size, 2) number of epochs, and 3) learning rate.

The batch size determines the number of inputs that is fed into the deep learning network for one iteration. The batch size depends on the memory of the machine. When the batch size is too high, it consumes a lot of memory.

The number of epochs determines how many times the whole input set is fed to the neural network. Thus, if there

is an increase in the number of epochs, then the accuracy of the network will increase.

Finally, the learning rate determines how frequently the deep learning network changes decision. If the learning rate is too high, then the deep learning network changes decisions very frequently. In most cases, a high learning rate doesn't bring any drastic changes to the accuracy. Also, a low learning rate may take too long to make a change in decision.

The goal of this research is to examine the following hypotheses:

H₁: The batch size is inversely proportional to the network accuracy. If we increase the batch size, then accuracy is decreased.

H₂: The accuracy of a deep learning network is directly proportional to the number of epochs it passes through.

H₃: Lower and higher learning rate do not bring any change to the network. The network performs best within a preferred learning rate

A. Dataset Description

This research utilized the dataset from Kimia Path24 [3]. This dataset has a total of 22,591 training images from 24 different categories. This dataset has a total of 1,325 testing images that include all categories. The image size is 1000 x 1000 for all the images.

B. Data Preparation

The Keras library from python was used to build the neural network. First, the image was loaded using the OpenCV library and then resized the image to 28 x 28. Then, the dataset was divided into two parts: 1) training and 2) testing. The training set contained 75% of the data, and the testing set contained 25% of the data. The testing set is used for validating the network. The labels were converted from integer to vectors and data augmentation was performed. The rotation range was set to 30, width shift range to 0.1, height shift range to 0.1, shear range to 0.2, zoom range to 0.2, horizontal flip to true, and fill mode to nearest.

C. Building Network Model

After completing the data augmentation, the neural network was built. Fig. 3 contains the code snippet of the neural network. Our deep learning model contains 3 convolution networks. A small portion of the convolution layer is used as the input of each node, and the size of the small portion is often 3 x 3 or 5 x 5 [36]. The activation function is the rectified linear unit followed by a flat layer, and finally, two dense layers. The first layer has the activation function rectified linear unit (ReLU) and the output layer has the softmax. The softmax helps to keep the output between zero and one. Next, the adam optimizer was set with a learning rate of 10⁻³ and the model with a batch size of 32, the number of epochs as 25 was set.

```
class LeNet:
    @staticmethod
    def build(width, height, depth, classes):
        # initialize the model
        model = Sequential()
        inputShape = (height, width, depth)

        # if we are using "channels first", update the input shape
        if K.image_data_format() == "channels_first":
            inputShape = (depth, height, width)

        # first set of CONV -> RELU -> POOL layers
        model.add(Conv2D(20, (5, 5), padding="same",
            input_shape=inputShape))
        model.add(Activation("relu"))
        model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))

        # second set of CONV -> RELU -> POOL layers
        model.add(Conv2D(50, (5, 5), padding="same"))
        model.add(Activation("relu"))
        model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))

        # Third set of CONV -> RELU -> POOL layers
        model.add(Conv2D(64, (5, 5), padding="same"))
        model.add(Activation("relu"))
        model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))

        # first (and only) set of FC -> RELU layers
```

Fig. 3: Building Deep Learning Neural Network with Keras

D. Training the Model

After setting all the parameters, we ran the model with the fit_generator. Then, we saved the model and the learning curve plot with loss, accuracy, validation loss, and validation accuracy.

E. Tesingt the Model

After we finished training our model, we used a different script to test the model. We loaded the saved model and ran the prediction methods with the test data and generated the confusion matrix to find the accuracy, recall, precision, and f1-score.

V. RESULTS

The change of accuracy with respect to the batch size is displayed in Table 1. From the table, we can see that the accuracy decreases when we increase the batch size, which is satisfying our first hypothesis. Fig. 4 shows that the decrease in accuracy is linear with respect to the batch size.

TABLE 1. CHANGE OF ACCURACY WITH BATCH SIZE

Epoch	Batch size	Learning rate	Accuracy
50	16	0.001	0.6491
50	32	0.001	0.6204
50	64	0.001	0.5585
50	128	0.001	0.4989
50	256	0.001	0.4777
50	512	0.001	0.4438
50	1024	0.001	0.4513

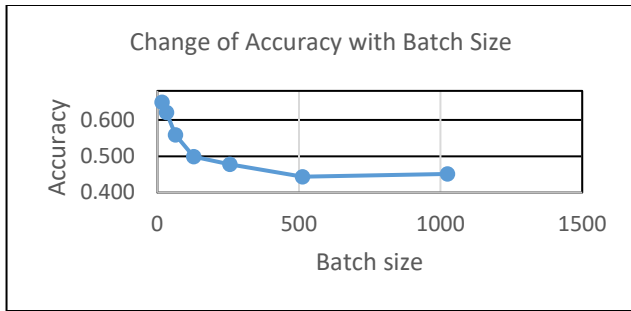


Fig. 4: Change of Accuracy with Batch Size

The change of accuracy with the change of epoch is displayed in Table 2. From the table, we can see that the accuracy increases with the number of epochs, which satisfies our second hypothesis. Fig. 5 shows that the accuracy is not linearly correlated with the number of epochs when the value is small. With a larger epoch size, the accuracy increases linearly with respect to the number of epochs.

TABLE 2. CHANGE OF ACCURACY WITH NUMBER OF EPOCHS

Epoch	Batch size	Learning rate	Accuracy
10	32	0.001	0.5668
15	32	0.001	0.5970
25	32	0.001	0.5298
35	32	0.001	0.6128
50	32	0.001	0.6204
70	32	0.001	0.6491
100	32	0.001	0.6732

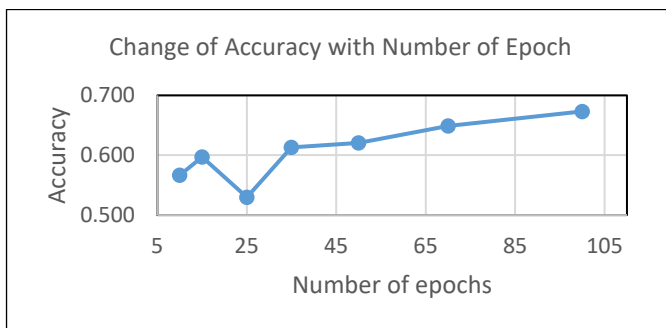


Fig. 5: Change of accuracy with number of epochs

From Table 3, we can see that the accuracy increases when we increase the learning rate from 0.1 to 0.001. Then after 0.005, the accuracy starts decreasing. At the learning rate of 0.0001, the accuracy is 45.6%, which is less than the accuracy of the learning rate of 0.001 (61.9%). The accuracy starts decreasing after the learning rate of 0.0001. This experiment validates our third hypothesis that the higher learning rate and lower learning rate do not always bring better changes to the accuracy.

TABLE 3. CHANGE OF ACCURACY FOR LEARNING RATE

Epoch	Batch size	Learning rate	Accuracy
50	32	0.1	0.0528
50	32	0.05	0.0302
50	32	0.01	0.0566
50	32	0.005	0.5925
50	32	0.001	0.6189
50	32	0.0005	0.6272
50	32	0.0001	0.4558

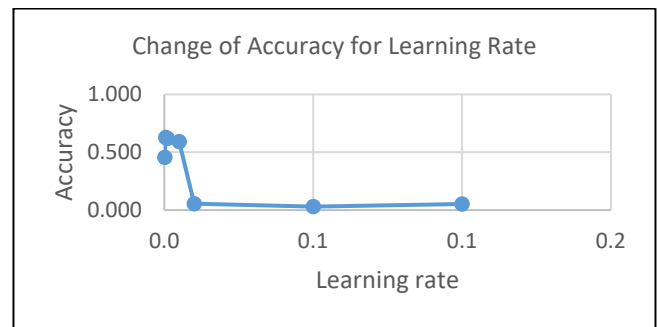


Fig.6: Change of Accuracy for Learning Rate

Fig. 6 shows the change in accuracy, with respect to learning rate. The relation between learning rate and accuracy is not linear. The accuracy does not increase or decrease linearly with the change in learning rate.

VI. DISCUSSION

In this experiment, we have done an empirical study of image processing with deep learning. We have conducted the experiments on the KIMIA path24 dataset. We have simply designed the architecture and used the feature selection. The image processing mechanism was used for manipulating the dataset. We have used rotation, width shifting, height shifting shear range, horizontal flip, and fill mode. The network was trained and validated by a total of 22,591 images from the KIMIA path24 dataset. The running time of the network was 6-7 hours depending on the parameters. We used a Google collaboratory training environment equipped with two core Intel's @2.3 GHz Xeon processor with a NVIDIA Tesla K80 (GK210 chipset) [37].

In this experiment, we used two serial 2-dimensional CNN architectures for segmentation. We used ReLU for the convolution layer and softmax for the fully connected layer. During the compilation, we used the adam optimizer and set loss to binary_crossentropy. The learning rate, the batch size, and the number of epochs was set according to hypotheses testing.

As a limitation of our experiment, various types of network architecture for the segmentation are possible. As

proof of the concept to test those hypotheses of our deep learning network parameter, the structure of the architecture was empirically designed after studying the image prototype. An adjustment of our proposed network, including the number of convolution layers, the number of nodes, the activation function, and other parameter upgrades, are possible. As a future work, an optimization of the network architecture and the parameter upgrades could increase the runtime and improve the performance. As our architecture and other parameters were set to test those hypotheses against the KIMIA path24 dataset, this design can be optimized and retrained for other image datasets.

One of the issues of the deep convolutional neural network is sample images. The number of images required to test and validate the network is not always available. The scale of the medical data available for studies is insufficient for machine learning and computer vision. This can affect the performance of deep learning for medical data [38]. In general, the convolutional neural network requires a large number of training data for each specific category. In our current dataset, we have a total of 24 categories. Some categories have more than 1,000 images for training, whereas few categories have less than 100 images for training and validation. These categories decline in accuracy and affect the overall performance. Using a proper dataset for all categories can improve the performance and provide more degree of freedoms.

VII. CONCLUSION

In this experiment, we built a simple deep convolutional neural network and conducted hypothesis testing on three hypotheses on the Kimia Path24 dataset. 2D convolutional neural networks were used for image classification. Our three hypotheses were proved correct. We used an image feature selection and image processing mechanisms for moderating the dataset. The parameter tuning and network structure designs provide essential outcomes, which ensures the credibility of our model.

REFERENCES

- [1] J. Ho, A. V. Parwani, D. M. Jukic, Y. Yagi, L. Anthony, and J. R. Gilbertson, "Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies," *Human pathology*, vol. 37, no. 3, pp. 322-331, 2006.
- [2] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, pp. 331-359, 2013.
- [3] M. Babaie *et al.*, "Classification and retrieval of digital pathology scans: A new dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 8-16.
- [4] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," vol. 42, pp. 60-88, 2017.
- [5] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. J. E. S. w. A. de Leon Ferreira, "Deep learning for biological image classification," vol. 85, pp. 114-122, 2017.
- [6] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks," *arXiv preprint arXiv:1502.02506*, 2015.
- [7] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps*: Springer, 2018, pp. 323-350.
- [8] Z. Wang, J. Chen, S. C. J. I. T. o. P. A. Hoi, and M. Intelligence, "Deep learning for image super-resolution: A survey," 2020.
- [9] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, 2014: Springer, pp. 184-199.
- [10] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094-2107, 2014.
- [11] S. Chetlur *et al.*, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.
- [12] N. BLAS. <http://www.netlib.org/blas/> (accessed).
- [13] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.
- [14] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *International Conference on Machine Learning*, 2015, pp. 1737-1746.
- [15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
- [16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689-696.
- [17] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016: IEEE, pp. 372-387.
- [18] T. Brosch and R. Tam, "Manifold learning of brain MRIs by deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013: Springer, pp. 633-640.
- [19] S. M. Plis *et al.*, "Deep learning for neuroimaging: a validation study," vol. 8, p. 229, 2014.
- [20] A. Payan and G. J. a. p. a. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks," 2015.

- [21] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, 2014: IEEE, pp. 1015-1018.
- [22] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18-31, 2017.
- [23] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International conference on medical image computing and computer-assisted intervention*, 2013: Springer, pp. 411-418.
- [24] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.
- [25] T. Brosch, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014: Springer, pp. 462-469.
- [26] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *Neuroimage*, vol. 124, pp. 127-146, 2016.
- [27] H.-I. Suk, S.-W. Lee, D. Shen, and A. s. D. N. Initiative, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569-582, 2014.
- [28] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fMRI," *NeuroImage*, vol. 129, pp. 292-307, 2016.
- [29] W. H. Pinaya *et al.*, "Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia," *Scientific reports*, vol. 6, p. 38897, 2016.
- [30] S. Bao and A. C. Chung, "Multi-scale structured CNN with label consistency for brain MR image segmentation," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 113-117, 2018.
- [31] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 20-28.
- [32] H. Chen, Q. Dou, L. Yu, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation," *arXiv preprint arXiv:1608.05895*, 2016.
- [33] H. Choi and K. H. Jin, "Fast and robust segmentation of the striatum using deep convolutional neural networks," *Journal of neuroscience methods*, vol. 274, pp. 146-153, 2016.
- [34] A. Karpathy. "CS231n: Convolutional Neural Networks for Visual Recognition." <http://cs231n.github.io/neural-networks-1/#actfun> (accessed).
- [35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2015.
- [36] J. Liu *et al.*, "Applications of deep learning to MRI images: A survey," vol. 1, no. 1, pp. 1-18, 2018.
- [37] T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas Filho, "Performance analysis of google colaboryatory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61677-61685, 2018.
- [38] S. Robertson, H. Azizpour, K. Smith, and J. J. T. R. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," vol. 194, pp. 19-35, 2018.

Biomechanical Prosthesis with EMG Signal Acquisition for Patients with Transradial Amputation

Luis Alberto Huamán Lévano
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
 72043692@continental.edu.pe

Deyby Huamanchahua
Department of Mechatronics Engineering
Universidad Continental
Huancayo, Perú
 dhuamanchahua@continental.edu.pe

Abstract — In the present work, a biomechanical prosthesis with the acquisition of EMG signals will be designed for patients who have lost their right upper extremity due to a transradial forearm amputation, with a short stump before the elbow and an affected psychological factor. For these reasons, the EMG signal acquisition circuit will be designed from operational amplifiers (OPAMPs), which will obtain, filter, and amplify the electromyographic signals acquired from the stump and encoded on the Arduino nano board. being these used to perform the main movements with Maxon motors of pronation and supination in the biomechanical wrist, as well as the movements of the fingers to release and grab an object of the biomechanical hand, where the motor stops turning when it grabs an object if it meets the thermoresistor included in the tip of the index finger. Then, this research will serve for the development of right upper limb prototypes that have the same morphology and modeling of this biomechanical prosthesis, to improve the quality of life of patients who, due to this type of amputation, have been affected in their motor capacity. and psychological.

Keywords—*Biomechanical prosthesis, transradial forearm amputation, EMG, OPAMPs, motor capacity.*

I. INTRODUCTION

Today people are suffering various very severe accidents in their daily lives, which may have repercussions in the not so distant future or at that very moment, they may already lose some part of their body as a result of these incidents, is this, that many times doctors and surgeons come to have the decision to amputate the affected limb and if it is from the forearm to the hand, the so-called "transradial forearm amputation" will be performed, which is a type of amputation where "the extirpates the upper limb from the union between the middle third and the distal third of the forearm, leaving a small stump at the end of the amputated part" [1], therefore this amputation is the most frequent, since "Within of major upper-limb amputations, transradial amputation and radiocarpal disarticulation are performed in about 35% of surgical interventions" [2].

However, sometimes, when patients learn of the loss of their upper limb, they can be psychologically affected by this new lifestyle that they have to live, because "The amputation of an upper limb at any level will have repercussions decisive in the life of the person and most cases it will be unexpected, causing serious functional, aesthetic, psychological and socio-occupational repercussions" [3], then it must be taken into account that a patient who has come out of an intervention surgery like this, apart from having motor deficiencies, it will

be psychologically affected in his daily life, "making prosthetic devices fall into the so-called Uncanny Valley, which can be defined as a feeling of strangeness and non-belonging that users have when they see or use a device that is very similar to something human but that behaves unexpectedly" [4].

Therefore, the objective of this article is to design a biomechanical prosthesis with EMG signal acquisition for patients who have lost their right upper extremity due to a transradial forearm amputation, therefore, by leaving a small stump before the elbow, it is known that "a myoelectric prosthesis is very influential in a transradial level amputation because it makes a greater functional contribution" [2], then this circuit will be composed of operational amplifiers OPAMPs, which by obtaining the various electromyographic signals emitted by the muscles of the forearm, they will carry out various filtering stages to obtain a completely clean and functional signal.

For the acquisition of the EMG signals, a certain process must be followed to obtain them, since "the EMG signals are captured by electrodes on the surface of the skin and are generated by the contraction of the muscles, they range between 5 to 20 mV, and require a signal filtering and amplification system" [5], therefore, it begins with the first stage of initial amplification, where a gain for the input signal will be determined, then in the second stage a Notch filter will be performed, in which the noise in the input signal is rejected, as a third stage, a bandpass filter was used to attenuate the signals that exceed the established frequency ranges, so that the final amplification of the signal is then carried out without inverting it and therefore, as the last stage Offset was established for the interaction of the EMG signals and the Arduino Uno board, achieving the main movements in the biomechanical wrist of pronation and supination. and the finger movements of releasing and grasping an object of the biomechanical hand.

Being for this reason, that in this way it will be possible to support the patient to improve his motor capacity when using the biomechanical forearm prosthesis daily with the acquisition of EMG signals and when presenting an aesthetic finish, it will be able to function normally in his daily life without being affected. for a psychological factor.

II. MATERIALS AND METHOD

For the development of the biomechanical prosthesis, various materials will be used within its structure and composition whose main characteristics are light, resistant,

flexible, and easy to find today, therefore, the materials that are recommended for use in The construction of prostheses are like our prototype must be thermoplastic sheets, silicone, linings, and covers, which will be used in various parts of our biomechanical prosthesis, as described below:

1. Nylon Filament (Polyamides):

This filament is going to be used in 3D printing of the biomechanical hand and forearm since this material is made up of semi-crystalline structures, in addition to this, the filament has a good balance of chemical and mechanical characteristics that offer good stability, rigidity, flexibility, shock resistance and offers a high level of detail.

- Fingers of the hand: The design of each of the fingers (see Fig. 1), was made from the anthropometric measurements obtained, through the following Table I:

TABLE I. AVERAGE VALUES AND DEVIATION OF THE LENGTH OF THE PHALANXES OF THE ADULT HAND

	Phalanx lengths (mm)		
	Proximal phalanx	Medial Phalanx	Distal Phalanx
Little finger	37 ±10	23 ±5	20 ±3
Cancel	45 ±11	30 ±6	23 ±3
Means, medium	57 ±13	37 ±8	26 ±2
Index	50 ±11	32 ±7	25 ±3
Thumb	43 ±6	NA	32 ±4

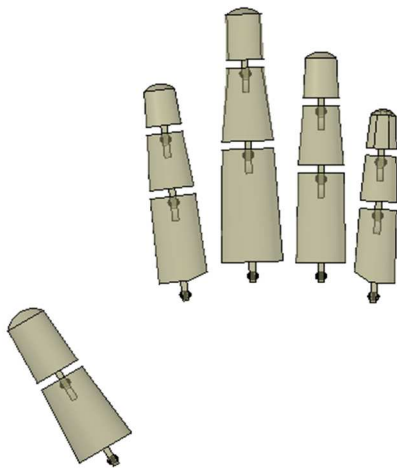


Fig. 1. 3D design in AutoCAD of the five fingers of the biomechanical hand – Type of view: X-Ray

- Hand: Developing the 3D modeling of the hand (see Fig. 2), was made from the data obtained from the following Table II:

TABLE II. HAND AND FOOT DIMENSIONS OF ADULT MEN AND WOMEN

		I	J	K	L
95	in.	8.07	4.63	3.78	9.11
	cm.	20.5	11.8	9.6	23.1
5	in.	7.00	3.92	3.24	7.89
	cm.	17.8	10.0	8.2	20.0

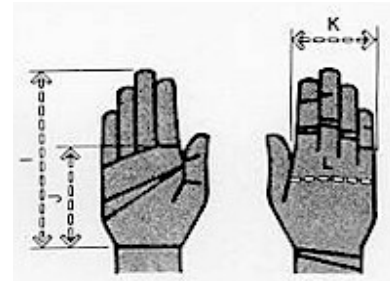


Fig. 2. Hand and foot dimensions of adult men and women, in inches and centimeters, according to percentile selection, with data from (Binvignat et al., 2012)

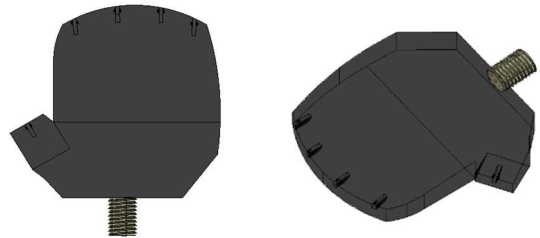


Fig. 3. 3D AutoCAD design of the palm and screw-type socket for the biomechanical wrist – View Type: X-Ray

- Forearm joint: Developing the 3D modeling of the forearm (see Fig. 4):

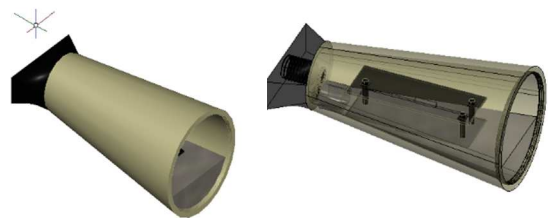


Fig. 4. AutoCAD 3D design of the forearm and accessories for biomechanical wrist movement – View Type: X-Ray and Realistic

2. Polyethylene plates:

This type of material will be used in the connection of the prosthesis with the stump, since having a low density it is a soft and flexible thermoplastic, which can be remodeled by applying hot air to achieve the necessary temperature in the

pressure area that is you want to shape it into a socket for the biomechanical prosthesis.

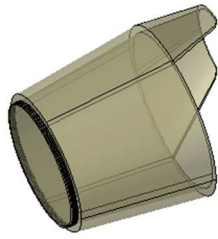


Fig. 5. AutoCAD 3D drawing of the forearm junction and residual limb socket – View Type: X-Ray

3. Silicone - Type Iceross 3S:

It is going to be used as a filler material since it has resistance to humidity, radiation, and ozone, in addition to that, it has great resistance to compression deformation, therefore, a suspension system for the socket will be developed. silicone suction for gripping the prosthesis with the stump.

4. Lining -Pelite:

This material is going to be used as a soft socket for the contracting limb since it is a closed-cell polyethylene foam material available in various durometers, which can be molded in a plaster mold with heat to make a better fit. residual limb comfort.

5. Sheath – Latex, and nylon:

The nylon sheath will be used outside the prosthesis for a more aesthetic finish by coloring the prosthesis according to the patient's skin type and the latex sheath will be used to provide suspension, since being a sheath non-porous, the suspension will be achieved by a combination of suction and mechanical friction between the stump and the prosthesis socket.

6. 3D design of the forearm prosthesis:

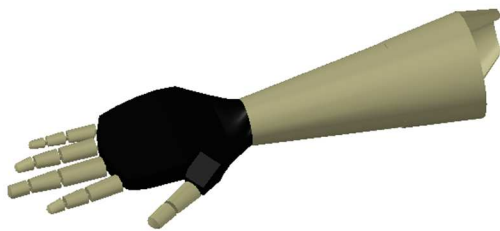


Fig. 6. 3D design in AutoCAD of the reinforced forearm prosthesis – Type of view: Realistic

On the other hand, within the prosthesis, the use of a silver-silver chloride electrode will be taken into account due to its availability, non-invasiveness, and ease of use, since it will be located in the muscle called flexor carpi radialis to perform the main movements in the biomechanical wrist of pronation and supination, utilizing a Maxon 12 mm DCX motor, which is

considered the most suitable actuator for this type of prosthesis, since its movements are carried out without jerks, in addition to presenting high power densities, a wide range of speed values and a long useful life.

Besides, a thermoresistor was used in the tip of the index finger for the control and displacement of the fingers of the biomechanical hand, concerning the signals received and encoded from the voice interface through the Nano Arduino board, these movements are performed by another Maxon 12 mm DCX motor, which according to the rotation mode will be doing the movements of the fingers to release and grasp an object employing industrial rubber bands, used for their high resistance to deformation and a very influential elasticity constant for this type of work.

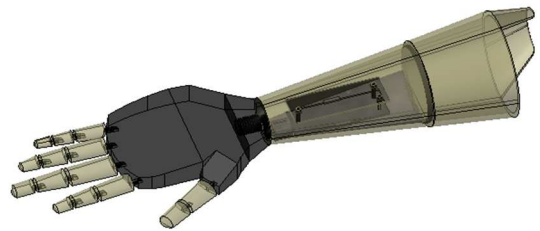


Fig. 7. 3D design in AutoCAD of the reinforced forearm prosthesis – Type of view: X-Ray

III. EMG SIGNAL ACQUISITION CIRCUIT

To develop the EMG signal acquisition circuit, it was made from OPAMPs and calculations mathematical, to determine the gains that we want, thus being able to estimate the values of both the resistors and the capacitors that are going to be included and the cutoff frequencies determined for the various filters that are going to be used, for which it is going to divide to the circuit into six stages, which will be represented in a flow chart.

- **First Stage, Initial Amplification:** As is known, "the amplitude of the EMG signals (electromyography) varies from μV to a low range of mV (less than 10 mV)" [6], therefore, to perform good filtering in the following stages of the circuit and there is not much distortion in the input signal, an instrumentation application stage was carried out, which was done with a Gain (G) of 100. Then, to calculate the gain resistance (R_{G1}), the following formula was used:

$$G = \frac{49,4K\Omega}{R_{G1}} + 1 \tag{1}$$

Using (1), the value of the resistance gain is equal to 498.98 Ω .

- **Second Stage, Notch Filter:** It is designed to reject a frequency that interferes with the circuit, since "myoelectric signals are of low value, noise, or artifacts such as ambient noise or, to a greater extent, line noise (50Hz – 60Hz) can cause a false interpretation of the results" [6], therefore, this filter will reject frequencies of 60 Hz using a cut-off resistor:

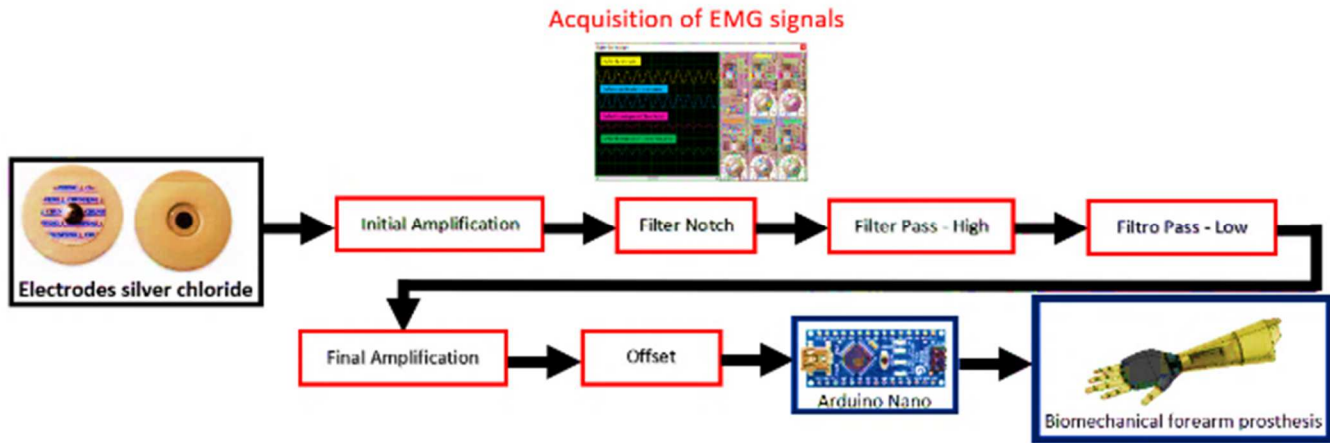


Fig. 8. Flow chart for the acquisition and acquisition of EMG signals for the biomechanical forearm prosthesis

$$F_c = \frac{1}{2\pi RC} \tag{2}$$

Knowing the cutoff frequency (Fc) is 60 Hz, a value of 100 nanofarads (C) is assumed for the two capacitors. Using (2), the values of R₂ and R₃ are 26.52KΩ.

- Second Stage, Bandpass filter:** Here you have to take into account that a “Clean” EMG signal must be obtained, therefore, a bandpass filter will establish a certain range of frequencies and attenuate the rest, since a bandpass “Allows the passage of frequencies between two frequencies ω₁ and ω₂ (ω₁ < ω₂), called lower cutoff frequency and upper cutoff frequency, blocking the rest”[8], then a “ Sallen-Key ” arrangement for this filter would be very valuable due to its simplicity and being established as second-order, not too much noise is generated and with a Butterworth type curve the Butterworth type frequency (Q) was determined, Q = 0.7071.
- Third Stage, High-pass filter:** In this stage, the caps let the high frequencies pass and the resistors are the ones that absorb the low frequencies, having a cutoff frequency (Fc) of 5Hz of Butterworth type and second-order.

$$K = 3 - \frac{1}{Q} \tag{3}$$

To obtain the value of the filter gain (K), in (3) was used which indicates that the value of K = 1.5858. Considering a value of R₅ as 1KΩ.

$$K = \frac{R_5}{R_6} + 1 \tag{4}$$

Using (4), the value of R₆ is 585.8Ω. Knowing the cutoff frequency (Fc) is 60 Hz and the value of 1 nanofarad (C). Using (2), the value of R is 31.831KΩ.

- Fourth Stage, Low-pass Filter:** In this stage, the opposite happens, the resistors allow the low frequencies to pass, and the capacitors are the ones that absorb the high frequencies,

having a cutoff frequency (Fc) of 1.3 kHz and a Sallen-Key arrangement of second-order and Butterworth type.

The value of the filter gain (K) is 1.5858 in (3). Considering a value of R₁₁ as 680Ω in (4), the value of R₁₂ is 394.34Ω. Knowing the cutoff frequency (Fc) is 1.3 KHz and the value of 1.2KΩ (R). Using (2), the value of C is 0.102 nanofarads.

- Fifth Stage, Final Amplification:** For the final amplification stage in the calculation, all the resistances found in the circuit will be needed, and a gain according to it, using a non-inverting amplifier, since it does not want the signal waves to be inverted and it determines the resistance that will be used for the gain.

Considering a value of R₁₄ as 10KΩ in (4), the value of R₁₃ is 3.3KΩ.

- Sixth Stage, OFFSET:** The Arduino is working in a range from 0 to 5 volts, therefore, it is necessary to raise the signal, since this circuit is in a signal of 2.5V in a positive and negative value, for this it will be added to a DC signal, using a potentiometer to introduce 5V to the source and be able to regulate the offset, to the necessary signal, with a gain (G) of 1.

Considering a value of reference resistor (R_f) as 10KΩ, the values of the resistance in parallel (R_p) is 5KΩ and of the design resistance (R_d) is 10KΩ.

Therefore, after having obtained the filtered and amplified EMG signals, they will be transferred from these to the Arduino Uno board, where the voltage ranges will be defined to perform the movements in the biomechanical wrist of pronation and supination, therefore both thanks to the offset, the average voltage that should be exceeded were determined, at the moment that the muscle is contracting and the motor begins to rotate 90° to perform the supination movement, putting the biomechanical hand in a gripping position, therefore, the estimated voltage is 3V or 6mV and in analog value, within the Arduino programming, it was 154.

- H-bridge circuit for Maxon motor:** The main function of this circuit is to allow our Maxon motors to rotate in both directions utilizing BJT transistors and diodes, which will allow the passage of current concerning the digital outputs established by the Arduino Uno board to define the rotation that it wants. and this would be expressed in the following Table III.

TABLE III. Digital values applied to each of the H-bridge inputs

S1	S2	Turn Orientation
1	0	Horary
0	1	Counterclockwise
1	1	Short circuit
0	0	Short circuit

Then, by joining all these simulations, the following EMG signal acquisition circuit would be obtained as a result:

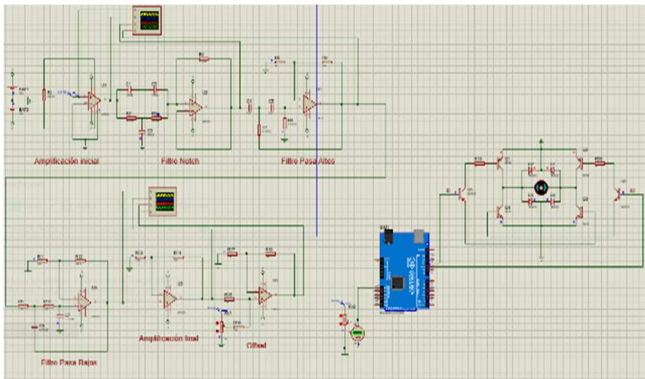


Fig. 9. Design of the entire acquisition circuit processing of simulated EMG signals in Proteus.

IV. DISCUSSION

For a biomechanical design of a prosthesis with the acquisition of EMG signals, it must be considered that “There are two methods to acquire this type of biological signals, the first is the invasive method that requires surgery that involves risks such as pneumothorax, hematoma, and discomfort. The second approach is a non-invasive method or surface EMG.” [9], then by considering the second method as the best approach and incorporating a series of previously stipulated filters into the biomechanical prosthesis, it will be possible to develop more accurate anthropometric movements.

Therefore, the developed prototype managed to analyze the pulses of the stump and amplify them so that they can be analyzed by the microcontroller, this was achieved by using the operational amplifiers that allowed the filter of the pulses since the interference of the environment does not allow an adequate analysis to achieve an appropriate anthropomorphic movement performing proper supination and pronation of the arm, then the correct operation of the following filters can be seen:

- Notch filter:** To attenuate the unwanted signal, the NOTCH filter was used. In one of the tests carried out on this circuit with a signal of 5mV and 60 Hz, it is verified in the red wave, when it is with a frequency of 60 Hz, it was attenuated and therefore the following signals will also be attenuated.

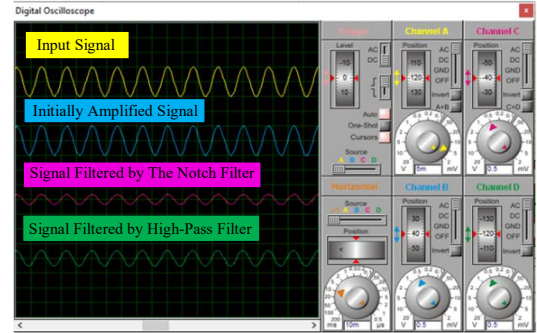


Fig. 10. EMG signals attenuated by the Notch filter and emitted in the oscilloscope simulation in Proteus

- High-Pass Filter:** To attenuate the low-frequency pulses, the High-Pass filter was used, so in this case, a 5 mV and 3 Hz signal was used to test the behavior of this circuit, in the green pulse shown by the oscilloscope such behavior is observed.

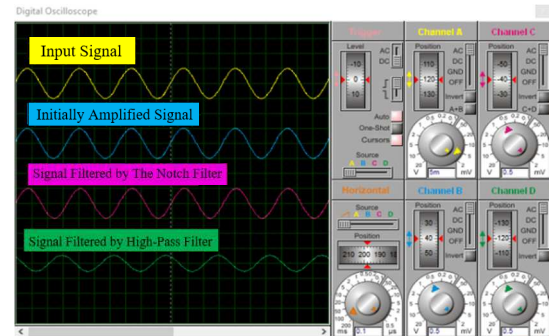
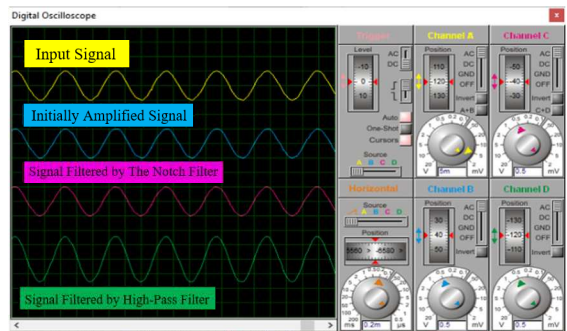


Fig. 11. EMG signals attenuated by the High-Pass filter and emitted in the oscilloscope simulation in Proteus

- Low-Pass Filter:** To attenuate the highest frequencies of the EMG pulses, its low-pass filter behavior was verified by applying a 5mV and 1.6kHz signal and the oscilloscope showed us such behavior in the celestial pulses



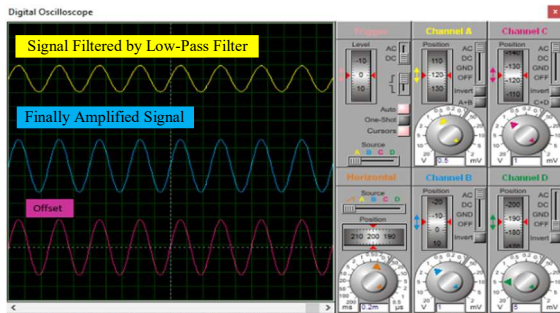


Fig. 12. EMG signals attenuated by the Low-Pass filter and emitted in the simulation of the oscilloscope in Proteus

V. CONCLUSIONS

The acquisition of EMG signals by the forearm prosthesis was established through five stages, the initial amplification stage, Notch filter, bandpass filter, final amplification, and an offset, which together managed to establish filtering optimal for processing these signals on the Arduino Nano board.

The aesthetic design that was used in the forearm prosthesis will allow the patient to restore his psychological state affected by the loss of his right upper limb due to this type of transradial amputation, therefore, the patient will be able to rejoin and carry out his daily life. as if nothing had happened.

This work will undergo possible changes in the future since when this situation improves, EMG signal acquisition tests can be carried out on real patients who have suffered a transradial amputation of the forearm, in addition to this work will be done on the incorporation of a voice interface with IA that will provide the patient with a rapid recovery from the psychological condition suffered by this type of amputation.

VI. REFERENCES

[1] "Amputees - The challenges faced by Gaza-strip amputees in seeking medical treatment [EN/AR/HE] - occupied Palestinian territory", 2016 ReliefWeb. Disponible en: <https://reliefweb.int/report/occupied-palestinian-territory/amputees-challenges-faced-gaza-strip-amputees-seeking-medical>

[2] B. Peerdeman, et al. "Myoelectric forearm prostheses: state of the art from a user-centered perspective", 2011 The Journal of Rehabilitation Research and Development. 2011, 48(6), 719. ISSN 0748-7711, doi:10.1682/jrrd.2010.08.0161

[3] Rehabilitación del amputado de miembro superior, 4 de marzo de 2009. Disponible en: <https://www.andade.es/dra-celia-lopez-cabarcos/item/rehabilitacion-del-amputado-de-miembro-superior>

[4] M. Mori, K. Macdorman and N. Kageki. "The uncanny valley", 2012 IEEE Robotics & Automation Magazine. 2012, 19(2), 98–100. ISSN 1070-9932, doi:10.1109/mra.2012.2192811

[5] K. Akazawa, R. Okuno y M. Yoshida. "Biomimetic EMG-prosthesis-hand", 1996, 18th annual international conference of the IEEE engineering in medicine and biology society. IEEE, 1996. ISBN 078033811, doi:10.1109/iembs.1996.651851

[6] Electromiografía (EMG) - dalcame. Inicio – dalcame, 2005. Disponible en: <https://www.dalcame.com/emg.html#.YieccnrMLIV>

[7] F. Miyara, "Filtros activos", 2004 Sitio Oficial de la F C E I A, abril de 2004. Disponible en: <https://www.fceia.unr.edu.ar/enica3/filtros-t.pdf>

[8] A. Suberbiola et al. "Arm orthosis/prosthesis movement control based on surface EMG signal extraction. International Journal of Neural Systems", 2015, 25(03), 1550009. ISSN 1793-6462, doi:10.1142/s0129065715500094

[9] R. Billions, et al. "Prototyping a prosthetic arm for ulnar and radial deviation", Paper presented at the 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2020, doi:10.1109/HNICEM51456.2020.9400108 Retrieved from www.scopus.com

[10] V. Alvarado, et al. "Adquisición de señales SEMG con electrodos secos para el control de movimiento de dedos en una prótesis robótica fabricada en una impresora 3D", 2019 Ingeniare. Revista chilena de ingeniería, 2019, 27(3), 522–536. ISSN 0718-3305, doi:10.4067/s0718-33052019000300522

[11] C. Ortega, A. Ibarra, E. Viveros and D. Mayorca, "Prototipo Para La Adquisición y Caracterización De Señales Electromiográficas Superficiales Del Movimiento De Flexión-Extensión De Los Dedos De La Mano". 2020 Revista Ibérica De Sistemas e Tecnologías De Informação, 08, pp. 52-64 ProQuest Central. ISSN 16469895.

[12] R. Russo, J. Fernández and R. Rivera, "Algorithm of Myoelectric Signals Processing for the Control of Prosthetic Robotic Hands". 2018 Journal of Computer Science and Technology, 04, vol. 18, no. 1 ProQuest Central. ISSN 16666046. DOI <https://doi.org/10.24215/16666038.18.e04>.

[13] E. Chávez, B. Sifuentes y R. Vidaland, "Processing of myoelectric signals in a microcomputer for identification of movements intention and the cost reduction in the purchase of prosthesis in Peru". 2021 Journal of Physics: Conference Series, 2021, 1780(1), 012035. ISSN 1742-6596, doi:10.1088/1742-6596/1780/1/012035

[14] E. Lopez, R. Méndez y A. Vilchis, "Diseño de una prótesis de mano para uso en teclados con interfaz sEMG". 2019 RECIBE, Revista Electrónica De Computación, Informática, Biomédica y Electrónica, 2019, 8(1), B1-1-1-23. ISSN 2007-5448, doi:10.32870/recibe.v8i1.119.

[15] C. Miozzi, V. Errico, G. Saggio, E. Gruppioni and G. Marrocco, "UHF RFID-Based EMG for Prosthetic Control: preliminary results" 2019 IEEE International Conference on RFID Technology and Applications (RFID-TA), 2019, pp. 310-313, doi: 10.1109/RFID-TA.2019.8891964.

[16] A. Nastuta and C. Agheorghiesei. "Monitoring hand gesture and effort using a low-cost open-source microcontroller system coupled with force sensitive resistors and electromyography sensors". Scopus. 2018, 660, 261–269. ISSN 21945357.

[17] T. Kapelner et al. "Decoding Motor Unit Activity From Forearm Muscles: Perspectives for Myoelectric Control", 2018 IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2018, 26(1), 244–251. ISSN 1558-0210, doi:10.1109/tnsre.2017.2766360

[18] S. Herle, "Movement intention detection from SEMG signals using time-domain features and discriminant analysis classifiers," 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), 2018, pp. 1-6, doi: 10.1109/AQTR.2018.8402774.

[19] S. Yang, Y. Chai, J. Ai, S. Sun and C. Liu, "Hand Motion Recognition Based on GA Optimized SVM Using sEMG Signals," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 2018, pp. 146-149, doi: 10.1109/ISCID.2018.10134.2

[20] M. Dyson, J. Barnes and K. Nazarpour. "Myoelectric control with abstract decoders. Journal of Neural Engineering", 2018, 15(5), 056003. ISSN 1741-2552, doi:10.1088/1741-2552/aacbf6

[21] I. Batzianoulis et al. "Decoding the grasping intention from electromyography during reaching motions. Journal of NeuroEngineering and

Rehabilitation”, 2018, 15(1). ISSN 1743-0003, doi:10.1186/s12984-018-0396-5

[22] M. Markova, D. Shestopalov, and A. Nikolaev. “Estimation of Features Informativeness of the EMG Signal in the Problem of Forearm Prosthesis Controlling”, 2018 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT. SCOPUS, doi:10.1109/USBEREIT.2018.8384548.

[23] K. Schoepp et al. “Design and integration of an inexpensive wearable mechanotactile feedback system for myoelectric prostheses”, 2018 IEEE Journal of Translational Engineering in Health and Medicine. 2018, 6, 1–11. ISSN 2168-2372, doi:10.1109/jtehm.2018.2866105

[24] K. Bakshi, M. Manjunatha y C. KUMAR. “Estimation of continuous and constraint-free 3 DoF wrist movements from surface electromyogram signal using kernel recursive least square tracker”, 2018 Biomedical Signal Processing and Control. 2018, 46, 104–115. ISSN 1746-8094, doi: 10.1016/j.bspc.2018.06.012

[25] L. HAN, P. Kamalanathan and H. Al-dahhan. “Liquid and solids phase backmixing in a bubble and slurry bubble column using a virtual tracer response methodology based on the trajectory data of the radioactive particle tracking (RPT) technique”, 2021 The Canadian Journal of Chemical Engineering. 2021, ISSN 1939-019X, doi:10.1002/cjce.24187

[26] A. Dwivedi et al. “A learning scheme for EMG based decoding of dexterous, in-hand manipulation motions”, 2019 IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019, 27(10), 2205–2215. ISSN 1558-0210, doi:10.1109/tnsre.2019.2936622

[27] J. CAO, T. Zhongming y W. Zhengtao. “Hand gestures recognition based on one-channel surface EMG signal”, 2019 Journal of Software Engineering and Applications. 2019, 12(09), 383–392. ISSN 1945-3124, doi:10.4236/jsea.2019.129023

[28] N. Unanyan and A. Alexey. “Design of upper limb prosthesis using real-time motion detection method based on EMG signal processing. Biomedical Signal Processing and Control”, 2021, 70, 103062. ISSN 1746-8094, doi: 10.1016/j.bspc.2021.103062

[29] J. Mayor et al. “Dexterous hand gestures recognition based on low-density sEMG signals for upper-limb forearm amputees”, 2017 Research on Biomedical Engineering. 2017, 33(3), 202–217. ISSN 2446-4740, doi:10.1590/2446-4740.08516

[30] J. Fan et al. “Improving sEMG-based motion intention recognition for upper-limb amputees using transfer learning”, 2021 Neural Computing and Applications. 2021, ISSN 1433-3058, doi:10.1007/s00521-021-06292-0

Bangla Handwritten Character Recognition Method

Lutfun Nahar

Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
Email: lutfacecu@iiuc.ac.bd

Abstract— Recognizing and extracting handwritten character information is still a challenge in the scanning process. This research describes a method for OCR applications where Bengali handwritten characters can be recognized effectively. This method mainly focuses on post processing steps like feature extraction and building classification model to find a favorable accuracy rate. For feature extraction LBP method is used. Local Binary Pattern is a coming of age feature extracting method which is applied very few times in Bengali Language. This work is also an experimental approach of confirming what occurs when LBP is used in Bengali Characters. To classify Random Forest algorithm is applied, which is also a unique classification method. The datasets are gathered by collecting Bengali characters written in various fashions. Initially, scanned images of Bengali characters are given as input and by applying LBP required features are extracted. Principal Component Analysis (PCA) is applied on the collected feature vectors to reduce the dimension. Finally, RF algorithm is implied on the output to generate a recognition rate. Support vector machine (SVM) is also used as a classifier to evaluate and compare.

Keywords—Local Binary Pattern (LBP);Random Forest Algorithm (RF);Feature Extraction;SVM;Principle Component Analysis (PCA)

I. INTRODUCTION

The method of recognizing handwritten input from photographs, documents and other images by computers is known as handwriting recognition systems. There are online and offline hand writing recognition methods. Online is when characters are recognized at the time of writing. In this method, the pen trajectories and direction of movements are recorded by touch pads and later the writing and recorded directions are used to recognize characters or words. But offline method on the other hand is quite hard ,as the trajectories or directions of the cursives are not available to detect characters. To determine offline characters some factors must be considered like: Aspect ratio, Percent of pixels above the horizontal half way point, Percent of pixels above the vertical half way point ,Number of strokes ,Reflected by Y axis , reflected by X axis and distance away from center.. Postal automation, bank cheque, drafting and all sorts of documentation can be made easier by hand writing recognition systems.

Bengali is the seventh most spoken native language in the world with 205 million speakers. Despite that Bengali language has poor number of machine recognizer for its vast database, complex and look alike characters. It consists of 50 basic characters including 11 vowels (Shoroborno) and 39 consonant (banjonborno) characters and 10 numerals. In Bangla, the concept of upper case or lower case letter is not present. Bangla basic characters have characteristics that differ from other languages [1] It has horizontal lines over some characters called “matra” also it has symbolic forms of consonants which are used alongside the consonants. Since the 80’s the journey for inventing an efficient and accurate handwriting recognition system has begun. It achieved a milestone by creating CEDAR-FOX, first automatic writer recognition system. The trials have not stopped since then by the researchers to establish an almost accurate recognizer for most spoken languages. This thesis paper offers post preprocessing method that can recognize Bengali character from an image. It is an effort to better the Bengali recognition system.

As mentioned earlier, after the basic scanning of image, here Local binary pattern known as LBP method is applied for feature extraction. The reason behind choosing LBP is mainly because the texture based classifications have been an eye catcher lately. This research attempts to discover the outcome of LBP as feature extraction method on Bengali Language. Then the extracted features’ dimensions are reduced using Principle component analysis process. PCA is able to take out the core structures of an image and can quicker the execution time. Then, SVM and Random forest algorithm is applied to build a classification model and generate an accuracy rate. Though both the methods are efficient, Random forest algorithm has shown a better upshot than Support vector machine here.

Thus the present section introduces the problem addressed in this research, while literature review is elaborated in the Section II. Methodology is described in section III. The results are presented in Section IV and Section V concludes the paper.

II. LITERATURE REVIEW

Generous amount of effort has been dedicated for robust BHR methods .The known methods are mostly by using neural networking , HMM, etc. Some unique approaches are in the following: MLP based recognition scheme using stroke features for Bangla basic characters by Bhowmik et

al.[2], here, a stroke-feature-based recognition scheme was proposed, where character images are identified and 10 certain features are extracted from each of them. MLP classifier is trained using back propagation algorithm. Respectively, 350 and 90 sample images are used for 50 Bangla basic characters. Bhattacharya et al.[3] also designed a two-stage MLP based recognition system using shape based features with a large dataset, in 2006. A novel multi stage approach by A.F.R. Rahman in 2001[4], the analysis shows the features, how to detect various high-level features might help to formulate successful multistage OCR designs by proposing multistage scheme for the recognition of Bengali handwritten characters. Super imposed matrix for character recognition by Ahmed Shah Mashiyat in 2004[5] was proposed.

In this system, the Bengali text accepted as an image file and after segmented into each character boundary of each character is determined. The characters are then stored in a 32*32 matrix and the matrix is compared with a knowledge base where all recognized character is stored in superimposed form. Finally depending on the similarity, the system recognizes the character and the accuracy.

III. METHODOLOGY

.Fig.1 illustrates the overview of Bengali handwritten recognition procedure.

This research work is form on the post processing method of Bengali character recognition. After preprocessing conduct the scanned character image is taken as input. Then the features are extracted implementing LBP (Local binary pattern) algorithm and for better features PCA (Principle component analysis) is activated which reduce the dimension of the feature vector

For classification process Random forest algorithm is applied which produce a satisfactory accuracy rate. A diagram is given to explain the procedure of this research in a nutshell. Every step is described into 3 phases

1. Pre-processing
2. Feature extraction.
3. Recognition

It is suggested that the procedure with standard datasets can lead to a favorable output for a Bangla OCR system

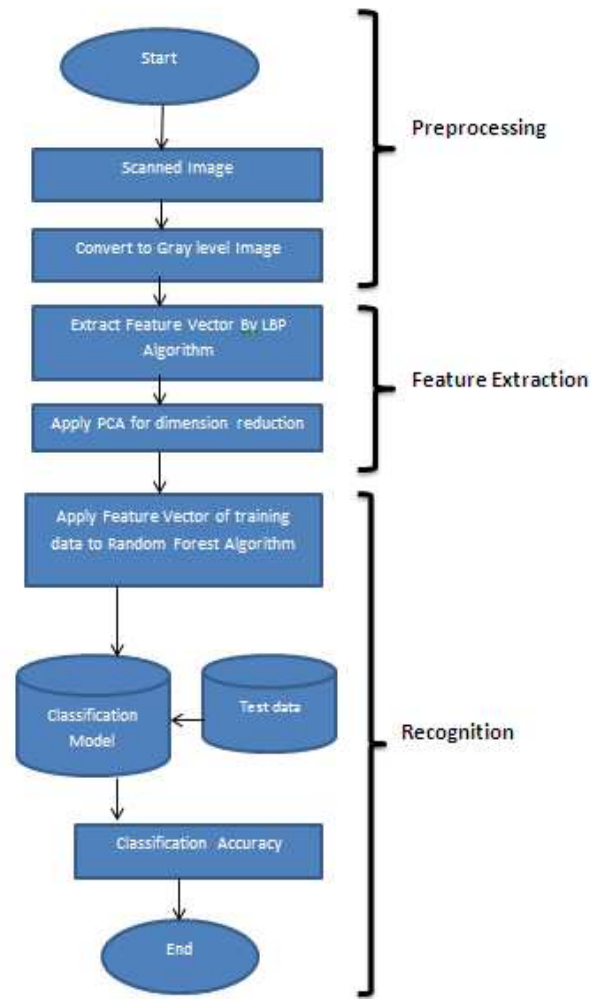


Fig.1.Flowchart of Bangla Handwritten Character recognition

A. LBP

This algorithm is applied here for the required feature extraction phase. It is a method that reconstructs an image into an array that interprets the small scale emergence of the image. Images are possessed of micro patterns and the LBP algorithm works with the uniform patterns of image. The important distinctive of LBP is its heftiness to monotonic gray scale transpose. It is also easy to compute which makes it preferable for real time applications. In this work the basic LBP is applied which works in a 3*3 pixel block by thresh holding its bordering pixel with the value of the center pixel. The output comes into binary 0s that point out the neighboring gray scale value is less than the center value and 1 indicates the opposite. After comparison the binary 8 bit string is converted to decimal number which length is 256, This process is explained with the following mathematical expressions:

IV. RESULTS

A. Datasets

Dataset refers to a collection of images used by researchers to evaluate programs.[11].In this process, the Bengali handwritten character images written in various styles are added and later put into training. Both vowels and consonants are taken also to apply in the system.

The datasets are gathered in a well-organized and systematic manner. Dataset are collected from online Bangla Lekha-Isolated [13], which contains sample images of numeric digits, Bangla characters and compound characters. We used 50 basic Bangla character with 3000 samples of each characters .After preprocessing 1,50,000 handwritten character images were selected. Moreover collection of Bengali characters of different handwriting styles are collected and put into the database. In training phase 70% data are used and 30% are used for validation.

The vowels and consonants have been casted-off separately for recognition. The vowels of Bengali Characters have less differences. So, the classification gets rougher in that zone. The data are tested by choosing random characters and storing them. Later the better given results are taken for observation.

B. Summary of Findings

This paper evaluated the datasets of Bangla handwritten characters by applying into 2 separate classification methods: SVM (support vector machine), Random Forest Algorithm. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given supervised learning data.[12] the algorithm outputs an optimal hyperplane which categorizes new examples. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set. However, like most machine learning algorithms, they are generally applied using a randomly selected training set classified in advance. But in this field so far, the Random forest algorithm has given the most productive results. The foremost reason behind SVM’s poor performance is perhaps the nature of the datasets here. SVM works better with binary classification. In this experiment every character is identified as different classes itself, which is a disadvantage for SVM. On the other hand, RF algorithm’s strongest suit is to work with categorized data.



Fig.4.Some fragments of the Datasets.

Some sample data are given in fig 4 and fig 5.

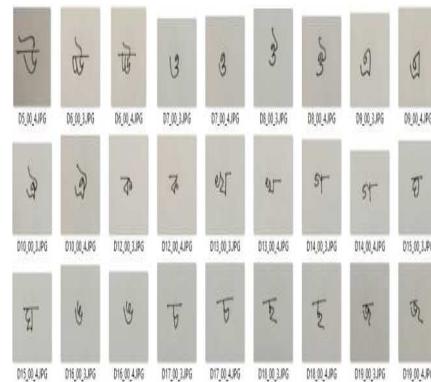


Fig.5. Some portions of training sets.

Hence, the result has gone in the favor of Random Forest algorithm here. Findings are explained in the following table I.

Table I. Accuracy rate by applying SVM.

Class	Accuracy rate
39(consonant)	60%
11(vowel)	77%

Here, five-fold cross validation for the classifier using the parameters C = 100 and gamma = 0.1

Table II. Accuracy rate by applying RF algorithm

Class	Cross validation	Accuracy Rate
39(consonant)	2	78%
11(vowel)	2	80%
39(consonant)	5	80%
11(vowel)	5	85%

C. Interpretation

Despite having an optimistic consequence, the results could be uplifted with better resolution of the character images, vast samples and different values of cross validations.

D. Comparative Study

Every research has their own collection of data sets. So, comparative analysis does not harvest an evocative outcome. There are pre-conditions like addressing noise detection and cleaning phase in some papers. However, a comprehensive solution for leaving zero noise in images is not obtained yet. In the discussed paper, SVM and Random forest algorithm both are applied as classification method to compare and examine which method produce a better result. Same set of data sets are implied to them and cross validated using number of testing and training sets.

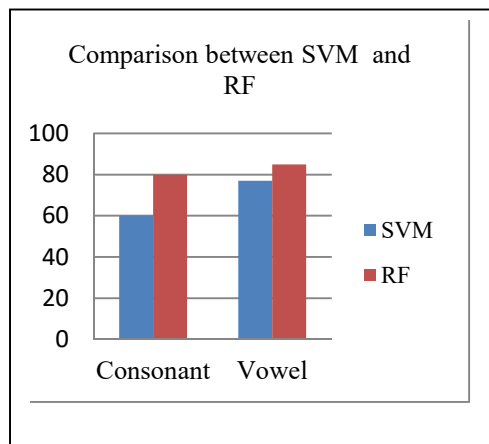


Fig.6. Comparing SVM with RF algorithm for various datasets

Comparison between the SVM and RF algorithm while using same feature extraction and dataset is given above in fig 6.

E. Accomplishments:

The aim of this paper is to assist a OCR system for developing an operative recognition model, and it can be said that, the objective was successfully met. In this effort, a new and completely unique choice of methods is used. Previous works on Bengali character recognitions use methods like HMM (Hidden Markov Model), Free chain code for feature extracting, and ANN (Artificial neural network), MLP (Multi-layer Perception) for classification. But here for feature extraction LBP (Local Binary Pattern) is used and Random Forest Algorithm is used for classification schemes. RF algorithm is used for recognizing handwritings

of various international languages but perhaps not for Bengali Language. The purpose of this exertion is to observe

what outcomes they could bring combined. 85% accuracy is gained applying LBP and RF method with a minimum error.

V. CONCLUSION AND FUTRE WORK

Regardless of being one of the most spoken languages, Bengali language has poor number of machine recognizer for its vast database, cursives and similar nature symbols. Bengali handwritten character recognition can facilitate the progress in: Postal automation, Bank draft, Passport verification, Signature approving etc. In fine, all types of documented works that needed to be computerized can be recognized with Bengali Language also, if sufficient progress in this field has been made. There is ample of works done already on English, Chinese and Hindi. Despite of being one of the most spoken languages the digitalization of Bengali language is yet to be made fully. The aim of this thesis is to achieve satisfactory recognition rate to narrow down this gap. In future we will work on Compound Characters and also attention can be compensated for angular alphabets.

REFERENCES

- [1] Nawab, N.B. and Hassan, M.M., 2012, May. Optical Bangla character recognition using chain-code. In Informatics, Electronics & Vision (ICIEV), 2012 International Conference on (pp. 622-627). IEEE.
- [2] Bhowmik, T.K., Bhattacharya, U., Parui, S.K.: Recognition of Bangla handwritten characters using an MLP classifier based on stroke features. In: Proceedings of the 11th International Conference on Neural Information Processing (ICONIP), India, pp. 814– 819 (2004)
- [3] Bhattacharya, U., Parui, S.K., Shaw, B. and Bhattacharya, K., 2006, October. Neural combination of ANN and HMM for handwritten Devanagari numeral recognition. In Tenth international workshop on frontiers in handwriting recognition.
- [4] Rahman, F.R.,Rahman,R.,Fairhurst,M.C.:Recognition of Handwritten Bengali characters: A novel multistage approach. Pattern Recognition. 35(3), 997–1006 (2002)
- [5] Mashiyat, Ahmed Shah, Ahmed Shah Mehadi, and Kamrul Hasan Talukder. "Bangla off-line handwritten character recognition using superimposed matrices." Proc. 7th International Conf. on Computer and Information Technology. 2004.
- [6] Hassan, T. and Khan, H.A., 2015, May. Handwritten bangla numeral recognition using local binary pattern. In Electrical Engineering and Information Communication Technology (ICEEICT), 2015 .
- [7] https://research.aston.ac.uk/portal/files/116199/NCRG_97_010.pdf
- [8] http://www.montefiore.ulg.ac.be/~kbessonov/present_data/GBIO0002-1_GenAndBioinf2015-16/lectures/L6/Suppl2_Ringner2008.pdf
- [9] Horning, N., 2010. Random Forests: An algorithm for image classification and generation of continuous fields data sets. New York.
- [10] <http://amateurdatascientist.blogspot.com/2012/01/random-forest-algorithm.html>
- [11] <http://datasets.visionbib.com/info-index.html#TT19>.
- [12] Tong, S. and Koller, D., 2001. Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov), pp.45-66.
- [13] Mar 2017. [Online]. Available:<https://www.sciencedirect.com/science/article/pii/S2352340917301117>

The Iso-RSA Cryptographic Scheme

Mamadou I. Wade
 Department of Electrical Engineering and Computer Science
 Howard University
 Washington, DC, U.S.A
 wademamadoui@gmail.com or mamadou.wade@howard.edu

Abstract—This paper shows one way that can be used to strengthen the RSA cryptographic scheme which was published in 1978 by Ron Rivest, Adi Shamir, and Len Adleman. We use the isonumber theory of the first kind which was developed by Santilli where the unit one of normal arithmetic is replaced with a new unit called an isounit. The proposed cryptosystem possesses both characteristics of symmetric and asymmetric encryption and therefore can be seen as a hybrid public-key and private-key cryptographic system that benefits from both types of the cryptographic method with a substantial increase in security of the associated ciphers. Additionally, applications and implementations of the Iso-RSA cryptosystem for the encryption and decryption of information encoded into positive integers or binary digits of 0s or 1s are implemented, and simulation results show that our proposed cryptographic system produces highly secure encrypted information. A wide range of other applications in security mechanisms and security services which include data confidentiality, authentications, data integrity, digital signature, key exchange, and others can be implemented using our proposed Iso-RSA cryptographic scheme.

Index Terms—Iso-RSA, Iso-RSA Cryptographic Scheme, RSA, Iso-RSA Encryption Function, Iso-RSA Decryption Function, Isonumber Cryptosystems.

I. INTRODUCTION

A Number of public-key cryptographic schemes have been proposed in the literature during the past few decades. One of the ideas behind a public-key cryptographic system is to have two keys, one made public and accessible to anyone and can be used to encrypt a secret message, while the other key is kept private and can be used for decryption to recover the original message. Among the public-key cryptographic schemes, one can find the RSA cryptographic scheme which was published in 1978 by Ron Rivest, Adi Shamir, and Len Adleman. The RSA is one of the most widely used public-key cryptographic systems, and it is a block cipher whereby both the plaintext and ciphertext consist of integers in the range 0 to $n - 1$ for some positive integer n [9], [11]. Another public-key cryptographic scheme proposed in 1984 is the ElGamal cryptographic system; it is based on the Diffie-Hellman key exchange protocol [8], [10], and its security rest on the

difficulty of solving a Discrete Logarithms (DL) problem over finite fields. Among public-key cryptographic schemes, one can also include the Paillier cryptographic system proposed in 1999, and it is based on the composite residuosity class problem [5]. This paper also includes concepts of the number theory of the first kind which was proposed by Santilli [6], [7], whereby the unit 1 of normal arithmetic is replaced by a new unit called an isounit.

In our proposed cryptographic approach, we set the isounit to be both an encryption and decryption secret isokeys as is the case in symmetric cryptosystems. In addition, when the secret isokey is randomly generated, the Iso-RSA cryptographic scheme is a probabilistic and asymmetric cryptographic scheme with a public encryption isokey and a private decryption isokey; therefore, can be seen as a hybrid public-key and private-key cryptographic system. Because of its extremely large isokey sizes, which can be more than 3,000 digits or about 10,000 bits, with a minimum added computational cost, the Iso-RSA cryptosystem produces highly secure encrypted information.

II. SANTILLI'S THEORY OF ISONUMBERS

A brief introduction to Santilli's Theory of Isonumbers is provided in this section. This theory was discovered and used by Santilli in his representation of antiparticles as particles moving backward in time, and it provides a rigorous foundation for an idea proposed by Feynman [1], [6], [7].

The real numbers is a field where 0 is the unit for addition and 1 is the unit for multiplication, and it can be represented by a five-tuple $(\mathbb{R}, +, \cdot, 0, 1)$. Santilli observed a sense of direction from left to right for \mathbb{R} that mathematicians implicitly use, so it could be represented as a six-tuple $(\mathbb{R}, +, \cdot, 0, 1, \rightarrow)$. So, practically our real number system is not symmetric. In the Isonumber Theory of the Second Kind, Santilli's proposed a new representation of the real numbers $(\hat{\mathbb{R}}, +, *, 0, \hat{1}, \leftarrow)$ which he called the isotopic dual representation of the real numbers, where the unit for addition remained 0, but the new unit for multiplication also called isounit is $\hat{1} = -1$ and is in the field (i.e. $(-1) = \hat{I} \in \mathbb{R}$). Thus,

$$\hat{a} * \hat{b} = (-a)(-1)(-b) = -ab = \hat{ab}. \quad (1)$$

Our interest in this paper is in Santilli's Isonumber Theory of the First Kind, where the new unit or isounit \hat{I} is not in the field (i.e. $\hat{I} \notin \mathbb{F}$). Let the field be given by $\mathbb{F}(a, +, \cdot)$ (for instance, Z_p , with p prime) and let the isounit $\hat{I} = \hat{T}^{-1} \notin \mathbb{F}$ be an invertible quantity. Let a new definition of multiplication be defined on the field \mathbb{F} using $*$ = $\hat{T} = \hat{I}^{-1}$ [1]. Now we can define a new field which is an isofield of the first kind $\hat{\mathbb{F}}(\hat{a}, +, *)$, with elements called isonumbers and rules for multiplication and addition given by:

$$\begin{aligned} \hat{a} &= a\hat{I}, \\ \hat{a} * \hat{b} &= (a\hat{I})\hat{I}(b\hat{I}) = ab\hat{I} = \widehat{ab} \\ \hat{a} + \hat{b} &= (a\hat{I}) + (b\hat{I}) = (a+b)\hat{I} = \widehat{a+b}. \\ \hat{0} &= 0\hat{I} = 0 \end{aligned} \quad (2)$$

III. UNDERLYING MATHEMATICS

This section provides explanations about the underlying mathematics needed to develop and construct the Iso-RSA cryptographic scheme. It includes the Isodivision Algorithm, the Euclidean-Senegalese Algorithm, the Extended Euclidean-Senegalese Algorithm, and the Euler-Senegalese Isototient Function.

A. Isodivision Algorithm

Recall that the division algorithm describes the relation obtained when a nonnegative integer a is divided by a positive integer n . This relation is given by $a = q \times n + r$, where the remainder r and the quotient q satisfy $0 \leq r < n$ and $q = \lfloor \frac{a}{n} \rfloor$, the symbols $\lfloor x \rfloor$ gives the largest integer less than or equal to x , which is also called the Floor of x .

The isodivision algorithm is an equivalent version of the division algorithm using isonumbers.

Proposition 1: Isodivision Algorithm

Let \hat{n} be a positive isointeger and \hat{a} be any nonnegative isointeger, and let \hat{I} be the isounit, if \hat{a} is isodivided by \hat{n} , we obtain an isointeger quotient \hat{q} and an isointeger remainder \hat{r} satisfying the following equation:

$$\hat{a} = \hat{q} * \hat{n} + \hat{r} = \hat{q}\hat{I}\hat{n} + \hat{r} \quad (3)$$

where

$\hat{0} \leq \hat{r} < \hat{n}$; $\hat{q} = \lfloor \frac{\hat{a}}{\hat{n}} \rfloor \hat{I}$ and $*$ = $\hat{T} = \hat{I}^{-1}$. The symbol $\lfloor \hat{x} \rfloor$ represents the largest isointeger less than or equal to \hat{x} . Eq. (3) can also be written as

$$\begin{aligned} \hat{a} &= \hat{q} * \hat{n} + \hat{r} \\ \hat{a} &= \hat{q}\hat{I}^{-1}\hat{n} + \hat{r} \\ \hat{a} &= \lfloor \frac{\hat{a}}{\hat{n}} \rfloor \hat{I}\hat{I}^{-1}\hat{n} + \hat{r} \\ \hat{a} &= \lfloor \frac{\hat{a}}{\hat{n}} \rfloor \hat{n} + \hat{r} \end{aligned} \quad (4)$$

Note that the isodivision algorithm is a generalization of the division algorithm whereby the unit 1 for multiplication is replaced by the isounit \hat{I} . If the isounit \hat{I} in (3) and (4) is replaced by 1, and the $*$ operator is replaced by the regular

multiplication operator \times , we obtain the traditional division algorithm.

B. Euclidean-Senegalese Algorithm

The Euclidean-Senegalese algorithm can be used to find the greatest common isodivisor (*gcd*), \hat{d} , of two isointegers \hat{a} and \hat{b} .

Definition: mod Operator

Given an isointeger \hat{a} and a positive isointeger \hat{n} , the remainder \hat{r} when \hat{a} is isodivided by \hat{n} is defined as $\hat{a} \bmod \hat{n}$. we write

$$\hat{r} = \hat{a} \bmod \hat{n} \quad (5)$$

Using mod operator, the isodivision algorithm in (4) can be written as

$$\hat{a} = \lfloor \frac{\hat{a}}{\hat{n}} \rfloor \hat{n} + \hat{a} \bmod \hat{n} \quad (6)$$

TABLE I: The Euclidean-Senegalese Algorithm

Assume that $\hat{a} \geq \hat{b} > \hat{0}$		
Remainder (\hat{r}_i)	Quotient (\hat{q}_i)	Isodivision Algorithm Verified
$\hat{r}_1 = \hat{a} \bmod \hat{b}$	$\hat{q}_1 = \lfloor \frac{\hat{a}}{\hat{b}} \rfloor \hat{I}$	$\hat{a} = \hat{q}_1 * \hat{b} + \hat{r}_1$ $= \lfloor \frac{\hat{a}}{\hat{b}} \rfloor \hat{b} + \hat{r}_1$; $\hat{0} < \hat{r}_1 < \hat{b}$
$\hat{r}_2 = \hat{b} \bmod \hat{r}_1$	$\hat{q}_2 = \lfloor \frac{\hat{b}}{\hat{r}_1} \rfloor \hat{I}$	$\hat{b} = \hat{q}_2 * \hat{r}_1 + \hat{r}_2$ $= \lfloor \frac{\hat{b}}{\hat{r}_1} \rfloor \hat{r}_1 + \hat{r}_2$; $\hat{0} < \hat{r}_2 < \hat{r}_1$
$\hat{r}_3 = \hat{r}_1 \bmod \hat{r}_2$	$\hat{q}_3 = \lfloor \frac{\hat{r}_1}{\hat{r}_2} \rfloor \hat{I}$	$\hat{r}_1 = \hat{q}_3 * \hat{r}_2 + \hat{r}_3$ $= \lfloor \frac{\hat{r}_1}{\hat{r}_2} \rfloor \hat{r}_2 + \hat{r}_3$; $\hat{0} < \hat{r}_3 < \hat{r}_2$
⋮	⋮	⋮
$\hat{r}_{n-1} = \hat{r}_{n-3} \bmod \hat{r}_{n-2}$	$\hat{q}_{n-1} = \lfloor \frac{\hat{r}_{n-3}}{\hat{r}_{n-2}} \rfloor \hat{I}$	$\hat{r}_{n-3} = \hat{q}_{n-1} * \hat{r}_{n-2} + \hat{r}_{n-1}$ $= \lfloor \frac{\hat{r}_{n-3}}{\hat{r}_{n-2}} \rfloor \hat{r}_{n-2} + \hat{r}_{n-1}$ $\hat{0} < \hat{r}_{n-1} < \hat{r}_{n-2}$
$\hat{r}_n = \hat{r}_{n-2} \bmod \hat{r}_{n-1}$	$\hat{q}_n = \lfloor \frac{\hat{r}_{n-2}}{\hat{r}_{n-1}} \rfloor \hat{I}$	$\hat{r}_{n-2} = \hat{q}_n * \hat{r}_{n-1} + \hat{r}_n$ $= \lfloor \frac{\hat{r}_{n-2}}{\hat{r}_{n-1}} \rfloor \hat{r}_{n-1} + \hat{r}_n$ $\hat{0} < \hat{r}_n < \hat{r}_{n-1}$
$\hat{r}_{n+1} = \hat{r}_{n-1} \bmod \hat{r}_n$ $= \hat{0}$	$\hat{q}_{n+1} = \lfloor \frac{\hat{r}_{n-1}}{\hat{r}_n} \rfloor \hat{I}$	$\hat{r}_{n-1} = \hat{q}_{n+1} * \hat{r}_n + \hat{0}$ $= \lfloor \frac{\hat{r}_{n-1}}{\hat{r}_n} \rfloor \hat{r}_n + \hat{0}$ $gcd(\hat{a}, \hat{b}) = \hat{d} = \hat{r}_n$

Using (6) result, (5) can also be written as

$$\hat{r} = \hat{a} \bmod \hat{n} = \hat{a} - \lfloor \frac{\hat{a}}{\hat{n}} \rfloor \hat{n} \quad (7)$$

Proposition 2:

Given a nonnegative isointeger \hat{a} , a positive isointeger \hat{b} and assuming that $\hat{a} \geq \hat{b} > \hat{0}$, we have the following relation:

$$gcd(\hat{a}, \hat{b}) = gcd(\hat{b}, \hat{a} \bmod \hat{b}) \quad (8)$$

Proposition 3: Euclidean-Senegalese Algorithm

The Euclidean-Senegalese algorithm allows us to find the greatest common isodivisor (*gcd*), \hat{d} , of two isointegers \hat{a} and \hat{b} , or $\hat{d} = gcd(\hat{a}, \hat{b})$. To find \hat{d} , the isodivision algorithm in (3), or its equivalent version in (5), or (8) can be applied until a zero remainder is obtained at some iteration $n+1$. This process of finding $\hat{d} = gcd(\hat{a}, \hat{b})$ is summarized in Table I. As can be seen in Table I, when a zero remainder is obtained at iteration $n+1$, we have found the *gcd* of \hat{a} and \hat{b} , and is given by

$$gcd(\hat{a}, \hat{b}) = \hat{d} = \hat{r}_n. \quad (9)$$

This approach is a systematic way of finding the *gcd* of two isointegers \hat{a} and \hat{b} , and it is very useful when isointegers are encrypted using the Iso-RSA cryptographic scheme.

Note that the traditional Euclidean algorithm is a special case of the Euclidean-Senegalese algorithm when the isounit \hat{I} is set to 1 and the $*$ = $\hat{T} = \hat{I}^{-1}$ operator is replaced with the multiplication operator \times , and the isonumbers are replaced with integer counterparts.

C. Extended Euclidean-Senegalese Algorithm

The Extended Euclidean-Senegalese algorithm allows us to find not only the greatest common isodivisor (*gcd*), \hat{d} , of two isointegers \hat{a} and \hat{b} , but also two isointegers \hat{x} and \hat{y} satisfying the following equation.

$$\hat{a} * \hat{x} + \hat{b} * \hat{y} = \hat{d} = gcd(\hat{a}, \hat{b}) \quad (10)$$

The Extended Euclidean-Senegalese algorithm is a generalization of the classical Extended Euclidean algorithm whereby unit 1 is replaced by the isounit \hat{I} and the traditional multiplication operator \times is replaced by the operator $*$ = $\hat{T} = \hat{I}^{-1}$, in addition to other minor adjustments.

In order to find the two isointegers \hat{x} and \hat{y} satisfying (10), assume that at each iteration step, i , of the Euclidean-Senegalese algorithm in Table I, we can compute \hat{x}_i and \hat{y}_i satisfying the following equation,

$$\hat{r}_i = \hat{a} * \hat{x}_i + \hat{b} * \hat{y}_i \quad (11)$$

The recurrence relations for \hat{x}_i and \hat{y}_i needed to construct the Extended Euclidean-Senegalese algorithm are given by

$$\hat{x}_i = \hat{x}_{i-2} - \hat{I}^{-1} \hat{q}_i \hat{x}_{i-1} \quad (12)$$

and

$$\hat{y}_i = \hat{y}_{i-2} - \hat{I}^{-1} \hat{q}_i \hat{y}_{i-1} \quad (13)$$

Equations (12) and (13) provide the relations we need in order to find \hat{x}_i and \hat{y}_i at each iteration step on the Extended Euclidean-Senegalese algorithm that will be discussed next in Table II.

Similar to the approach used to construct the Euclidean-Senegalese algorithm described in Table I, (8) can be applied until a zero remainder is obtained at iteration $n+1$. One of the goal is to find \hat{x}_i and \hat{y}_i at each iteration.

At iteration $i = -1$.

We initialize $\hat{r}_{-1} = \hat{a}$. So, (11) can be written as

$$\hat{r}_{-1} = \hat{a} * \hat{x}_{-1} + \hat{b} * \hat{y}_{-1} = \hat{a} \quad (14)$$

Eq. (14) implies that $\hat{y}_{-1} = \hat{0}$ and $\hat{x}_{-1} = \hat{I}$ since $\hat{a} * \hat{I} = \hat{a}$. At iteration $i = 0$.

We initialize $\hat{r}_0 = \hat{b}$. So, (11) can be written as

$$\hat{r}_0 = \hat{a} * \hat{x}_0 + \hat{b} * \hat{y}_0 = \hat{b} \quad (15)$$

Eq. (15) implies that $\hat{x}_0 = \hat{0}$ and $\hat{y}_0 = \hat{I}$ since $\hat{b} * \hat{I} = \hat{b}$.

The results for iterations $i = 1$ through $i = n+1$ are summarized in Table II.

Note that columns 3 and 4 of Table II are calculated using (12) and (13), respectively; while column 5 is obtained from (11).

As can be seen on the results on the last row of Table II, the iteration process ends when a remainder of zero is obtained at some iteration $i = n+1$. At this iteration, we have found the *gcd*(\hat{a}, \hat{b}) which is equal to $\hat{d} = \hat{r}_n$, the isointeger $\hat{x} = \hat{x}_n$, and $\hat{y} = \hat{y}_n$, all satisfying (10).

In addition if $gcd(\hat{a}, \hat{b}) = \hat{I}$, then we also have found the isomultiplicative inverse of \hat{b} in (10), which is given by $\hat{b}^{-1} \text{ mod } \hat{a} = \hat{y} = \hat{y}_n$.

In order to illustrate the use of the Euclidean-Senegalese algorithm in Table I, and the Extended Euclidean-Senegalese algorithm presented in Table II, consider the following example. Given $\hat{a} = \hat{20}$, $\hat{b} = \hat{7}$, and $\hat{I} = \hat{30}$, find the *gcd*(\hat{a}, \hat{b}) = \hat{d} , the value of \hat{x} , \hat{y} , and the isomultiplicative inverse of $\hat{b} \text{ (mod } \hat{a})$ if applicable(when $\hat{d} = \hat{I}$). Results of this example is summarized in Table III .

Also, note that the isointegers $\hat{a} = \hat{20}$ and $\hat{b} = \hat{7}$ are relatively isoprime, so we would expect their *gcd* to be equal to \hat{I} as shown in the result obtained in Table III.

Also note that $\hat{a} \text{ mod } \hat{b} = a\hat{I} \text{ mod } b\hat{I} = (a \text{ mod } b)\hat{I}$ implies that $gcd(\hat{a}, \hat{b}) = gcd(a, b)\hat{I} = d\hat{I} = \hat{d}$, where $d = gcd(a, b)$.

D. Euler-Senegalese's Isototient Function

The Euler-Senegalese's isototient function is the equivalent version of the Euler's totient function when using isomubers. The Euler-Senegalese's isototient function generalized the Euler's totient function when the unit 1 is replace by the isounit \hat{I} and the multiplication operator \times is replaced with the operator $*$ = $\hat{T} = \hat{I}^{-1}$. So, the Euler's totient fuction is a special case of the Euler-Senegalese's isototient function. We can recall the Euler's totient function as follows:

Definition: Euler's Totient Function ($\phi(n)$)

The Euler's totient function is defined as the number of positive integers that are less than n and relatively prime to n . Its notation is $\phi(n)$, and we define $\phi(1) = 1$ [10].

Note that if p and q are two distinct prime numbers and $n = p \times q$, then

$$\phi(p) = (p - 1) \quad (16)$$

$$\phi(n) = \phi(p \times q) = \phi(p)(q) = (p - 1) \times (q - 1) \quad (17)$$

Definition: Euler-Senegalese's Isototient Function ($\phi(\hat{n})$)
 Given the isounit \hat{I} , we define the Euler-Senegalese's isototient function as the corresponding isonumber of the number of positive isointegers that are less than \hat{n} and relatively isoprime to \hat{n} . Its notation is $\phi(\hat{n})$, and we define $\phi(\hat{I}) = \hat{I}$.

Proposition 6

Consider the following relations:

It \hat{p} is an isoprime number, then

$$\phi(\hat{p}) = \hat{p} - \hat{I} \quad (18)$$

If \hat{p} and \hat{q} are two isoprime numbers that are distinct, and $\hat{n} = \hat{p} * \hat{q}$, then

$$\phi(\hat{n}) = \phi(\hat{p} * \hat{q}) = \phi(\hat{p}) * \phi(\hat{q}) = (\hat{p} - \hat{I}) * (\hat{q} - \hat{I}) = (\hat{p} - \hat{I})\hat{I}^{-1}(\hat{q} - \hat{I}) \quad (19)$$

If $\hat{n} = \hat{p} * \hat{q}$, where \hat{p} and \hat{q} are two isoprime numbers, then the Euler-Senegalese's Isototient function $\phi(\hat{n})$ is related to the Euler's Totient $\phi(n)$ by

$$\phi(\hat{n}) = \phi(n)\hat{I} \quad (20)$$

The cost of implementing (20) is less than the cost implementing (19).

IV. THE ISO-RSA CRYPTOGRAPHIC ALGORITHM

This section describes our proposed Iso-RSA cryptographic algorithm, as well as the traditional RSA Algorithm. We introduce the traditional RSA encryption scheme and the extensions that lead to the Iso-RSA scheme.

A. *Traditional RSA Cryptographic System*

The traditional RSA cryptographic system was published by Ron Rivest, Adi Shamir, and Len Adleman in 1978. The RSA algorithm has stood the test of time in terms of security and efficiency of computation, especially when implemented using the fast exponentiation algorithm or other similar methods; it is one of the most widely used public-key cryptographic systems. The RSA is a block cipher whereby both the ciphertext (C) and plaintext (M) consist of integers between 0 and $n - 1$. The RSA algorithm consists of a key generation step, an encryption step, and a decryption step. Each of these steps is as follows [9], [10].

1) **The RSA Public and Private Key Generation:** Two prime numbers p and s are randomly selected in order to compute the public key $n = p \times s = ps$. For the encryption scheme to be computationally secure, p and s must be as large as possible so that factoring n to obtain p and s is computationally infeasible. This value of n is used to compute the Euler's Totient function defined by (17) as $\phi(n) = \phi(p \times s) = \phi(p) \times \phi(s) = (p - 1) \times (s - 1)$. Another integer e greater than 1, relatively prime to $\phi(n)$, and less than

$\phi(n)$ is randomly selected. So, we have the greatest common divisor (gcd) of $\phi(n)$ and e is 1; or $gcd(\phi(n), e) = 1$ and $1 < e < \phi(n)$. Now, the public key needed for encryption is a pair of positive integers given by $\{n, e\}$. It is important to note that the greatest common divisor of two integers can be computed using the Euclidean algorithm.

To find the private key needed for decryption, we compute the integer f that is the multiplicative inverse of $e \pmod{\phi(n)}$; which can be written as $e \times f \equiv 1 \pmod{\phi(n)}$, or $f \equiv e^{-1} \pmod{\phi(n)}$, where the symbol \equiv represents the congruent relation. We also note that $f < \phi(n)$, and is relatively prime to $\phi(n)$, meaning that $gcd(\phi(n), f) = 1$. The private key needed for decryption is given by the pair of positive integers $\{n, f\}$. We can also point out that f can be computed using the Extended Euclidean algorithm.

2) **The RSA Encryption Function:** The RSA encryption function is used to encrypt a message or plaintext represented by an integer M between 0 and $n - 1$. The encryption of the message M using the encryption function, E , the modulo operator $\pmod{}$, in conjunction with the public key $\{n, e\}$, to obtain the cipher C , is given by

$$C \equiv E(M) \equiv M^e \pmod{n} \quad (21)$$

3) **The RSA Decryption Function:** The RSA decryption function is used to decrypt the ciphertext represented by integer C between 0 and $n - 1$. The decryption of the ciphertext C using the decryption function, D , in conjunction with the private key f to obtain the message M , is given by

$$M \equiv D(C) \equiv C^f \pmod{n} \quad (22)$$

B. *Iso-RSA Cryptographic System*

This subsection presents in detail the Iso-RSA Cryptographic System. It is also a three steps process consisting of a key generation step, an encryption step, and a decryption step. At this point, we have developed the mathematical tools needed to carry out our cryptographic process in the context of isomubers; these mathematical tools could not be found anywhere in the literature, leading us to develop and apply them to our cryptographic problem. These mathematical tools are discussed in the underlying mathematics section, and they include Euclidean-Senegalese Algorithm, Extended Euclidean-Senegalese Algorithm, and the Euler-Senegalese's isototient function, and others.

1) **The Iso-RSA Public and Private key Generation:** The public and private isokeys generation for the Iso-RSA cryptographic system are described in this section. First, we randomly generate two prime numbers p and s , and randomly generate the isounit \hat{I} greater than the largest values of the message to be encrypted. After computing $\hat{p} = p\hat{I}$ and

$\hat{s} = s\hat{I}$, we calculate the public encryption isockey \hat{n} as

$$\hat{n} = \hat{p} * \hat{s} = \hat{p}\hat{T}\hat{s} = \hat{p}\hat{T}^{-1}\hat{s} = p\hat{I}\hat{T}^{-1}s\hat{I} = (ps)\hat{I} = n\hat{I} \quad (23)$$

If \hat{n} is known, the RSA public encryption key n can be obtained by $n = \hat{n}\hat{I}^{-1}$. Now we compute the Euler's isototient function given by (19) and summarized as

$$\phi(\hat{n}) = (\hat{p} - \hat{I})\hat{I}^{-1}(\hat{s} - \hat{I}) \quad (24)$$

Another way of computing $\phi(\hat{n})$ is $\phi(\hat{n}) = \phi(n)\hat{I}$ given by (20). When $\phi(\hat{n})$ is known, $\phi(n)$ can be obtained by $\phi(n) = \phi(\hat{n})\hat{I}^{-1}$.

To obtain the public encryption isockey, \hat{e} , we can randomly select the integer e and calculate $\hat{e} = e\hat{I}$ such that the following greatest common isodivisor (*gcid*) is verified, that is $gcid(\phi(\hat{n}), \hat{e}) = \hat{I}$, and $\hat{I} < \hat{e} < \phi(\hat{n})$. The $gcid(\phi(\hat{n}), \hat{e})$ can be calculated using the Euclidean-Senegalese algorithm summarized in Table I. Or, \hat{e} can simply be computed as $\hat{e} = e\hat{I}$, where e is the RSA public key.

On one hand, the encryption isokeys of the Iso-RSA cryptographic scheme is given by the triplet $\{\hat{n}, \hat{e}, \hat{I}\}$, where pair of isointegers $\{\hat{n}, \hat{e}\}$ is made public, while the isounit \hat{I} is the secret encryption and decryption isockey. On the other hand, the encryption isokeys of the Iso-RSA cryptographic scheme can be computed using the triplet $\{n, e, \hat{I}\}$, where n and e are the previous RSA public encryption keys. The values of \hat{n} and \hat{e} can be computed as $\hat{n} = n\hat{I}$ and $\hat{e} = e\hat{I}$, where again the isounit \hat{I} is the secret encryption and decryption isockey. Now, we can compute the private decryption isockey \hat{f} , which is the isomultiplicative inverse of $\hat{e} \pmod{\phi(\hat{n})}$. We can write

$$\hat{e} * \hat{f} \equiv \hat{I} \pmod{\phi(\hat{n})} \quad (25)$$

Equation (25) implies that

$$\hat{f} \equiv \hat{e}^{-1} \pmod{\phi(\hat{n})} \quad (26)$$

Three different methods can be used to calculate \hat{f} .

Method 1: The Extended Euclidean-Senegalese Algorithm. The Extended Euclidean-Senegalese Algorithm summarized in Table II is a systematic way that can be used to calculate \hat{f} given by the relation described in (25) or (26).

Method 2: Trial and Error

When the isonumbers are large, it is not the best method to compute \hat{f} because it involves a search of an isointeger \hat{q} satisfying (28). Equation (25) implies

$$(\hat{e} * \hat{f}) \pmod{\phi(\hat{n})} = \hat{I} \pmod{\phi(\hat{n})} = \hat{I} \quad (27)$$

Equation (27) implies that there exist an isointeger \hat{q} such that

$$\hat{e} * \hat{f} = \hat{q} * \phi(\hat{n}) + \hat{I} \quad \text{or} \quad \hat{e}\hat{T}\hat{f} = \hat{q}\hat{T}\phi(\hat{n}) + \hat{I} \quad (28)$$

Using (28), the goal is to use trial and error or search for the value of \hat{q} that will make \hat{f} to be an isointeger given by

$$\hat{f} = (\hat{q}\hat{T}\phi(\hat{n}) + \hat{I})\hat{T}\hat{e}^{-1} = \frac{\hat{q}\hat{I}^{-1}\phi(\hat{n}) + \hat{I}}{e}, \quad (29)$$

where $\hat{e}^{-1} = \frac{\hat{I}}{e}$ and $e = \hat{e}\hat{I}^{-1}$. Note that $\hat{T}\hat{I} = \hat{T}\hat{T}^{-1} = 1$.

Method 3: Direct method

Assuming that the value f describe in the traditional RSA Cryptographic system is available, \hat{f} is computed as

$$\hat{f} = f\hat{I} \quad (30)$$

The isointegers needed for decryption are $\{\hat{n}, \hat{f}, \hat{I}\}$, where \hat{f} is a private decryption isockey, and \hat{I} a secret encryption and decryption isockey, and \hat{n} is the public isokeys previously used for encryption.

For the encryption to be computationally secure, the values of \hat{p} , \hat{s} , \hat{e} , and \hat{f} must be as large as possible.

2) **The Proposed Iso-RSA Encryption Function:** Similar to the RSA encryption function used to encrypt a message or plaintext represented by an integer M between 0 and $n - 1$, the Iso-RSA encryption function is used to encrypt a message M with corresponding isointeger $\hat{M} = M\hat{I}$ between $\hat{0}$ and $\widehat{n-1}$. Using the RSA public key e , or given the public encryption isockey \hat{e} , one can compute the the public key $e = \hat{e}\hat{I}^{-1}$. So, assuming the secrete encryption and decryption isockey (isounit) \hat{I} is known to both sender and receiver, the encryption of \hat{M} using the encryption function, E , to obtain the ciphertext \hat{C} is given by

$$\hat{C} = E(\hat{M}) = E(M\hat{I}) = (M^e)\hat{I} \pmod{n\hat{I}} = (M^e)\hat{I} \pmod{\hat{n}} \quad (31)$$

Note that \hat{C} can also be computed as $\hat{C} = E(\hat{M}) = E(M)\hat{I}$, where $E(M) = C = M^e \pmod{n}$ is given by Eq. (21) and corresponds to the RSA encryption function using the public key $\{n, f\}$. So, the sender can encrypt the message M using either (31) or $\hat{C} = E(\hat{M}) = E(M)\hat{I}$ since it has access to the shared secrete isockey \hat{I} . If \hat{I} is a different randomly generated positive integer during each encryption operation, the encryption function $E(\hat{M})$ contains a random variable; therefore, it is probabilistic compared to the RSA encryption function described in (21). This will ensure that the encryption of the same message M using the same isockey $\{\hat{e}, \hat{n}\}$ with a different random isounit or secret isockey \hat{I} , will produce different results and make the associated ciphertext more secure.

3) **The Proposed Iso-RSA Decryption Function:** The Iso-RSA decryption function is used to decrypt the ciphertext represented by the isointeger \hat{C} between $\hat{0}$ and $\widehat{n-1}$. One can decrypt the ciphertext \hat{C} using the decryption function, D , in conjunction with the private isockey \hat{f} , to obtain the original message \hat{M} . Using the private isockey \hat{f} , we can compute the private key $f = \hat{f}\hat{I}^{-1}$, or f can be directly computed from the RSA algorithm, and we calculate the cipher $C = \hat{C}\hat{I}^{-1}$. The message M is given by

$$\hat{M} = D(\hat{C}) = D(C\hat{I}) = (C^f)\hat{I} \pmod{n\hat{I}} = (C^f)\hat{I} \pmod{\hat{n}} \quad (32)$$

The original message M can be obtained from (32) as

$$M = \hat{M}\hat{I}^{-1} = \frac{\hat{M}}{\hat{I}} \quad (33)$$

TABLE II: The Extended Euclidean-Senegalese Algorithm

Assume that $\hat{a} \geq \hat{b} > \hat{0}$				
Compute Remainder (\hat{r}_i)	Compute Quotient (\hat{q}_i)	Compute (\hat{x}_i)	Compute (\hat{y}_i)	Which Fulfills
$\hat{r}_{-1} = \hat{a}$	—	$\hat{x}_{-1} = \hat{I}$	$\hat{y}_{-1} = \hat{0}$	$\hat{r}_{-1} = \hat{a} * \hat{x}_{-1} + \hat{b} * \hat{y}_{-1} = \hat{a}$ $\hat{r}_{-1} = \hat{a}\hat{I}^{-1}\hat{x}_{-1} + \hat{b}\hat{I}^{-1}\hat{y}_{-1} = \hat{a}$
$\hat{r}_0 = \hat{b}$	—	$\hat{x}_0 = \hat{0}$	$\hat{y}_0 = \hat{I}$	$\hat{r}_0 = \hat{a} * \hat{x}_0 + \hat{b} * \hat{y}_0 = \hat{b}$ $\hat{r}_0 = \hat{a}\hat{I}^{-1}\hat{x}_0 + \hat{b}\hat{I}^{-1}\hat{y}_0 = \hat{b}$
$\hat{r}_1 = \hat{a} \text{ mod } \hat{b}$ Or $\hat{a} = \hat{q}_1 * \hat{b} + \hat{r}_1$ $\hat{a} = \lfloor \frac{\hat{a}}{\hat{b}} \rfloor \hat{b} + \hat{r}_1$ $\hat{0} < \hat{r}_1 < \hat{b}$	$\hat{q}_1 = \lfloor \frac{\hat{a}}{\hat{b}} \rfloor \hat{I}$	$\hat{x}_1 = \hat{x}_{-1} - \hat{I}^{-1}\hat{q}_1\hat{x}_0$ $\hat{x}_1 = \hat{x}_{-1} - \hat{0}$ $\hat{x}_1 = \hat{I}$	$\hat{y}_1 = \hat{y}_{-1} - \hat{I}^{-1}\hat{q}_1\hat{y}_0$ $\hat{y}_1 = \hat{0} - \hat{I}^{-1}\hat{q}_1\hat{I}$ $\hat{y}_1 = -\hat{q}_1$	$\hat{r}_1 = \hat{a} * \hat{x}_1 + \hat{b} * \hat{y}_1$ $\hat{r}_1 = \hat{a}\hat{I}^{-1}\hat{x}_1 + \hat{b}\hat{I}^{-1}\hat{y}_1$
$\hat{r}_2 = \hat{b} \text{ mod } \hat{r}_1$ Or $\hat{b} = \hat{q}_2 * \hat{r}_1 + \hat{r}_2$ $= \lfloor \frac{\hat{b}}{\hat{r}_1} \rfloor \hat{r}_1 + \hat{r}_2;$ $\hat{0} < \hat{r}_2 < \hat{r}_1$	$\hat{q}_2 = \lfloor \frac{\hat{b}}{\hat{r}_1} \rfloor \hat{I}$	$\hat{x}_2 = \hat{x}_0 - \hat{I}^{-1}\hat{q}_2\hat{x}_1$ $\hat{x}_2 = \hat{0} - \hat{I}^{-1}\hat{q}_2\hat{I}$ $\hat{x}_2 = -\hat{q}_2$	$\hat{y}_2 = \hat{y}_0 - \hat{I}^{-1}\hat{q}_2\hat{y}_1$ $\hat{y}_2 = \hat{I} - \hat{I}^{-1}\hat{q}_2(-\hat{q}_1)$ $\hat{y}_2 = \hat{I} + \hat{I}^{-1}\hat{q}_2\hat{q}_1$	$\hat{r}_2 = \hat{a} * \hat{x}_2 + \hat{b} * \hat{y}_2$ $\hat{r}_2 = \hat{a}\hat{I}^{-1}\hat{x}_2 + \hat{b}\hat{I}^{-1}\hat{y}_2$
$\hat{r}_3 = \hat{r}_1 \text{ mod } \hat{r}_2$ Or $\hat{r}_1 = \hat{q}_3 * \hat{r}_2 + \hat{r}_3$ $= \lfloor \frac{\hat{r}_1}{\hat{r}_2} \rfloor \hat{r}_2 + \hat{r}_3;$ $\hat{0} < \hat{r}_3 < \hat{r}_2$	$\hat{q}_3 = \lfloor \frac{\hat{r}_1}{\hat{r}_2} \rfloor \hat{I}$	$\hat{x}_3 = \hat{x}_1 - \hat{I}^{-1}\hat{q}_3\hat{x}_2$ $\hat{x}_3 = \hat{I} - \hat{I}^{-1}\hat{q}_3(-\hat{q}_2)$ $\hat{x}_3 = \hat{I} + \hat{I}^{-1}\hat{q}_3\hat{q}_2$	$\hat{y}_3 = \hat{y}_1 - \hat{I}^{-1}\hat{q}_3\hat{y}_2$ $\hat{y}_3 = -\hat{q}_1 - \hat{I}^{-1}\hat{q}_3\hat{y}_2$	$\hat{r}_3 = \hat{a} * \hat{x}_3 + \hat{b} * \hat{y}_3$ $\hat{r}_3 = \hat{a}\hat{I}^{-1}\hat{x}_3 + \hat{b}\hat{I}^{-1}\hat{y}_3$
\vdots	\vdots	\vdots	\vdots	\vdots
$\hat{r}_{n-1} = \hat{r}_{n-3} \text{ mod } \hat{r}_{n-2}$ Or $\hat{r}_{n-3} = \hat{q}_{n-1} * \hat{r}_{n-2} + \hat{r}_{n-1}$ $= \lfloor \frac{\hat{r}_{n-3}}{\hat{r}_{n-2}} \rfloor \hat{r}_{n-2} + \hat{r}_{n-1}$ $\hat{0} < \hat{r}_{n-1} < \hat{r}_{n-2}$	$\hat{q}_{n-1} = \lfloor \frac{\hat{r}_{n-3}}{\hat{r}_{n-2}} \rfloor \hat{I}$	$\hat{x}_{n-1} = \hat{x}_{n-3} - \hat{I}^{-1}\hat{q}_{n-1}\hat{x}_{n-2}$	$\hat{y}_{n-1} = \hat{y}_{n-3} - \hat{I}^{-1}\hat{q}_{n-1}\hat{y}_{n-2}$	$\hat{r}_{n-1} = \hat{a} * \hat{x}_{n-1} + \hat{b} * \hat{y}_{n-1}$
$\hat{r}_n = \hat{r}_{n-2} \text{ mod } \hat{r}_{n-1}$ Or $\hat{r}_{n-2} = \hat{q}_n * \hat{r}_{n-1} + \hat{r}_n$ $= \lfloor \frac{\hat{r}_{n-2}}{\hat{r}_{n-1}} \rfloor \hat{r}_{n-1} + \hat{r}_n$ $\hat{0} < \hat{r}_n < \hat{r}_{n-1}$	$\hat{q}_n = \lfloor \frac{\hat{r}_{n-2}}{\hat{r}_{n-1}} \rfloor \hat{I}$	$\hat{x}_n = \hat{x}_{n-2} - \hat{I}^{-1}\hat{q}_n\hat{x}_{n-1}$	$\hat{y}_n = \hat{y}_{n-2} - \hat{I}^{-1}\hat{q}_n\hat{y}_{n-1}$	$\hat{r}_n = \hat{a} * \hat{x}_n + \hat{b} * \hat{y}_n$
$\hat{r}_{n+1} = \hat{r}_{n-1} \text{ mod } \hat{r}_n$ $= \hat{0}$ Or $\hat{r}_{n-1} = \hat{q}_{n+1} * \hat{r}_n + \hat{0}$ $= \lfloor \frac{\hat{r}_{n-1}}{\hat{r}_n} \rfloor \hat{r}_n + \hat{0}$	$\hat{q}_{n+1} = \lfloor \frac{\hat{r}_{n-1}}{\hat{r}_n} \rfloor \hat{I}$	—	—	$\hat{d} = \text{gcd}(\hat{a}, \hat{b}) = \hat{r}_n$ $= \hat{a} * \hat{x} + \hat{b} * \hat{y}$ $\hat{x} = \hat{x}_n \text{ and } \hat{y} = \hat{y}_n$ If $\hat{d} = \hat{I}$, then $\hat{b}^{-1} \text{ mod } \hat{a} = \hat{y}_n = \hat{y}$

TABLE III: The Extended Euclidean-Senegalese Algorithm Example

Given: $a = 20; b = 7; I = 30$				
Compute Remainder (r_i)	Compute Quotient (q_i)	Compute (x_i)	Compute (y_i)	Which Fulfills
$r_{-1} = a = 20$	-	$x_{-1} = I = 30$	$y_{-1} = 0$	$r_{-1} = aI^{-1}x_{-1} + bI^{-1}y_{-1} = a$ $r_{-1} = 20I^{-1}I + 7I^{-1}0 = 20$
$r_0 = b = 7$	-	$x_0 = 0$	$y_0 = I$	$r_0 = aI^{-1}x_0 + bI^{-1}y_0 = b$ $r_0 = 20I^{-1}0 + 7I^{-1}I = 7$
$r_1 = a \bmod b$ $r_1 = 20 \bmod 7$ $r_1 = 6$	$q_1 = \lfloor \frac{a}{b} \rfloor I$ $q_1 = \lfloor \frac{20}{7} \rfloor (30)$ $q_1 = 2$	$x_1 = I = 30$	$y_1 = -q_1 = -2$	$r_1 = aI^{-1}x_1 + bI^{-1}y_1$ $r_1 = 20I^{-1}I + 7I^{-1}(-2)$ $r_1 = 20 - 14 = 6$
$r_2 = b \bmod r_1$ $r_2 = 7 \bmod 6$ $r_2 = 1$	$q_2 = \lfloor \frac{b}{r_1} \rfloor I$ $q_2 = \lfloor \frac{7}{6} \rfloor (30)$ $q_2 = 30 = I$	$x_2 = -q_2$ $x_2 = -I$	$y_2 = I + I^{-1}q_1$ $y_2 = I + I^{-1}I(2)$ $y_2 = I + 2$ $y_2 = 3$	$r_2 = aI^{-1}x_2 + bI^{-1}y_2$ $r_2 = 20I^{-1}(-I) + 7I^{-1}(3)$ $r_2 = -20 + 21$ $r_2 = I$
$r_3 = r_1 \bmod r_2$ $r_3 = 6 \bmod 1$ $r_3 = 0$ STOP (We Got a 0 Remainder)	-	-	-	$n = 2$ $d = \text{gcd}(a, b) = a * x + b * y = r_n$ $\text{gcd}(20, 7) = 20 * x + 7 * y = r_3 = I$ $x = x_2 = -I$ and $y = y_2 = 3$ Since $d = I$, we have $b^{-1} \bmod a = 7^{-1} \bmod 20 = y_2 = 3$

There is another way the receiver who already has access the public isokeys $\{\hat{n}, \hat{e}\}$, the private isokeys $\{\hat{n}, \hat{f}\}$, and the secrete isokey \hat{I} , can recover the original message or plaintext M from the ciphertext \hat{C} . Using the received Iso-RSA ciphertext \hat{C} , the receiver can compute the RSA ciphertext C as $C = \hat{C}\hat{I}^{-1}$, and then compute M using the RSA decryption function as $M = D(C) = C^f \pmod n$ given by (22), where $f = \hat{f}\hat{I}^{-1}$, $n = \hat{n}\hat{I}^{-1}$ or $n = p \times s$.

V. APPLICATIONS OF THE ISO-RSA CRYPTOGRAPHIC SCHEME

This section shows some possible applications and implementations of the Iso-RSA Cryptographic scheme. The results from this section are obtained using simulation programs that we have written using a MATHEMATICA 12 software package. Examples 1 and 2 are for illustration purposes with smaller values of the encryption and decryption isokeys. These examples give insight through simple computations using the Iso-RSA cryptographic scheme. In example 3, the encryption and decryption isokeys have extremely large number of digits and therefore making the corresponding ciphertext computationally secure.

The Iso-RSA Cryptographic scheme can be used to encrypt any data or information that are numbers or bits, or that can be encoded/mapped into numbers or bits. These two major cases are described in the next two subsections.

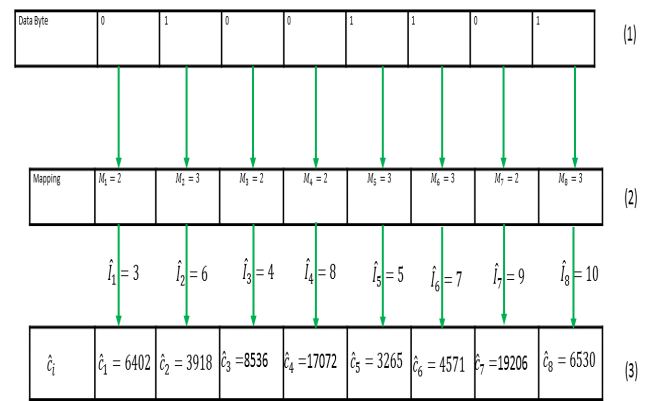
A. Binary Encryption/Decryption Case Applications

The Iso-RSA Cryptographic scheme can be used to encrypt a binary digit or bit represented by 0 or 1. Recall that the encryption of the message M and its corresponding isonumber $\hat{M} = M\hat{I}$ to produce the ciphertext \hat{C} is given by (31) as $\hat{C} = E(\hat{M}) = E(M\hat{I}) = (M^e)\hat{I} \pmod{\hat{n}}$.

On one hand, we can obviously observe that if message M takes the value of bit 1 or $M = 1$, the corresponding ciphertext

$\hat{C} = (1^e)\hat{I} \pmod{\hat{n}} = \hat{I} \pmod{\hat{n}} = \hat{I} \pmod{n\hat{I}} = \hat{I}$. To avoid having the same results when encrypting bit 1, we can change the value of \hat{I} when encrypting each bit value of 1. Or/and, M can be mapped to a positive integer before encryption, for example we can set $M = 3$. So, it is not recommended to map M to integer 1 because it will lead to having $\hat{C} = \hat{I}$ which will be easy to infer by a hacker.

On the other hand, when M takes the value of bit 0 or $M = 0$, the corresponding ciphertext $\hat{C} = (0^e)\hat{I} \pmod{\hat{n}} = 0 \pmod{\hat{n}} = 0 = \hat{0}$. To avoid having the same result of 0, when encrypting bit 0, before encryption each bit zero is map to a different value such as 2 for example, and/or we can also change the value of \hat{I} for each bit value of 0 encryption as shown in next example.



(1) Original Data Byte
(2) Data Encoding, bit 0 mapped to 2, bit 1 mapped to 3
(3) Encrypted data using corresponding isounit \hat{I}_i

(a)

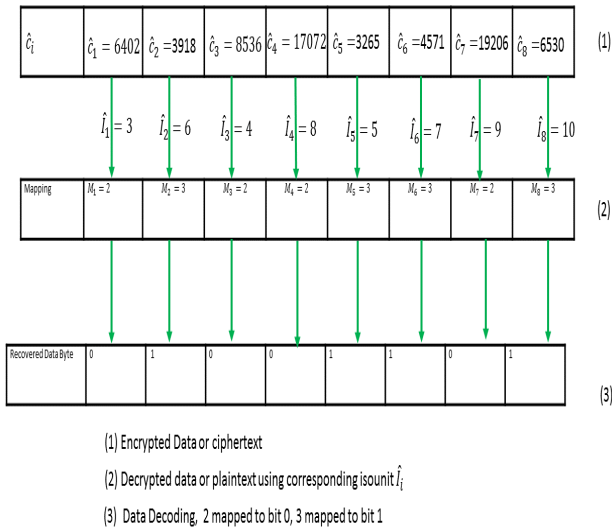
Fig. 1: Data byte Encryption using the RSA-Senegalese Cryptographic Scheme

Example 2: Binary Encryption/Decryption Case

Encrypt the data byte 01001101 using the Iso-RSA cryptographic scheme, then decrypt the corresponding ciphertext to obtain the original data byte.

We randomly generate our private keys (prime numbers) $p = 59$ and $s = 41$ and compute the associated isokeys \hat{p} and \hat{s} , \hat{n} , and then select the public key $e = 181$. Also, each corresponding isounit \hat{I} is randomly generated. Before encryption, each bit 0 is mapped/encoded to integer 2, while each bit 1 is mapped to integer 3.

During the encryption step shown in Fig. 1, the first bit of 0 is mapped to integer 2 then encrypted. Using the isounit $\hat{I}_1 = 3$, we compute the public isokey $\hat{n} = (ps)\hat{I}_1 = 7257$, and use



(a)

Fig. 2: Data byte Encryption using the Iso-RSA Cryptographic Scheme.

the public isokey $\hat{e} = e\hat{I}_1 = 543$ to encrypt the message $M = 2$ and obtain the ciphertext $\hat{C}_1 = 6402$.

During the decryption step, we compute the private isokey $\hat{f} = 423$ and use the isounit \hat{I}_1 to decrypt the ciphertext \hat{C}_1 and obtain the message $\hat{M} = 6$ and compute $M = 2$, which is then mapped back to bit 0. Other bits are encrypted and decrypted in a similar way. Fig 1. summarizes the encryption results for all bits, while Fig. 2 provides the decryption results.

If there are more bytes to be encrypted, instead of choosing a different set of isounits from I_1 through I_8 , we can just rotate left or rotate right by a number of units $n > 1$ only known to legitimate users. This 8-bit binary encryption and decryption process can be extended to data with 16-bit, 32-bit, 64-bit length, or data with a higher bit length.

B. Integer Encryption/Decryption Case

The Iso-RSA cryptographic scheme can be also used to encrypt any data or information represented by a positive integer. For illustration purposes, we can revisit example 1, but this time we use encryption and decryption isokeys with large number of digits making the associated ciphertext highly secure.

Example 3: Encrypt the message $M = 5$ and decrypt the corresponding ciphertext using the Iso-RSA cryptographic scheme.

Isokeys Generation Phase:

We randomly generate a 1000 digits integer for the isounit which is the secret encryption and decryption isokey given by $\hat{I} = 6705642340830151337605366935497966688761425655519559843828657577090172913994408414332899431464255251388681687534843573142127832857164988057162140679344430849550623415758144002805955676539779144413081666427672244649977065294448511646168508466258719140780943313500356137350913117761841587541072046842140628578142955389931592357685235132792156575417572258336611013384887932596891993802189171620277603372515155995685792629671939427207687000922309664920659259876090557494905671340147267632922771751460911880985048796749243916894295401320178931291607476629090848825487365429867396914733651503131872003548735801648508726573810037409657943941822983944431554434714883289287312299035367249105842302709469538697906678497506604733618656495369998344420895459600227346026009124090746307330643089753441656918158490417608926105825031300108627198097128996366457145879916542747941929967421611549556420177407112578351666448385326256147185085435692846830946101537995091381866029559339800844475788961790423761031689997212.$

We also randomly select the private prime numbers p and s with 600 digits each, and then compute $\hat{p} = p\hat{I}$ and $\hat{s} = s\hat{I}$ with each 1600 digits. We calculate the public isokey \hat{n} using (23) as $\hat{n} = \hat{p} * \hat{s} = (ps)\hat{I}$ and obtain an integer with 2200 digits and 7306 bits given by

$\hat{n} = 1109216243454912743255660804559755004992834351052549014434500755060694914920200996577882430565433539641869697102491569052297674823868086998199560135651691594957298370714132063739061821791803207988321495979594408956545176572589015144119810493695404488904405680564916683333596928393294898786558701309102221401712780094412140704909038502539457683582022798068290980535955553785180716488393563018578881751094343338891584052410147809531282088123538394726280171016082963794320734360516317648813962962780800934399476222188432466539146061878670708207983647501856537882288556013464631705390717255749584465337016401836697427890195826459762916456895516754030118458581092636520574931526905190381332479492206946367964630058601133769052759298783996249265509806842618411084857843882914028275819081096379797089087208855906639572410927479993381334712064914730146796662143649287774856712317316238345712123544747990816502245187292759305394322350447786369389086618768325272946484626068871128067323130629977878745318556925894051661536006947697179977737828205566141809076529221828083227500120615169946942131168434714020481311823501180969554575520869268523087616789521121148550329632524419438461296544170303432877039617079928717113711859746417811$

4130765056386157791275293246836706380154642224799
 3209964722641078494333670644051118234727239791793
 1948488586670306356265251696084685844644820366670
 0005774375160728871810850929614687553988869223262
 6914051289840140189035369276217721325072239275226
 5665847459044268864072109667233826301859759393745
 2288647349894315366235693844460740073457780039824
 6780562998449270151792590992703254010962659596279
 3237431244275777138353128642855883461368595190103
 9423592261104832674770219447163400425459575748943
 7856698796970273028082171227710458332232984566509
 9813612497748485487371331971816175639710675562106
 9630074224665345684199530889831097259935263286063
 6288050915841593567848035452588690967065796690001
 4440073248411420140050887363650136284817018960498
 7925529432507006108513851046920049661728896342367
 6874275159033432197545543754241923889943917914518
 3675807666011624988202060746498659127644263280693
 6257180960762402144539278294226413114320975756422
 69733765027688070276284472244369655207825705236.

Next, after computing $\phi(n)$ and $\phi(\hat{n})$, we randomly select the integer e and calculate \hat{e} such that $gcd(\phi(\hat{n}), \hat{e}) = \hat{I}$, and $\hat{I} < \hat{e} < \phi(\hat{n})$. The public isokey \hat{e} has 2199 digits and given by

$\hat{e} = 1534522875345678222432311017089006757547524615$
 73592509899379309647028015045341188530529340118911
 59334099947005942706023359266103289504694773931271
 89886536010475951886427460261244664624763366282808
 96065936847779752483529139578994705158522749346091
 48937482312993090818589168807557661122506923369098
 25128762612346719902587934210666183871847105242413
 64343494794524508151029158506293982168254317173397
 39840904998031518400935701248441882843758364361802
 15274307819994769229642602591759469612862000091975
 00687942798382269949536422937557261528004131424765
 5052556228079827458681501143722277221545330814163
 21112041644949048001984578362124522448503324621421
 15853290239365615575022166838924970685650302585059
 58594486872773215203913943518010364281589064124774
 63479355055541614807969492382605486055214825799097
 47956988668067817687748357398389823592285418196681
 42362880682491089746961905686586508442292367226484
 92106687205249203208405188503712813492463963711264
 53210068800373540848716064318902522702019674082157
 33745375923544511297335561165245905123906313918128
 06635962877404509287843557883679487983659131441691
 67903965445461748195361194260757274616298678928437
 60236337350619957714628744050546169997024014752082
 07728100583270944601433321559696212334033942739536
 57050724322081181100011975969915330347059997701811
 43876564098748594207555245819338268158652193910410
 60741204520003797068760611319832566798066253332293

33976561830869875089619520091709712399906259549667
 89929433185889785199965271896531754209952068297021
 79344414417568668882299550834292094620156636822946
 24506928646203025942504525417485610161593988081813
 64873354415459214501062380748439156797708633617439
 22803753406882973333382087539907906464587582022719
 48731003625202939892510334649450748889876378004723
 21638218863901781904051996523257715588720767000466
 37706664829799494662162276009299042842034953439095
 44722607894082394643331471542582532223621681481894
 32420161197202609619845300012127427043313236763489
 38646126902817389483074396707432829712961367201695
 72197358375185220237679291558274956923297046248072
 97207155112503107620169202131221743916422417495218
 42892619824005634431793613151627531279266985726673
 51255033570172760614170325852914036879720915307373
 092.

So, we now have the public encryption isokey $PU = \{\hat{n}, \hat{e}\}$ and the secret encryption isokey (isounit) $PR = \{\hat{I}\}$.

We calculate the private decryption isokey \hat{f} which has 2199 digits but not shown here because of number of pages limitations.

So, the private decryption isokey is $PR = \{\hat{f}, \hat{I}\}$

Encryption Phase:

The message $M = 5$ is encrypted using (31) to obtain the ciphertext \hat{C} with 2199 digits given by

$\hat{C} = 383331694654418271994382897533298668520133104$
 5072557055016654109772284531409051214424330802144
 159913559972098620323464670930140205048959104556
 265672990909100103260391516732566988794204394085
 671241427027797904441707755208441944569562391279
 177426417096087370983318073574116063547033358059
 854973224713999850379408868450926389654555670760
 643574282904992964719730791552937812322942710431
 817627776641494493288487214756278025936012065011
 234715824271913496834543275290132039119184442148
 374669472309796339021850748944034050634954509637
 269074920900005204577123438005741767030333935264
 231210472914337294890627542132103948311111363438
 977706487333810663110599531455382544712479347444
 330188444219105762431583399997902298931033891559
 335670998611791364575107354437799687897940595907
 086965219914005797502587923430503639841021567325
 662724506263681996144025269289970196795945132449
 287181611599110690456926208462097986285663473298
 992134079043615790447810581204467260556669690619
 005939822469218633774910540547535859062555410285
 298298041404478539240351121280982075935329878116
 713270313213702746079294465367470878949261999557
 756573680073416625057574997535575534634296753250
 516146445916483278035358720095513800759836950081
 537754568590449025230719387781191265209211704500

14666998171228099279298501460188022644501241571959
 57953728555482280663789579428811017246918294965516
 71109410257847349636187127416176712974867105707396
 33350944144908513962635265458798152067106168273352
 68609253191011505502160217664891992824042505679695
 47789971518233278372637341355749815404264089209856
 61594218720558698137706969083758980831332481301428
 31504548564007535192655673689702240320178596887129
 06035378634244324004330179963866746375186659009072
 23480150066032538225043057487064731734997851977030
 13620638105102108677194159116195042139829297694941
 74084923203657109183817466285084100760023116181264
 42148279574453025664362132165634941142568277415825
 82896285850263398471128924550133305720402462402666
 92553486235092818570927322897597535993139706690155
 04693770815926811163740351440189945060620285509354
 34860295855202795031402575317066030771679454785261
 51884620254064086293854359280057376781077304717598
 91150723591049343063009468056360599117262538440019
 580.

This value of \hat{C} represents the secure encrypted value that is transmitter or/and store.

It only takes 0.0547881 seconds to encrypt the message M using a regular desktop computer.

Decryption Phase:

The above ciphertext \hat{C} is decrypted using (32) to obtain \hat{M} with 1001 digits.

The corresponding message M is computed using (33), and is given by

$$M = 5.$$

It only takes 0.040612 seconds to decrypt the ciphertext \hat{C} using the same regular desktop computer.

As can be seen in the final result, the original message $M = 5$ is obtained after decryption.

The Iso-RSA cryptographic scheme produces highly secure encrypted information because of its isokeys sizes that can be extremely large, and other characteristics.

The number of digits for \hat{p} , \hat{s} , \hat{n} , \hat{e} , and \hat{I} is arbitrary, but the higher the number of digits for each, the more computationally secure the public isokeys and ciphertexts become, but also the computational cost is higher.

Also, note that the ciphertext \hat{C} produced by the Iso-RSA encryption scheme has a very large number of digits, and therefore will require more memory for storage and a longer transmission time. This large size of \hat{C} can be overcome with extra processing that will be discussed in our next paper.

VI. CONCLUSIONS

This paper introduces the Iso-RSA cryptographic scheme which possesses characteristics of symmetric and asymmetric cryptographic approaches, and therefore can be seen as a

hybrid private-key and public-key cryptosystem. It consists of a key generation step, an encryption step, and a decryption step. An isounit represented by a positive integer replacing the unit one of normal arithmetic is used as both secret encryption and decryption isokey, and it provides an additional degree of freedom of unit. Because the encryption and decryption isokeys can be extremely large, for example, more than 3000 digits or about 10,000 bits or more, the Iso-RSA cryptographic produces highly secure encrypted information. The paper also presents examples of possible applications of the Iso-RSA cryptosystem. The scheme can be used to encrypt and decrypt any information that can be encoded into positive integers or bits such as a binary 0 or 1. The Iso-RSA cryptographic scheme can also be used in a wide range of other applications in security mechanisms and security services which include data confidentiality, authentications, data integrity, digital signature, key exchange, and others. Our contribution includes the strengthening of the RSA cryptosystem using an isounit as a secret encryption and decryption isokey, leading to a hybrid public-key private-key cryptosystem that produces highly secure encrypted information. Our future work includes investigating how well the Iso-RSA cryptographic scheme can resist quantum computing-based attacks.

REFERENCES

- [1] C. X. Jiang, *Foundations of Santilli's Isonumber Theory*, Fundamental Open Problems in Science at the End of the Millennium, Proceeding of the Beijing Workshop, August 1997, Hadronic Press, Palm Harbor, FL 34682-1577, U.S.A, ISBN 1-57485-029-6, PP. 105-139.
- [2] M. I. Wade, H. C. Ogworonjo, M. Gul, M. Ndoye, M. Chouikha, W. Paterson *Red Green Blue Image Encryption Based on Paillier Cryptographic Cryptographic Approach*, World Academy of Science, Engineering and Technology, International Journal of Electronics and Communication Engineering Vol:11, No:12, 2017, <https://waset.org/abstracts/79232>
- [3] M. I. Wade, *Distributed Image Encryption Based On a Homomorphic Cryptographic Approach*, 10th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, Columbia University, New York City, USA 10- 12 October 2019 (IEEE UEMCON 2019).
- [4] M. I. Wade, T. Gill, "Isonumbers and RGB Image Encryption," Hadronic Journal, Algebras, Groups and Geometries, Vol. 37, 103-119 (2021.)
- [5] P. Paillier, *Public-Key Cryptosystems Based on Composite Degree Residuosity Classes*, J. Stern (Ed.): EUROCRYPT'99, LNCS 1592, pp. 223-238, 1999. @ Springer-Verlag Berlin Heidelberg 1999
- [6] R. M.Santilli-1, *Foundations of Theoretical Mechanics II*, Springer-Verlag New York, U.S.A, (1978), ISBN 0-387-08874-1.
- [7] R. M.Santilli-2, *Isonumbers and Genonumbers of Dimensions 1, 2, 4, 8, their Isoduals and Pseudoduals, and "Hidden Numbers," of Dimension 3, 5, 6, 7*, Algebras, Groups and Geometries, Vol. 10, 273 (1993).
- [8] R. Ranasinghe and P. Athukorala, *A Generalization of the ElGamal public-key cryptosystem*, Journal of Discrete Mathematical Sciences and Cryptography, DOI: 10.1080/09720529.2020.1857902
- [9] R. L. Rivest, A. Shamir, L. Adleman, *A Method for Obtaining Digital Signatures and Public-Key Cryptosystems*, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MD 02139 E-mail address rivest@theory.lcs.mit.edu
- [10] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 2003 Pearson Education, Inc.,
- [11] Yi. Xun, P. Russell, and B. Elisa, *Homomorphic Encryption and Applications*, 2014 XII, 126 p. 23 illus., <http://www.springer.com/978-3-319-12228-1>

The Iso-ElGamal Cryptographic Scheme

1st Mamadou I. Wade

Department of Electrical Engineering and Computer Science Department of Electrical Engineering and Computer Science
Howard University

Washington, DC, U.S.A

wademamadoui@gmail.com or mamadou.wade@howard.edu

2nd Tepper Gill

Department of Electrical Engineering and Computer Science
Howard University

Washington, DC, U.S.A

tgill@howard.edu

Abstract—In this paper, we use Santilli's isonumber theory of the first kind to add a new degree of freedom of unit to the ElGamal cryptographic system leading to a new encryption scheme. In this scheme, the key generation, encryption, and decryption steps are performed in an isofield with an isounit replacing unit one of normal arithmetic. This isounit is used as a shared secret isokey for encryption and decryption as is the case in a symmetric cryptographic system. In addition, our proposed cryptographic system is asymmetric with a public isokey used for encryption and a private isokey used for decryption. We thus obtain a hybrid public-key private-key cryptographic system with a substantial increase in security due to this additional freedom of unit. The private isokeys and ciphertexts produced by this scheme are highly secure. In addition, the ability to use isokeys of ten thousand bits or more with minimum additional computational cost makes this cryptographic scheme produce highly secure encrypted information. Also, applications and implementations of the Iso-ElGamal scheme on messages represented by positive integers and data or information encoded into binary digits or bits are presented. A wide range of other applications in security services and security mechanisms, which include authentication, data confidentiality, data integrity, digital signature, and others can be implemented using our proposed scheme.

Index Terms—Iso-ElGamal, Iso-ElGamal Cryptographic Scheme, ElGamal Cryptographic System; Isonumber Cryptosystems.

I. INTRODUCTION

A Wide range of public-key encryption schemes has been proposed in the past few decades. Using a public-key cryptographic system, one key can be made public for encryption, while a different key is made private for decryption, eliminating the key distribution problem encountered in symmetric encryption schemes, where the same shared secret key is used for both encryption and decryption. One widely used public-key encryption scheme is the ElGamal cryptographic system [9], [11] which was proposed by Taher ElGamal in 1984. This approach is based on the Diffie-Hellman key exchange protocol in which security

relies on the difficulty of solving the Discrete Logarithm (DL) problem over a finite field. Another widely used public-key cryptographic system is the RSA [10], [11] first introduced by Ron Rivest, Adi Shamir, and Len Adleman in 1978. The RSA is a block cipher where both the ciphertext and plaintext consist of integers. The Paillier cryptographic system [1], [5], [6], [11] is another highly favored public-key encryption scheme.

In this paper, we use Santilli's Isonumber Theory of the first kind to insert an additional degree of freedom of unit into the encryption process for the ElGamal system, which we call the Iso-ElGamal scheme. In this approach, the unit 1 of normal arithmetic is replaced by a new unit called an isounit. When the isounit is used as both encryption and decryption keys (isokeys), in addition to the public and private encryption and decryption isokeys, our proposed scheme becomes a hybrid public-private-key cryptographic system, providing the strengths of both asymmetric and symmetric schemes [3], [4]. Large isokey sizes of 3,000 digits (about 10,000 bits) or more can be used with minimal additional computational cost. This scheme can also be used for authentications, confidentiality, and data integrity.

The paper is summarized as follows: In Section II, we introduce Santilli's isonumber theory of the first kind. In Section III, we combine the ElGamal cryptographic scheme with isonumber theory to produce our new scheme which we called the Iso-ElGamal Cryptographic Scheme. In Section IV we look at the application of this system using integers and binary data. Section V is devoted to our conclusions.

II. SANTILLI'S THEORY OF ISONUMBERS

An important tool in science and engineering is symmetry. R. M. Santilli was the first to observe that new symmetry relations can be obtained for a fixed physical system via a change in the definition of multiplication in the underlying algebra. He called these new algebras Lie-isotopes, by analogy with a similar phenomenon in nuclear physics where the same atom can have a varying number of neutrons. In this section, we introduce Santilli's Theory of Isonumbers, which is among the tools for this paper (see [2], [7] and

[8]).

Let $(\mathbb{R}, +, \cdot)$ be the real numbers with addition and multiplication. We define the isodual numbers by inverting the direction of the real number line. Define \hat{a} by $\hat{a} + a = 0$. Let $(-1) = \hat{1}$ be the new unit, so that $\hat{1} * \hat{1} = \hat{1}$ and $\hat{a} * \hat{b} = (-a)(-1)(-b) = \hat{a}\hat{b}$. Thus, $(\hat{\mathbb{R}}, +, *)$ is also a representation of the same real numbers.

$$\leftarrow \hat{\mathbb{R}} : 0 : \mathbb{R} \rightarrow$$

A related example arises if we replace our unit 1 by any nonzero number $a = \hat{1}$. Since $\hat{1} * \hat{1} = \hat{1}$, we must have $*$ = a^{-1} . It is easy to check that $\hat{b} = ab$ and $\hat{b}^{-1} = ab^{-1}$. Thus, $(\hat{\mathbb{R}}, +, *)$ is yet another representation of the real numbers. However, now we have the option of choosing a in \mathbb{R} , in \mathbb{C} or any other field \mathbb{F} . In the first case, we obtain a theory of the second kind. In the other cases, we have Santilli's theory of the first kind.

Our scheme is based on the theory of the first kind, where the new unit is not in the field. Let a field $(F, +, \cdot)$ be given (e.g., Z_p , with p prime) and chose an invertible element $\hat{I} = \hat{T}^{-1} \notin F$ for our new the isounit. We introduce a new definition of multiplication on \hat{F} using $*$ = \hat{T} . This allows us to define a new field $(\hat{F}, +, *)$, isofield of the first kind, with rules given by:

$$\begin{aligned} \hat{a} &= a\hat{I}, & \hat{a} * \hat{b} &= (a\hat{I})\hat{T}(b\hat{I}) = ab\hat{I} = \widehat{ab} \\ \hat{a} + \hat{b} &= (a\hat{I}) + (b\hat{I}) = (a + b)\hat{I} = \widehat{a + b}. \end{aligned} \quad (1)$$

$$\hat{0} = 0\hat{I} = 0$$

III. THE ISO-ELGAMAL CRYPTOGRAPHIC SCHEME

This section describes our proposed Iso-ElGama cryptographic scheme. We first introduce the traditional ElGamal system and add the modifications that lead to the Iso-ElGamal scheme.

A. Traditional ElGamal Cryptographic System

The ElGamal cryptographic system is a widely used public-key system that was proposed by Taher ElGamal in 1984. It is based on the Diffie-Hellman key exchange protocol and its security depends on the difficulty of solving the Discrete Logarithms (DL) problem over a finite field, where the plaintext and ciphertext are integers less than some prime number q . The next few sections explain each of these steps [9], [11].

1) **The ElGamal Public and Private Key Generation:** We assume a communication session is taking place between a sender S and a receiver R . First, a large prime numbers q is

randomly selected on the receiver side, and a primitive root g modulo q is chosen. The receiver selects a private-key x such that $1 \leq x < (q - 1)$ to compute $a \equiv g^x \pmod{q}$, where the symbol \equiv represents the congruent relation; x is also later used for decryption. The public-key is given by $PU = \{q, g, a\}$, while the private is $PR = \{x, y\}$, where y represents the private encryption key selected by the sender such that $1 \leq y < (q - 1)$.

2) **The ElGamal Encryption Algorithm:** On the sender's side, let M be the message or plaintext to be encrypted using the ElGamal scheme with $M < q$. We encrypt the message M using the public-key a, g, q , and the private-key y , obtaining ciphertexts C_1 and C_2 . These values are used to compute:

$$K \equiv a^y \pmod{q}, C_1 \equiv g^y \pmod{q} \text{ and } C_2 \equiv M \times K \pmod{q}$$

3) **The ElGamal Decryption Algorithm:** The decryption of the ciphertext (C_1, C_2) using the private-key x and the public-key q to obtain the plaintext M is performed at the receiver side. If K^{-1} is the multiplicative inverse of K modulo q , the message M is obtained by computing $K \equiv (C_1)^x \pmod{q}$, and $M \equiv C_2 \times K^{-1} \pmod{q}$.

B. Iso-ElGamal Cryptographic System

The Iso-ElGamal cryptographic system is described in this section. Similarly, consider a communication session between a sender S and a receiver R . The Iso-Elgamal cryptographic scheme consists of key generation, encryption, and decryption phases described below.

1) **Iso-ElGamal Public and Private key Generation:** The public and private isokeys generated for the communication between sender S and receiver R uses a randomly chosen large prime number q and a primitive root g modulo q . In addition, a large positive integer isounit \hat{I} is chosen to represent the shared secret encryption-decryption isokey and used to compute the following:

$$\hat{q} = q\hat{I} \quad \text{and} \quad \hat{g} = g\hat{I} \quad (2)$$

After selecting a private decryption key x such that $1 \leq x < (q - 1)$ and computing $\hat{x} = x\hat{I}$ on the receiver side, one computes

$$\hat{a} = (g^x)\hat{I} \pmod{\hat{q}} \quad (3)$$

The sender selects a private encryption key y with $1 \leq y < (q - 1)$ and uses it to calculate the private encryption isokey $\hat{y} = y\hat{I}$. The public isokey is given by $PU = \{\hat{q}, \hat{g}, \hat{a}\}$, and the private and secret isokeys are $PR = \{\hat{x}, \hat{y}, \hat{I}\}$.

The values x, y , and \hat{I} can be selected to be large prime numbers in order the make the factorization of $\hat{x} = x \times \hat{I}$

to obtain x and \hat{I} from the private isokey \hat{x} more difficult for unauthorized users who have access to \hat{x} . Similar for the the factorization of $\hat{y} = y \times \hat{I}$.

2) **The Iso-ElGamal Encryption Algorithm:** The message $M < q$ with corresponding isonumber $\hat{M} = M\hat{I} < \hat{q}$ can be encrypted by the sender who has access to the public isokeys $\hat{a}, \hat{q}, \hat{g}$, the private key y and isokey \hat{y} , in addition to the secret isokey (isounit) \hat{I} . In addition, the sender can compute $a = \hat{a}\hat{I}^{-1}$ and $g = \hat{g}\hat{I}^{-1}$. With this information, the sender and calculate \hat{K} :

$$\hat{K} = (a^y)\hat{I} \pmod{\hat{q}} \quad (4)$$

With \hat{K} known and using $*$ = \hat{I}^{-1} , we obtain the ciphertext \hat{C}_1 and \hat{C}_2 :

$$\hat{C}_1 = (g^y)\hat{I} \pmod{\hat{q}} \quad (5)$$

and

$$\hat{C}_2 = (\hat{M} * \hat{K}) \pmod{\hat{q}} \quad (6)$$

The values of \hat{C}_1 in (5) and \hat{C}_2 given by (6) represent the pair of ciphertext (\hat{C}_1, \hat{C}_2) obtained using the Iso-ElGamal encryption algorithm.

Alternatively, if the values of K, C_1 , and C_2 obtained from the ElGamal encryption algorithm are first calculated, the values of \hat{K}, \hat{C}_1 , and \hat{C}_2 can be computed as

$$\hat{K} = K\hat{I} \text{ and } \hat{C}_1 = C_1\hat{I} \text{ and } \hat{C}_2 = C_2\hat{I} \quad (7)$$

3) **The Iso-ElGamal Decryption Algorithm:** The Iso-ElGamal decryption algorithm is implemented at the receiver side who has access to the private key x , isounit \hat{I} , and computed isokey \hat{x} , in addition to the public isokey $PU = \{\hat{q}, \hat{g}, \hat{a}\}$. We also assume that the ciphertext \hat{C}_1 and \hat{C}_2 have been transmitted to the receiver. The receiver already has access to the private key x and can compute the quantities

$$q = \hat{q}\hat{I}^{-1} \text{ and } C_1 = \hat{C}_1\hat{I}^{-1} \text{ and } C_2 = \hat{C}_2\hat{I}^{-1} \quad (8)$$

and use them to find

$$\hat{K} \equiv (C_1)^x \hat{I} \pmod{\hat{q}}. \quad (9)$$

Using \hat{K} , the receiver can calculate

$$K = \hat{K}\hat{I}^{-1} \quad (10)$$

and use it to compute

$$Z = K^{-1} \pmod{q} \quad (11)$$

and

$$\hat{Z} = Z\hat{I}. \quad (12)$$

The value of \hat{M} is given by

$$\hat{M} = (\hat{C}_2 * \hat{Z}) \pmod{\hat{q}\hat{I}} = (\hat{C}_2\hat{I}^{-1}\hat{Z}) \pmod{\hat{q}} \quad (13)$$

where $*$ = $\hat{I} = \hat{I}^{-1}$.

Finally, the original message is obtained as

$$M = \hat{M}\hat{I}^{-1}. \quad (14)$$

Thus, the original message M encrypted on the sender side is retrieved.

IV. APPLICATION OF THE ISO-ELGAMAL CRYPTOGRAPHIC SCHEME

In this section, we explore a few implementations of the Iso-ElGamal scheme. First, we consider the encryption and decryption of messages represented by positive integers less than the chosen isounit \hat{I} . Secondly, we consider information or data represented by bits. Results for each were obtained using simulation software we have written using Mathematica 12 software package.

A. Integer Encryption/Decryption Application Case

The Iso-ElGamal is used here to encrypt and decrypt information encoded into an integer less than the isounit \hat{I} . Consider the encryption and decryption of the message $M = 24$ using the Iso-ElGamal system.

Isokeys Generation Phase:

The generation of the public isokeys $PU = \{\hat{q}, \hat{g}, \hat{a}\}$ and the private and secret isokeys $PR = \{\hat{x}, \hat{y}, \hat{I}\}$ needed for encryption and decryption are shown here. They are extremely large with a number of digits generally around 3,100 digits corresponding to a number of bits around 10,298 bits. These large number of bits are necessary for highly computationally secure isokeys and ciphertext against attacks such as brute force. Displaying all values from $PU = \{\hat{q}, \hat{g}, \hat{a}\}$ and $PR = \{\hat{x}, \hat{y}, \hat{I}\}$ will take many pages because of their large sizes.

So, only the values of \hat{I}, q, g, x and y are displayed here. Other values such $\hat{q}, \hat{g}, \hat{a}, \hat{x}$, and \hat{y} can be computed from these values as explained above.

We randomly generate a large positive integer with 3,000 digits representing the secret isokey (isounit) \hat{I} used for encryption and decryption. It is given by

$\hat{I} = 61881370267169220159050077054207864289330937$
 $579014488754331730779513369711860670570752991218$
 $203328984814229260849681302926468369935272897714$
 $913781912215240794551615431053101471501236194818$
 $615620998707784606204161687419094614024278468461$
 $338460034114343913181960857804983871734756332281$
 $202072234821486894849169482716562950393435022454$
 $903927923952741418872832292925952367899340557519$
 $384588492915415355635465657365139917284078126256$
 $902153735847912948587429823533334524216858764237$
 $466620478131975110753709260257095996374599822956$

396080483594566462786325030981891315592079856179
 279578541902757374721821160492419616404793039331
 512980625253763240434607163101589792475481704800
 180808970298023647418381929608472819721488858670
 224232419213274975934801925465441716403787466699
 718350923591983090622431606502216223377816315346
 257716845160269699602997615866606439456960277195
 172608411813124451716945845459957347624422295253
 314065854390029994769891660895886531926586964324
 666543478092048089429731728841703991421094064219
 316907865361957387386802250998523622321934820235
 511974168528430429105573910249489357846432324187
 419979890010052476397351673353243812401575199268
 903158245543230365012148488638838344602713437028
 892754759255287611446586256177588098634806969675
 119178793565835557610956550286881571834419701638
 370053088320081553492598944295686892770212104853
 646056522556503568518744320726057694170299261361
 876106549981391934524274903788767572694627774031
 852949267199958500374410355909551677181330583808
 864038822852514673124133950692310516885141297018
 095556788317259936089218753281638604143781530
 704548546680922829551120050709060612597121995270562
 651927318794641281433752687143155478038208859576803
 913889307813596920297882872040743123833071379236815
 933881678311527295901269135549344427575000313266349
 047261098301624964854803121046420384292108617692518
 225888350216651927628222013387350695788078954390688
 108817817653875772352985274320857805207531807013566
 687088783180583922544329301563460955614109587621637
 249432541569183269101190086239597131061106835589295
 170729379887434466287694194251744531981809121273759
 824584274350045386282978734580995956417097524326332
 599792588561293052665420305698048678182250360622401
 614001055976047861465753228249170424063101185061400
 246112493486142916445018740806604170257125355153262
 970911894357392824059441592827283598563125583167266
 501602037169251722484499831922779343109571102986202
 617386168138950688311032278345062795788687030358633
 732974912320614132806891491934543182530548361640510
 950938524122884511715789559803450206764209978490554
 59996821372361410428574932937023345404059188889380
 194369987048130846145178364908683504802551928728831
 957203092412297222754968312021419730904820971668862
 639441807179083335932844484198681451951316320655318
 559990043628023870508175339811481314322882751721867
 259079934665408845698696149292331588152152626075098
 628195234780957663347909679072078829416147973749267
 939708864175862537671548454044062051997705171845977
 7615850482318922471245844118671070132470742916.

A prime number q with 100 digits is randomly generated, and a primitive root g modulo q with 100 digits is selected.

The value of q is

$q = 9760942796514275765881544370070504554461264179752849592070224700654377607327139902694438634730098111$

The value of g is

$g = 2440235699128568941470386092517626138615316044938212398017556175163594401831784975673609658682524535$

The public isokey \hat{q} is calculated as $\hat{q} = q\hat{I}$, and it has 3, 100 digits and 10, 298 bits but not displayed here.

The public isokey $\hat{g} = g\hat{I}$ with 3, 100 digits and 10, 296 bits is also calculated but not displayed here.

The private key x with 99 digits and 3, 28 bits is randomly generated and given by

$x = 367119322642078637789512557541048667547257023148314118526182175382629287616953592747817873242399922.$

The private isokey \hat{x} is calculated as $\hat{x} = x \times \hat{I}$, and it has 3, 099 digits and 10, 293 bits, also not displayed here.

We also calculate $a = g^x \pmod{q}$, and it has 100 digits and 332 bits. It is given by

$a = 8670140956271752363721691551414417075934976627920447306324435831863875366409938603386352791932708352.$

The associated isokey \hat{a} is computed as $\hat{a} = a \times \hat{I}$, and it has 3100 digits and 10298 bits, also not displayed here.

The private encryption key y is randomly generated and has 99 digits and 329 bits, and it is given by

$y = 695946457771683057076364591958034612150810033638431945316822573781604326896217519115097920359131458.$

The corresponding private isokey \hat{y} has 3, 099 digits and 10, 294 bits, and it is computed as $\hat{y} = y \times \hat{I}$. Also, not displayed here. So, the public isokeys is $PU = \{\hat{q}, \hat{g}, \hat{a}\}$ and the private isokey $PR = \{\hat{x}, \hat{y}, \hat{I}\}$ are all computed and shown above.

Encryption Phase:

The message $M = 24$ is encrypted to obtain the ciphertext \hat{C}_1 given by (5) and \hat{C}_2 given by (6). The cipher \hat{C}_1 has 3100 digits and is given by

$\hat{C}_1 = 1714037909774206145505461105140523417249561350039986353561275336287536866609280839503888414711111059625655163348700768465667987236900261074357370323748859993359280657567931339004379094883734345930033424964532337140800538454009517906217429454282411897012849887271288537750813319644899978061163571842769671456170601013716821741717957125659538111473655636828821545250313215263073235$

514390781051978906520980958142339646235569184482
 018858115010991261203300538002764805766185950061
 075859292891478461986491295796737512681638027642
 935125498653677481750620848484746193248561072027
 216314607409933076950805568666801040
 191530975743734802262083575360598987888557482537633
 438151421834407830191265234209634649581417915995347
 639422320800305569285415630596315354962145855939261
 002475900879459761972868658647438550044686448080286
 718977303256322694904790726431261374461930448015775
 606306946058364935518081591098343850639617410586145
 225373952014297118713322739252628981845680577825221
 398767095703435754442035380002999136885449622161562
 806068645569778802956963083598603536413806225541287
 418755024873738427269472555268416454794730094113825
 792145653919316392746921631166632271882691697935918
 279207039280697227189756192036853440323762421234974
 778909068630952030301037328331930289704815788594801
 979444854227801427731551860595226062192322650251237
 00097101534625779055437522378895565996415185039118
 816417693228059506149480573095643695342986178125152
 604887088437202907725420745371482821913923616947300
 463318180716226621341422595335001767116651625391393
 449150955114758591346510212507865342968773740830438
 651621574448580667875600797683273413262684334571687
 906813800039290739122743820307260544884517342133332
 821722123076337004477554024398173308410081204140405
 569578374668984538799689116924745623529319583865387
 855033486792593848477103451997218565787076864854575
 080052527412118600369393876889882340353660769173807
 061217691652249421731209958346970012764812641988770
 589307496243993653955909328229765125174087876462467
 718575531788193599615614128750184250075520298830893
 081187844368209454174377000351461480456267898719530
 024576378357534640087756145854111964815347482534167
 956642573954559652358700898615741691599816411810244
 542077709620087879936088925341813580939592363817427
 060964586466163237891626583773518991854819255691000
 724265759163603374950727071707989561806069045094685
 697631482770559956873515579386018936764305710541058
 150847710637706625179684174754894532420146590020334
 031154561149615059853345407288235450577584458667780
 934713174800511252925147489487897373290374464625857
 024999043234800838831972371061323914492859411494949
 075535648582674516331487797498215718378453871091097
 998763553362635009196860567938933831912504542089914
 838792541283205237130914197702908530291669324658070
 485787534789671955235487524310399600923270879962076
 275733310573228990311672361771057422950129516994035
 934922806464438325446486732011341849305173019338593
 893383671510602628547321822615519162265459736815876
 223772734079579828803751244515174401169411005069391
 008835061298368010743473745144624101310524689501

042401049148567172474199115026633387361569517096.

The ciphertext \hat{C}_2 has 3099 digits and is given by
 $\hat{C}_2 = 46860827767964398094898339966979613025356$
 348293772629393436390022940905314350176068226283
 816303158352125469043862775441001856011677999173
 004872173416033611603907295073604033549882854421
 752198371196261079501905709430614645658570911217
 307965999992016336077649179316894018876700976035
 927000914751908395113543596943557449268088734440
 623769140972060324137760638835508103791103370030
 974690311691971071657095927675091332287260785036
 594907552290819661401729778165038315906300276975
 089993085824103340199079573463440332228109420549
 666209310941796477682715204699789980460705735451
 172485618741133178617445994626727554636492796792
 468835325904928859202014971588815228345951650743
 405587521401407960891426817608768772675321118609
 684509530288935791996333486569123696545472885785
 272009929754280139272064034233073926454483504324
 806009726756930488077134486057327660997714115015
 492099695260358951405001267188384350549359263903
 134359645839892997639552586418499585066133130916
 618743203127720702157643756799772747391287241887
 464718456091675177387417922065369783552090305128
 635137028172349892153610282332647667347057975464
 776146326041705526915795231725036198131112516947
 326082552116261878985220239789552536935564810440
 882369633694698076339285442812379487013026594419
 384526672788620620545075200845581144469574942407
 699936111510928293923133793967400366693383220125
 711376444162835016645242470621230625801854109871
 130503814711326905811866707926172869443342678471
 305794415816063061353831222411532662539810257732
 176410810619082948373964801525261015824321074412
 551474646720582802257672851825873306548652649237
 780909336051657668253453445387745384421828907856
 750331708856179634250392442057838763860314051111
 385261090459634797142422218378029095178311304390
 512716449874943550149255059211949732998706515475
 968101105684260716971276257109296380444067534636
 134416598922582923373002736074265951233964544342
 290721647960473190767025773547532022726241757462
 705783967248479006745867858203927656825557106133
 072530469944151699491727152621390966170702500207
 720087855042132297590498013100941628676210490131
 382506672373345128347753307379831350097828115329
 896040807596213230063957856870642174139361121986
 274847793190495074979320149537058412343305874279
 663219810331325890044516829924036704888385447554
 884709446991725732429972009792652617679621205151
 631331853538191389169721110105324403124582696052
 586104340051601236590957175404341925259061699687

664430210066383930121459850012337362294304860475
 452590715401249792772993122907273067193099895287
 965362127309915682883786055336744370648273039439
 348827635936448130681786500989496063032853023806
 186226654590523174452640529091809665049690495864623
 170212695940220866981221514911356286870466096502231
 617490649757311054129216944957418303318485040758249
 295743096335803058743095608453791949870243696190944
 502273421574415849298418286382774524553185404639556
 237310687563534896277709821572960632622002741496987
 442267670086683086601317278531961983014353006295803
 147465617805137243360590841804330911423933344565463
 501221435122841800595522092427226973883353053503105
 797902526332365409930859545712844021434202395363706
 4492.

The pair of ciphertext (\hat{C}_1, \hat{C}_2) represents the secure encrypted message that will be store or transmitted through a channel that may be unsecured.

It takes only 0.0022954 seconds to encrypt the message. The computational cost time-wise is not large despite the large number of digits associated with the isokeys. The isonumbers are formed by having generally about one extra multiplication by \hat{I} to an original value such as a key in order to produce an isokey, for instance. So, a very high security based on very large isokeys and ciphertext is achieved with minimum computational cost.

Note that the values of the ciphertexts \hat{C}_1, \hat{C}_2 , and others values are very large and require larger memory size to store, or larger transmission time. This problem can be overcome by further processing and reducing them to smaller sizes that will be discussed in our next paper.

Decryption Phase:

The decryption of the ciphertext (\hat{C}_1, \hat{C}_2) is achieved using (8) through (14) to obtain \hat{M} with 3,002 digits but not displayed here. The final decrypted message obtained is $M = 24$, which corresponds to the original message that was initially encrypted.

The Iso-ElGamal cryptographic scheme may possibly takes us one step closer to an encryption scheme that is resistant to quantum computing-based attacks because of its characteristics, isokeys and ciphertext with a very large sizes.

The number of digits for each of the public isokeys and private isokeys is arbitrary, but the higher the number of digits, the more secure the isokeys and associated ciphertexts, but also the higher the computational cost.

B. Binary Encryption/Decryption Case Applications

The Iso-ElGamal scheme can be used to encrypt a bit represented by 0 or 1. Before encryption, each bit must

be first mapped to a positive integer greater than 1 in order to avoid having the associated ciphertext \hat{C}_2 equals 0 when encrypting bit 0. The isounit \hat{I} can also be given a different value for each bit of encryption.

For illustration purposes, consider encrypting the data byte 10110011 and then decrypting the associated ciphertext to recover the original data byte.

The public keys q and g used for each bit encryption are the same, as well as the private keys x and y are also the same when encrypting and decrypting each bit. But the associated public isokey $PU = \{\hat{q}, \hat{g}\}$, private isokey $PR = \{\hat{x}, \hat{y}\}$, and \hat{a} are computed using each associated isounit $\hat{I}_i, i = 1, 2, \dots, 7, 8$ for each bit as shown in Fig. 1 and Fig. 2. Also, before encryption, a bit 0 is mapped or encoded into integer $M = 3$, while a bit 1 is mapped or encoded into $M = 2$ for example, other values different to 2 and 3 can be used as well. The mapping of bit 0 to a value not equal to zero will avoid obtaining the trivial value $\hat{C}_2 = 0$ for the encryption of all bit 0 as shown in (6). Similarly, the mapping of bit 1 to a value other than 1 will avoid having the same value of \hat{C}_2 for all bit 1, when the values of \hat{k} and \hat{q} are the same for each bit 1 as shown in (6).

The keys used for the encryption/decryption of each bit are $q = 17, g = 11, x = 6,$ and $y = 4$, and the corresponding isonumbers $\hat{q}, \hat{g}, \hat{x}, \hat{y}$ are computed using the associated isounit $\hat{I}_i, i = 1, 2, \dots, 7, 8$ shown in in Fig. 1 and Fig. 2. The values of these isokeys are small and only for illustration purposes, but they must be made as big as possible in order to produce secure ciphertexts \hat{C}_1 and \hat{C}_2 .

For instance, to encrypt the first bit of 1 in the data byte, it is first mapped to $M_1 = 2$ before encryption using $\hat{I}_1 = 5$ to obtain ciphertexts \hat{C}_{11} and \hat{C}_{21} , where the second index corresponds to the index of the isounit used, while the first index distinguishes between the first and second cipher. Using $\hat{I}_1 = 5$, the computed isokeys are $\hat{q} = q\hat{I}_1 = 85, \hat{g} = g\hat{I}_1 = 55, \hat{x} = x\hat{I}_1 = 30,$ and $\hat{y} = y\hat{I}_1 = 20, \hat{a} = a\hat{I}_1 = [g^x \pmod{q}]\hat{I}_1 = 40, \hat{C}_{11} = 20$ and $\hat{C}_{21} = 75$, as shown in Fig. 1. The corresponding decrypted value is $M_1 = 2$, which is then decoded as bit 1, as shown in Fig. 2. Other bit values in the data byte are encrypted and decrypted in a similar manner and results are summarized in Fig. 1 and Fig. 2. Results on Figs. 1 and 2 are obtained from the Mathematica 12 simulation program which we have written.

In case that more data bytes need to be encrypted, the isounits \hat{I}_1 through \hat{I}_8 can be rotated left or right by a number of units $n > 1$ instead of choosing difference random isounits.

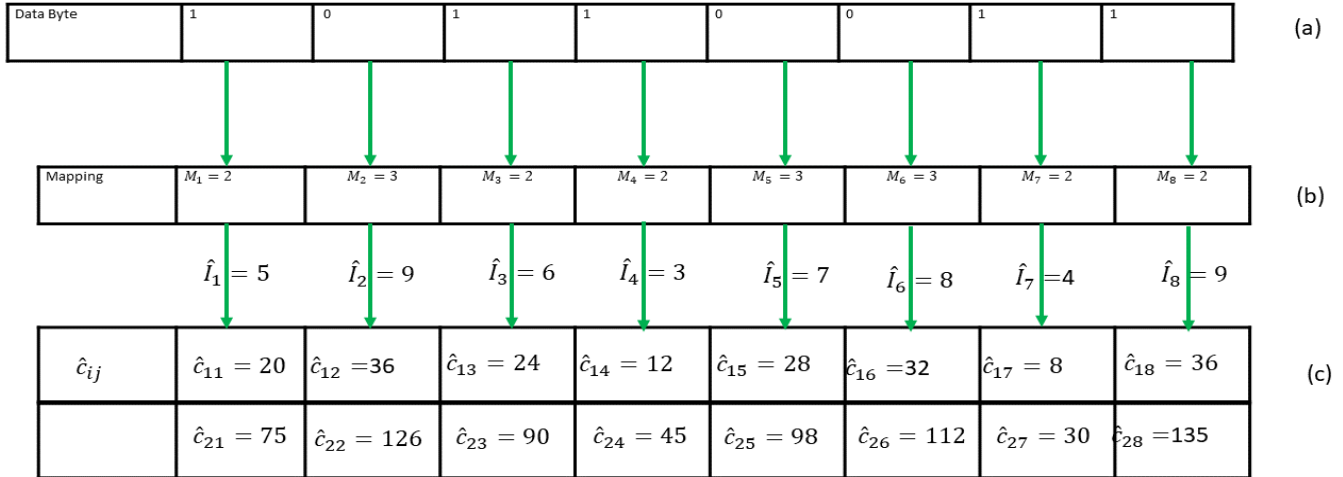


Fig. 1: Data byte Encryption using the ElGamal-Senegalese Cryptographic Scheme

- (a) Original Data Byte
- (b) Data Encoding, bit 0 mapped to 3, bit 1 mapped to 2
- (c) Encrypted data using corresponding isounit \hat{I}_i

Fig. 1: Data byte Encryption using the Iso-ElGamal Cryptographic Scheme

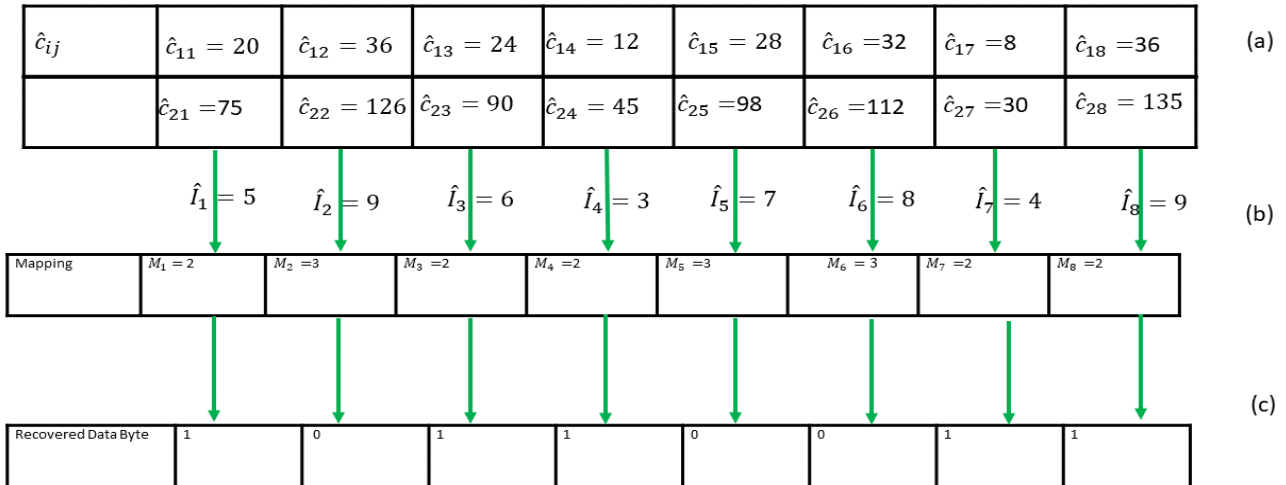


Fig. 2: Data byte Decryption using the ElGamal-Senegalese Cryptographic Scheme

- (a) Encrypted Data or ciphertext
- (b) Decrypted data or plaintext using corresponding isounit \hat{I}_i
- (c) Data Decoding, 3 mapped to bit 0, 2 mapped to bit 1

(a)

Fig. 2: Data byte Decryption using the Iso-ElGamal Cryptographic Scheme

V. CONCLUSIONS

In this paper, we have introduced the Iso-ElGamal cryptographic scheme which combines the traditional ElGamal Cryptographic System with Santilli's Isonumber Theory of the first kind to provide an additional degree of freedom of unit to the encryption-decryption of data encoded into positive integers or bits. Our approach includes both public-key or asymmetric encryption approach, and symmetric encryption using an isounit as share encryption and decryption isokey; therefore, leading to having a Hybrid public-key and private-key cryptographic system with a substantial increase in security of the associated ciphers. The cryptographic process which includes the key generation, encryption, and decryption phases is performed in an isofield whereby the new parameters are isonumbers. The paper also provides applications of the Iso-ElGamal scheme for the encryption of any data or information that can be encoded into positive integers, as well as bits. In addition, a wide range of applications such as confidentiality, data integrity, and digital signature can be implemented using this Iso-ElGamal cryptographic scheme. Our contribution includes the strengthening of the ElGamal cryptographic system using isonumber theory of the first through a mathematical description of the encryption and decryption approaches. Future work includes investigating how well the Iso-ElGamal scheme will resist quantum computing-based attacks.

REFERENCES

- [1] A. K. A. Hassan, Reliable Implementation of Paillier Cryptosystem, Iraqi Journal of Applied Physics, IJAP, Vol. 10, No. 4, October-December 2014, pp. 27-29
- [2] C. X. Jiang, Foundations of Santilli's Isonumber Theory, Fundamental Open Problems in Science at the End of the Millennium, Proceeding of the Beijing Workshop, August 1997, Hadronic Press, Palm Harbor, FL 34682-1577, U.S.A, ISBN 1-57485-029-6, PP. 105-139.
- [3] M. I. Wade, H. C. Ogworonjo, M. Gul, M. Ndoeye, M. Chouikha, W. Paterson Red Green Blue Image Encryption Based on Paillier Cryptographic Approach, World Academy of Science, Engineering and Technology, International Journal of Electronics and Communication Engineering Vol:11, No:12, 2017, <https://waset.org/abstracts/79232>
- [4] M. I. Wade, T. Gill, "Isonumbers and RGB Image Encryption.", Hadronic Journal, Algebras, Groups and Geometries, Vol. 37, 103-119 (2021.)
- [5] M. I. Wade, Distributed Image Encryption Based On a Homomorphic Cryptographic Approach, 10th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, Columbia University, New York City, USA 10- 12 October 2019 (IEEE UEMCON 2019).
- [6] P. Paillier, Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, J. Stern (Ed.): EUROCRYPT'99, LNCS 1592, pp. 223-238, 1999. @ Springer-Verlag Berlin Heidelberg 1999
- [7] R. M.Santilli-1, Foundations of Theoretical Mechanics II, Springer-Verlag New York, U.S.A, (1978), ISBN 0-387-08874-1.
- [8] R. M.Santilli-2, Isonumbers and Genonumbers of Dimensions 1, 2, 4, 8, their Isoduals and Pseudoduals, and "Hidden Numbers," of Dimension 3, 5, 6, 7, Algebras, Groups and Geometries, Vol. 10, 273 (1993).
- [9] R. Ranasinghe and P. Athukorala, A Generalization of the ElGamal public-key cryptosystem, Cryptology ePrint Archive, Report 2020/354, <https://eprint.iacr.org/2020/354>
- [10] R. L. Rivest, A. Shamir, L. Adleman, A Method for Obtaining Digital Signatures and Public-Key Cryptosystems, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MD 02139 E-mail address rivest@theory.lcs.mit.edu
- [11] W. Stallings, Cryptography and Network Security: Principles and Practice, 2003 Pearson Education, Inc.,

Connected and Autonomous Vehicles against a Malware Spread : A Stochastic Modeling Approach

Manal EL MOUHIB

Kamal AZGHIOU

Abdelhamid BENALI

Mohammed First University (UMP)
Oujda, Morocco
m.elmouhib@ump.ac.ma

Mohammed First University (UMP)
Oujda, Morocco
k.azghiou@ump.ac.ma

Mohammed First University (UMP)
Oujda, Morocco
a.benali@ump.ac.ma

Abstract—The proliferation of autonomous and connected vehicles on our roads is increasingly felt. However, the problems related to the optimization of the energy consumed, to the safety, and to the security of these do not cease to arise on the tables of debates bringing together the various stakeholders. By focusing on the security aspect of such systems, we can realize that there is a family of problems that must be investigated as soon as possible. In particular, those that may manifest as the system expands. Therefore, this work aims to model and simulate the behavior of a system of autonomous and connected vehicles in the face of a malware invasion. In order to achieve the set objective, we propose a model to our system which is inspired by those used in epidimology, such as SI, SIR, SIER, etc. This being adapted to our case study, stochastic processes are defined in order to characterize its dynamics. After having fixed the values of the various parameters, as well as those of the initial conditions, we run 100 simulations of our system. After which we visualize the results got, we analyze them, and we give some interpretations. We end by outlining the lessons and recommendations drawn from the results.

Index Terms—Security, Malware, Stochastic Model, Compartmental model, Autonomous and Connected Vehicle, CAV, Intelligent Transportation Systems.

I. INTRODUCTION

V2X is a generic term that specifies the interactions which a connected vehicle can have with its environment. The letter X in V2X can be V (V2V), I (V2I) or P (V2P). The V2V comprises the communication types occurring between vehicles. Whereas, V2I deals with the interactions that the vehicle can have with the infrastructure and vice versa. However, the last combination (V2P) represents the communication that can take place between the vehicle and the pedestrians.

All the X mentioned above can either be the target of one or more attacks, or their sources. Whatever the X in question, because it is part of an intelligent transport system, it is essential to protect it and protect oneself against it, otherwise the life of the human being is in danger [1]. However, apart from the human being, the autonomous and connected vehicle is both the most interesting and the most complex case to deal with. In addition to its mobile aspect, the complexity of such

a system, from the point of view of its security, lies in the extent of its attack surface that it presents [2] [3].

Despite all these fears, manufacturers are constantly producing new prototypes with increasingly sophisticated capabilities and functions [4]. Also, a segment of customers interested in this type of vehicle, not yet mature, is increasingly appreciated [5]. This means that converging towards a generation of society in which the word “driver’s license” would have no meaning is a certainty.

From what we have just said, implementing safety procedures and techniques dedicated to this new type of system, which will soon invade our roads, is a necessity. However, even if we have made enormous progress in this area, it is still insufficient [6]. Indeed, one can find works that have addressed the internal and local security of these systems as an isolated vehicle [7]. Others have addressed a more global aspect, given the local environment of such systems [8]. One can find other works which preferred to approach real cases [9]. Whatever the context of the work in question, it can be said that the conditions are not yet in place to draw definitive conclusions. Thus, researchers and engineers in such a situation always prefer to resort to modeling and simulation in order to predict what may happen in the future or in inaccessible situations (see sec. II).

The strength of the modeling and simulation approach is that, in most cases, we have tested them in areas relevant to the present for which we have enough experience. We can therefore think of reusing them, with precautions, to model phenomena relating to new domains for which we lack experience or access, as for the present work.

This work proposes to investigate the situation where most vehicles in circulation will be autonomous and connected vehicles. We consider the case of an environment where there is no trust. The problem that arises is, considering the conditions mentioned, what will be the dynamics of such a system if it is subjected to pandemic malware? To help answer this question, we propose in section III a stochastic model, based on what has been done for Covid-19, while adapting it to our case [10]. Solving the equations of the proposed model by simulation

will allow us to draw conclusions about certain requirements that a Connected and Autonomous Vehicle (CAV) system must meet in order to minimize the damage.

Besides the introduction, this article includes: (i) a section devoted to a review of the literature as close as possible to the subject treated (sec. II) and in which the research gap is expressed; (ii) a third section with several subsections all aimed at specifying quantities and operating the model (sec. III); (iv) the section (sec. IV) brings together everything concerning the simulation and the choice of the various parameters. Also, at the level of this section, one visualizes the results got after having numerically solved the model; (v) section V discusses the results got; (vi) In section VI, we end with a conclusion.

II. LITERATURE REVIEW

In our previous work [11], we studied a behavior of malware spreading over V2V channel by using a Multi-agent modeling approach. The results of the simulation show clearly that the density of traffic has a large impact in malware spreading. Then, the work on [12] presents an overview of the propagation of worm on the VANET network and refers to some research has dealt with the subject. The paper [13] examines the node movements in VANET's Network merged with the velocity dependent shadow fading model of wireless links between VANET nodes.

Besides, Authors of [14] provide an analytical study of the worm's propagation in the both static and dynamic traffic of urban system. They provide models for mobility, communication channel, Medium Access Control and worm propagation. And [15] focus on the spreading of worm and patches on highway corridor of 10km of length.

In [16], the authors discuss the parameters leading to active worm propagation on VANET. They describe a stochastic model based on the SIR epidemic model to define a worm's spreading on high and low traffic density. Although, other reseachers [17] propose a two-layer model named virus-traffic coupled dynamic model to study a virus propagation on V2V channel. this model combines the SIR model and cellular Automata.

Authors of [18] investigate the worm spreading through V2V communication and introduce a numerical model of worm spreading based on SIR model taking account Immunization. while those of [19] presents a SEIR-S Model based on VANET Network behavior and SIR Epidemic Model. A numerical analysis and Agent-based Modeling by Netlogo simulator are done to verify a proposed model.

To our knowledge, none of the previous work has solved the above problem using stochastic models reflecting more control measures. While the latter are widely used in the case of pandemic modeling, for example. We propose in this work to fill this gap by relying on epidemiologic modeling work while providing the necessary extensions and adaptations.

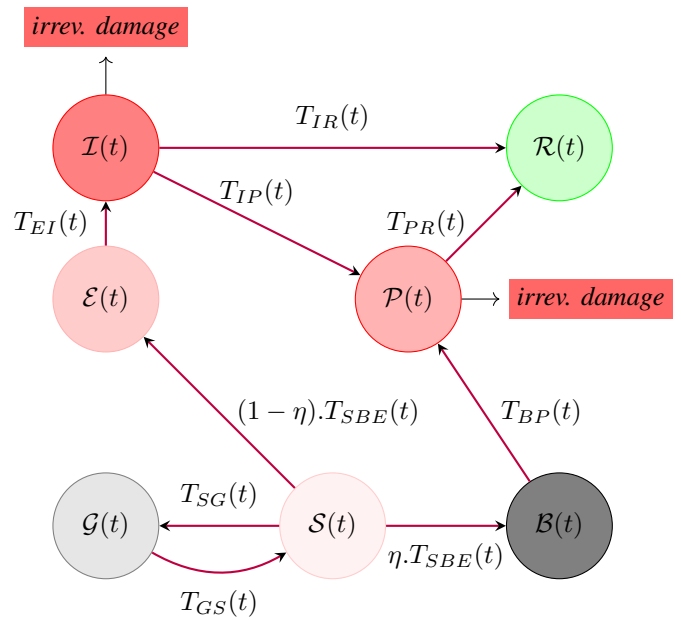


Fig. 1. Model dynamics graph

III. MODEL

A. Dynamics Graph

The vertices of the graph (Fig. 1) represent the sets in which a CAV could live. The *Susceptible* set, denoted \mathcal{S} , includes CAVs that are susceptible to certain malware-based attacks. This is the most populated set at the start of the malware spread. A CAV belonging to this set can evolve in three ways. Namely, it can transit to the *Blacklisted* set, denoted \mathcal{B} , if we can identify it as more likely to be infected than many other CAVs. Likewise, the *Graylisted* set, denoted \mathcal{G} , encloses CAVs facing moderate risk compared with the former which turned out to be false positives. Eventually, the exposed CAVs not belonging to either last both mentioned sets occupy the *Exposed* set, denoted by \mathcal{E} . Both the CAVs in the *Infected* set, denoted \mathcal{I} , and those in *Blacklisted* one can be members of the *Purging* set, denoted \mathcal{P} . The latest contains CAVs in sterilization stage. Upon achieving the purging process, and if the damage is reversible, the CAV becomes a new member of the *Recovered* set, denoted \mathcal{R} . we should mention that \mathcal{I} and \mathcal{B} are distinct from each other. On the one hand, the set \mathcal{I} comprises elements that are positively infected and under no control. Although \mathcal{B} may contain both infected and high-risk items, regardless of the situation, all of \mathcal{B} 's items are under control.

We labeled each vertex in the graph with the same letter of its underlying set as being a function of time. Indeed, it represents the cardinality of the considered set fixed at time t .

B. Inter-sets Transition Specification

A transition occurs from a set \mathcal{X} according to a stochastic process. We propose to model the number T_{ij} of CAVs leaving a set i and arriving at a set j , at time t , by a binomial

or Poisson process. In this work, we consider the progression of time t as a discrete sequence t_n . The following equations describe the transitions, as well as the various arguments and parameters involved.

Equation (1) gives the number of elements leaving set \mathcal{S} to arrive at set \mathcal{B} (resp. \mathcal{E}) at time t . Those who have probably been blacklisted. $\mathbb{P}_{S_{BE}}$ is the probability that a CAV leaves set \mathcal{S} and arrives at set \mathcal{B} (resp. \mathcal{E}).

The TABLE II gives the meaning of the other parameters.

$$T_{S_{BE}}(t) \sim Poisson(\mathcal{S}(t) \cdot \mathbb{P}_{S_{BE}}(t)) \quad (1)$$

Where

$$\mathbb{P}_{S_{BE}}(t) = 1 - \exp(-\beta \cdot c(t) \cdot h \cdot \mathcal{I}(t) / \mathcal{N}) \quad (2)$$

Equation (3) gives the number of elements leaving set \mathcal{S} to arrive at set \mathcal{G} at time t . Those who have probably been graylisted. \mathbb{P}_{SG} is the probability that a CAV leaves set \mathcal{S} and arrives at set \mathcal{G} .

$$T_{SG}(t) \sim Poisson(\mathcal{S}(t) \cdot \mathbb{P}_{SG}(t)) \quad (3)$$

Where

$$\mathbb{P}_{SG}(t) = 1 - \exp(\eta(1 - \beta) \cdot c(t) \cdot h \cdot \mathcal{I}(t) / \mathcal{N}) \quad (4)$$

Equation (5) gives the number of elements leaving set \mathcal{E} to arrive at set \mathcal{I} at time t . Those who have probably been infected. \mathbb{P}_{EI} is the probability that a CAV leaves set \mathcal{E} and arrives at set \mathcal{I} .

$$T_{EI}(t) \sim Bin(\mathcal{E}(t) \cdot \mathbb{P}_{EI}) \quad (5)$$

Where

$$\mathbb{P}_{EI} = 1 - \exp(-h \cdot \tau_{ei}) \quad (6)$$

Equation (7) gives the number of elements leaving set \mathcal{I} to arrive at set \mathcal{P} at time t . Those who have probably been purged. \mathbb{P}_{IP} is the probability that a CAV leaves set \mathcal{I} and arrives at set \mathcal{P} .

$$T_{IP}(t) \sim Bin(\mathcal{I}(t) \cdot \mathbb{P}_{IP}) \quad (7)$$

Where

$$\mathbb{P}_{IP} = 1 - \exp(-h \cdot \tau_{ip}) \quad (8)$$

Equation (9) gives the number of items leaving set \mathcal{I} to arrive at set \mathcal{R} at time t . Those who have probably been recovered. \mathbb{P}_{IR} is the probability that a CAV leaves set \mathcal{I} and arrives at set \mathcal{R} .

$$T_{IR}(t) \sim Bin(\mathcal{I}(t) \cdot \mathbb{P}_{IR}) \quad (9)$$

Where

$$\mathbb{P}_{IR} = 1 - \exp(-h \cdot \tau_{ir}) \quad (10)$$

Equation (11) gives the number of items leaving set \mathcal{G} to arrive at set \mathcal{S} at time t . Those who have probably been recovered. \mathbb{P}_{GS} is the probability that a CAV leaves set \mathcal{G} and arrives at set \mathcal{S} .

$$T_{GS}(t) \sim Bin(\mathcal{G}(t) \cdot \mathbb{P}_{GS}) \quad (11)$$

Where

$$\mathbb{P}_{GS} = 1 - \exp(-h \cdot \tau_{gs}) \quad (12)$$

Equation (13) gives the number of items leaving set \mathcal{B} to arrive at set \mathcal{P} at time t . Those who have probably been recovered. \mathbb{P}_{BP} is the probability that a CAV leaves set \mathcal{B} and arrives at set \mathcal{P} .

$$T_{BP}(t) \sim Bin(\mathcal{B}(t) \cdot \mathbb{P}_{BP}) \quad (13)$$

Where

$$\mathbb{P}_{BP} = 1 - \exp(-h \cdot \tau_{bp}) \quad (14)$$

Equation (15) gives the number of items leaving set \mathcal{P} to arrive at set \mathcal{R} at time t . Those who have probably been recovered. \mathbb{P}_{PR} is the probability that a CAV leaves set \mathcal{P} and arrives at set \mathcal{R} .

$$T_{PR}(t) \sim Bin(\mathcal{P}(t) \cdot \mathbb{P}_{PR}) \quad (15)$$

Where

$$\mathbb{P}_{PR} = 1 - \exp(-h \cdot \tau_{pr}) \quad (16)$$

Equations (17 and 18) represent the number of CAVs that experienced irreversible damage.

$$T_{Idam.}(t) \sim Bin(\mathcal{I}(t) \cdot \mathbb{P}_{Idam.}) \quad (17)$$

$$T_{Pdam.}(t) \sim Bin(\mathcal{P}(t) \cdot \mathbb{P}_{Pdam.}) \quad (18)$$

Where

$$\mathbb{P}_{Pdam.} = \mathbb{P}_{Idam.} = 1 - \exp(-h \cdot \delta) \quad (19)$$

C. Model Equations

Equations from (20) to (26) specify the dynamics of the proposed model. Being already mentioned that time is considered as a discrete variable, the step taken to solve them is Δt . It can be observed briefly that the evolution of the cardinality of each set is the difference between the number of CAVs having just joined this set at time t and that having just left it at the same time.

$$\mathcal{S}(t + \Delta t) = \mathcal{S}(t) - T_{S_{BE}}(t) - T_{SG}(t) + T_{GS}(t) \quad (20)$$

$$\mathcal{E}(t + \Delta t) = \mathcal{E}(t) + (1 - \eta) \cdot T_{S_{BE}}(t) - T_{EI}(t) \quad (21)$$

$$\mathcal{G}(t + \Delta t) = \mathcal{G}(t) + T_{SG}(t) - T_{GS}(t) \quad (22)$$

$$\mathcal{B}(t + \Delta t) = \mathcal{B}(t) + \eta T_{S_{BE}}(t) - T_{BP}(t) \quad (23)$$

$$\mathcal{I}(t + \Delta t) = \mathcal{I}(t) + T_{EI}(t) - T_{IR}(t) - T_{IP}(t) - T_{Idam.}(t) \quad (24)$$

$$\mathcal{P}(t + \Delta t) = \mathcal{P}(t) + T_{IP}(t) + T_{BP}(t) - T_{PR}(t) - T_{Pdam.}(t) \quad (25)$$

$$\mathcal{R}(t + \Delta t) = \mathcal{R}(t) + T_{PR}(t) + T_{IR}(t) \quad (26)$$

TABLE I
MODEL INITIAL VALUES

Symbol	Description	Value
\mathcal{N}_0	The Toatal Number of CAVs	10^6
\mathcal{E}_0	Initial Exposed CAVs	300
\mathcal{I}_0	Initial Infected CAVs	5
\mathcal{G}_0	Initial Graylisted CAVs	0
\mathcal{B}_0	Initial Blacklisted CAVs	0
\mathcal{P}_0	Initial Purging CAVs	0
\mathcal{R}_0	Initial Recovered CAVs	0

D. Isolation Function

Among the best strategies to adopt during an epidemic, or a pandemic, is the strategy of mobility restriction. We propose to adopt the same in the CAV community. However, as soon as we have not yet arrived at a generalist artificial intelligence comparable to that of humans, the isolation procedures may not be the same. One can, for example, think of memorizing all the CAVs with which the infected one has cooperated directly and repeat the same thing for a higher order of cooperation in order to apply the isolation. The questions that arise in this scenario and in others like it, with what latency should it be? And what impact will the choice of restriction time have on controlling the spread of malware? The injection of the function $c(t)$ specified in the equation (27), in the model allows us to know its usefulness.

$$c(t) = \begin{cases} c_0 & \text{if } t \leq t_{res} \\ (c_0 - c_{res})e^{-\kappa(t-t_{res})} + c_{res} & \text{if } t > t_{res} \end{cases} \quad (27)$$

IV. SIMULATION AND PARAMETERS ESTIMATION

A. Parameters and initial conditions Fixing

To simulate the model proposed in section III, we propose the values given in TABLE II and the initial conditions as specified in TABLE I. Because of the lack of data (futuristic scenario), we tried to fix the different values in the most rational way possible.

We implemented the simulation in a Python program. We have set the number of executions to 100. For each, we numerically solve the proposed model by specifying the potential evolution of each set or compartment. We plotted curves at the end of each simulation step (see Fig. 2). We should also mention that we automatically extracted the maximum values that appear in Fig. 2 and Fig. 3 from the simulation.

B. Simulation

Fig. 3 shows the evolution of the cardinality of the different sets considered in the model proposed in section III. The first remark which jumps to the eyes is that the evolution of the cardinality of the sets \mathcal{I} , \mathcal{E} , \mathcal{B} and \mathcal{G} on the one hand, and that of \mathcal{P} and \mathcal{R} have the same shape. However, the first family of these sets shows peaks, while the second tends towards a level from the thirtieth day. While we can clearly notice that the curve corresponding to the evolution of the cardinality of \mathcal{S} follows a completely different pace. In fact, the curve in question represents a degradation by oscillating until these oscillations attenuate and tend towards 0.

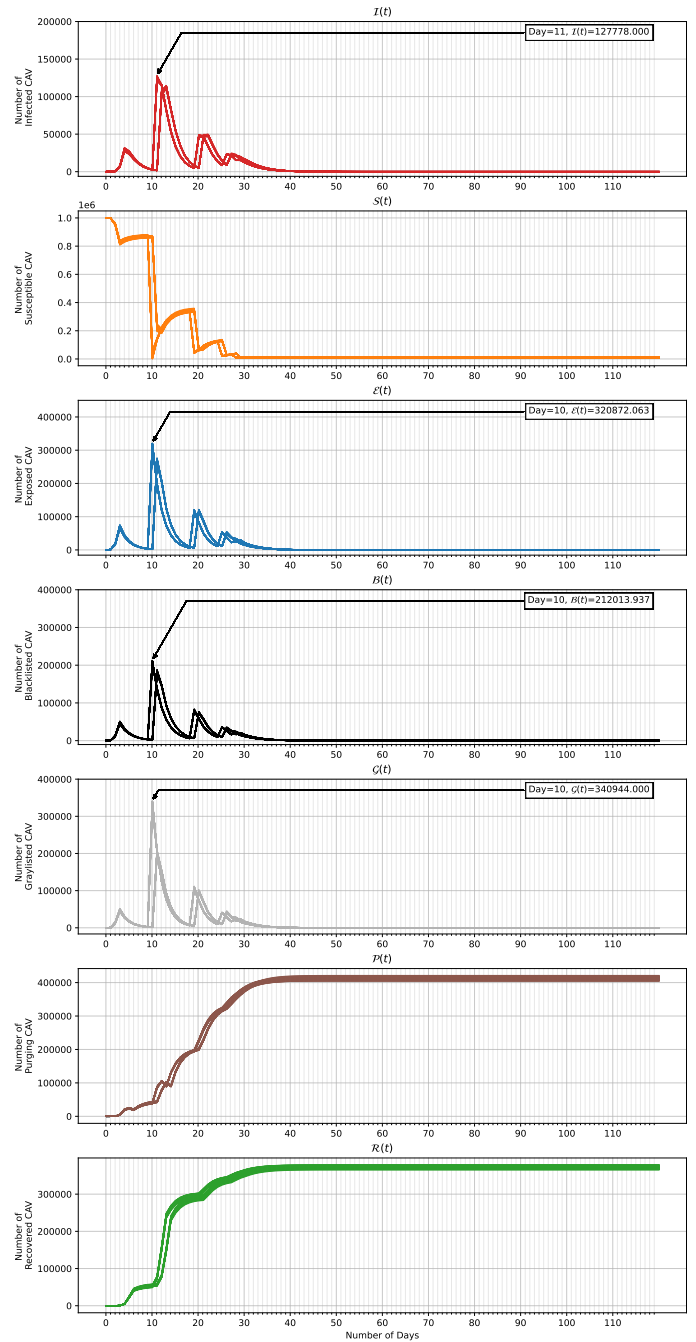


Fig. 2. Cardinality Evolution of the Sets \mathcal{I} , \mathcal{S} , \mathcal{E} , \mathcal{B} , \mathcal{G} , \mathcal{P} and \mathcal{R}

Let's inspect the collective behavior of the curves in Fig. 2. At time 0 until the third day, the curve $S(t)$ decreases so that it increases again until about the tenth day. Then, a sudden decrease is observed near this one. This pattern repeats until the set \mathcal{S} becomes empty. The drops are explained because the CAVs are leaving compartment \mathcal{S} to go towards \mathcal{G} , \mathcal{B} or \mathcal{E} . This corresponds to peaks in the curves of the latter. The increase in the cardinality of the host sets originates from a decrease at the level of the set \mathcal{S} , while the decrease of that of \mathcal{G} generates the increase observed in the curve of \mathcal{S} . The

TABLE II
MODEL PARAMETERS ESTIMATION

Symbol	Description	Range	Mean	Std	Baseline Value
c_0	Initial Cooperation Rate	[0, 800]	500	—	—
c_{res}	Cooperation Rate after Restriction	[0, 10]	0.9921	0.6468	1
k	Exponential decreasing speed of the cooperation rate	[0, 1]	0.0993	0.0101	0.1
β	Probability of Malware Injection	[0, 1]	0.5075	0.02908	0.5
η	Isolation Ratio of Exposed CAVs	[0, 1]	0.3975	0.0146	0.4
τ_{ei}	Transition Rate from Exposed to Infected State	[0, 1]	0.5000	0.0017	0.5
τ_{gs}	Transition Rate from Greylisted to Scuceptible State	[0, 1]	0.5	0.0023	0.5
τ_{ip}	Transition Rate from Infected to Purging State	[0, 1]	0.3001	0.0010	0.3
τ_{bp}	Transition Rate from Blacklisted to Purging State	[0, 1]	0.4501	0, 0009	0.45
τ_{ir}	Transition Rate from Infected to Restored State	[0, 1]	0.3	0.0010	0.6
τ_{pr}	Transition Rate from Purging to Restored State	[0, 1]	0.8000	0.0010	0.8
δ	Irreversible Damage Rate	[0, 1]	0.1999	0.0009	0.2

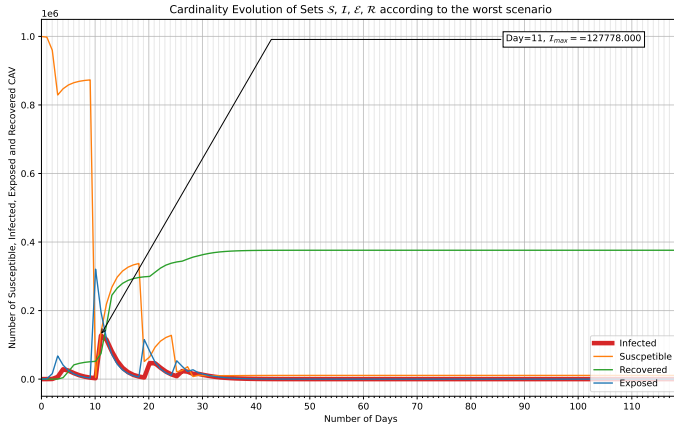


Fig. 3. Cardinality Evolution of Sets S , \mathcal{E} , \mathcal{I} , \mathcal{R} according to the worst scenario

cardinality of \mathcal{E} (resp. \mathcal{B}) generates an increase in that of \mathcal{I} (resp. \mathcal{P}). Set \mathcal{I} is unloaded in three ways: (i) The CAVs that have been detected as infected, transit to set \mathcal{P} to be purged, which explains the increase in $\mathcal{P}(t)$ in this period. (ii) the CAVs having been disinfected by means other than those recommended by the manufacturer or equivalent are found to belong to the set \mathcal{R} , which implies the increase of $\mathcal{R}(t)$. (iii) CAVs that have suffered irreversible damage are no longer part of the system. A part of \mathcal{P} can be lost forever while what remains is held by \mathcal{R} , hence the difference between the two levels, given the simulation parameters used.

The curves that we have just analyzed above are in fact families of curves corresponding to several simulation runs. The Fig. 3 represents the variation of the cardinality of the sets S , \mathcal{I} , \mathcal{E} and \mathcal{R} which are associated with the scenario manifesting the greatest peak of \mathcal{I} . This way of representing the curves supports the analysis that we made for Fig. 3. For example, the shift between the different curves shows the consistency that exists between the proposed model, more precisely, its dynamics, and the results got.

The reproduction number is widely used in epidemiologic [20]. In our context, it characterizes the power of propagation of a Malware within the framework of a CAV community. When we calculate this parameter in the presence of con-

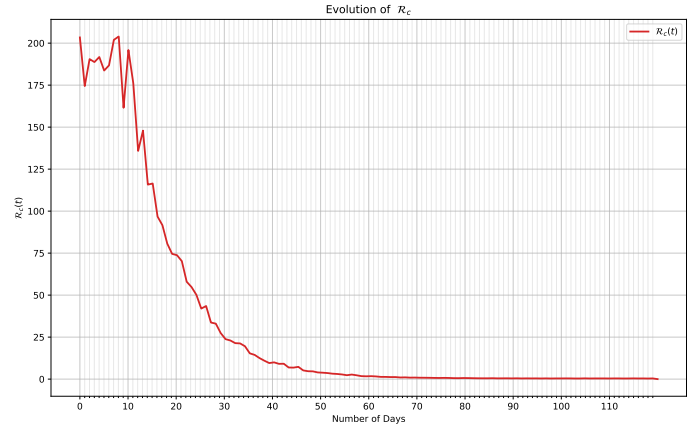


Fig. 4. Evolution of the R_c

trol measures, we rather speak of the effective reproduction number R_c . In this simulation, we have adopted the equation (28) which is like those which have been widely used in the literature while respecting its compatibility with our proposed specific model.

$$R_c(t) = \frac{\beta(t).c(t).(1 - \eta(t))}{\tau_{ip}(t) + \delta(t) + \tau_{ir}(t)} \quad (28)$$

In Fig. 4, we represent a sample of the stochastic behavior of the number $R_c(t)$. We can say that under the values taken by the parameters (see. TABLE II), the number of reproduction under control is $R_c(t) < 1$ around day 50. Which corresponds to about half of the period set for the simulation.

V. DISCUSSION

The proposed model and the simulation results got to make it possible to deduce the importance of the safety of transport systems and CAVs in particular. Also, we should add that this type of model also allows the prediction of the resources that we must put in place in order to control such scenarios. We can summarize the lessons to be drawn from this work in:

- (i) We recommend minimizing the T_{ij} numbers contributing to lead a CAV towards irreversible damage, such as fatal accidents.
- (ii) Minimizing the cardinality of \mathcal{G} by minimizing misidentification of CAVs that were suspected to be exposed.

This corresponds to minimizing targeting uncertainty. (iii) Increasing η increases the cardinality of \mathcal{B} , which means going from a less controlled system to a more controlled one. In this way, we reduce the cardinality of \mathcal{E} and subsequently that of \mathcal{I} . (iv) The dynamic of the model shows us that the sterilization stage undergone by the CAVs in \mathcal{P} requires an infrastructure dedicated to this type of operation in sufficient quantity and according to an adequate dispersion pattern. (v) The number T_{ir} sheds light on the fact that migration to \mathcal{P} is not the one and only possible choice. In very specific situations, such as communication failures, the probability of transition to \mathcal{P} can tend towards 0, which does not leave several choices to our CAV. So encouraging manufacturers to adopt the Open-CAV model, the twin of Open Source in software context, can be very useful.

VI. CONCLUSIONS

Through this work, we have investigated the case where a system of autonomous and connected vehicles would undergo the spread of malware at a certain infection rate. The usefulness of this modeling lies because a transport system is as critical as the life of the human being is.

This type of problem has been addressed in the literature in several ways. Namely, we can cite, among others, approaches based on multi-agent models, those based on epidemic models, either deterministic or stochastic. Finally, other researchers and engineers have preferred to investigate more realistic scenarios on prototypes or based on real incidents that may have occurred because of the premature injection of such systems into public roads.

We based our model on the SEIR model. To which we added two new sets that we named \mathcal{G} for Graylisted and \mathcal{B} Blacklisted. We have specified the variation of the cardinals of the different sets by binomial or even Poisson stochastic processes.

Afterwards, we ran 100 simulations of the proposed model for a certain set of parameters and initial conditions. The visualization of the results shows that the model controls the infectious situation well and truly under the stated hypotheses.

Finally, In the discussion section (sec. V), we drew some lessons and gave some suggestions for future work, such as the development of standards allowing third parties to invest in autonomous and connected vehicle systems.

REFERENCES

- [1] Vinayak V Dixit, Sai Chand, and Divya J Nair. Autonomous vehicles: disengagements, accidents and reaction times. *PLoS one*, 11(12):e0168054, 2016.
- [2] Amrita Ghosal and Mauro Conti. Security issues and challenges in v2x: A survey. *Computer Networks*, 169:107093, 2020.
- [3] Aljawharah Alnasser, Hongjian Sun, and Jing Jiang. Cyber security challenges and solutions for v2x communications: A survey. *Computer Networks*, 151:52–67, 2019.
- [4] DianGe Yang, Kun Jiang, Ding Zhao, ChunLei Yu, Zhong Cao, ShiChao Xie, ZhongYang Xiao, XinYu Jiao, SiJia Wang, and Kai Zhang. Intelligent and connected vehicles: Current status and future perspectives. *Science China Technological Sciences*, 61(10):1446–1471, 2018.
- [5] Manuel Alector Ribeiro, Dogan Gursoy, and Oscar Hengxuan Chi. Customer acceptance of autonomous vehicles in travel and tourism. *Journal of Travel Research*, 61(3):620–636, 2022.

- [6] Piergiuseppe Mallozzi, Patrizio Pelliccione, Alessia Knauss, Christian Berger, and Nassar Mohammadiha. Autonomous vehicles: state of the art, future trends, and challenges. *Automotive Systems and Software Engineering*, pages 347–367, 2019.
- [7] Jiayan Zhang, Fei Li, Haoxi Zhang, Ruxiang Li, and Yalin Li. Intrusion detection system using deep learning for in-vehicle security. *Ad Hoc Networks*, 95:101974, 2019.
- [8] Jian Wang, Yameng Shao, Yuming Ge, and Rundong Yu. A survey of vehicle to everything (v2x) testing. *Sensors*, 19(2):334, 2019.
- [9] Curtis R Taylor, Jason M Carter, Shean Huff, Eric Nafziger, Jackeline Rios-Torres, Bob Zhang, and Joseph Turcotte. Evaluating efficiency and security of connected and autonomous vehicle applications. In *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, pages 236–239. IEEE, 2022.
- [10] Sha He, Sanyi Tang, Libin Rong, et al. A discrete stochastic model of the covid-19 outbreak: Forecast and control. *Math. Biosci. Eng.*, 17(4):2792–2804, 2020.
- [11] Manal El Mouhib, Kamal Azghoui, and Abdelwahed Tahani. Analysis of the impact of traffic density on the compromised cav rate : a multi-agent modeling approach. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–6, 2021.
- [12] Azzedine Boukerche and Qi Zhang. Countermeasures against worm spreading: A new challenge for vehicular networks. *ACM Computing Surveys (CSUR)*, 52(2):1–25, 2019.
- [13] Maziar Nekovee. Modeling the spread of worm epidemics in vehicular ad hoc networks. In *2006 IEEE 63rd Vehicular Technology Conference*, volume 2, pages 841–845. IEEE, 2006.
- [14] Jian Wang, Yanheng Liu, and Kevin Deng. Modelling and simulating worm propagation in static and dynamic traffic. *IET Intelligent Transport Systems*, 8(2):155–163, 2014.
- [15] Lin Cheng and Rahul Shakya. Worm spreading and patching in inter-vehicle communications. *International Journal of Communication Networks and Information Security*, 2(1):50, 2010.
- [16] Syed A Khayam and Hayder Radha. Analyzing the spread of active worms over vanet. In *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, pages 86–87, 2004.
- [17] Lei Wei, Hongmao Qin, Yunpeng Wang, Zhao Zhang, and Guizhen Yu. Virus-traffic coupled dynamic model for virus propagation in vehicle-to-vehicle communication networks. *Vehicular communications*, 14:26–38, 2018.
- [18] Oscar Trullols-Cruces, Marco Fiore, and Jose M Barcelo-Ordinas. Worm epidemics in vehicular networks. *IEEE Transactions on Mobile Computing*, 14(10):2173–2187, 2014.
- [19] Duc Tran Le, Khanh Quoc Dang, Quyen Le Thi Nguyen, Soha Alhelal, and Ammar Muthanna. A behavior-based malware spreading model for vehicle-to-vehicle communications in vanet networks. *Electronics*, 10(19):2403, 2021.
- [20] Sanyi Tang, Yanni Xiao, Youping Yang, Yicang Zhou, Jianhong Wu, and Zhien Ma. Community-based measures for mitigating the 2009 h1n1 pandemic in china. *PLoS one*, 5(6):e10911, 2010.

CASSAVA LEAF DISEASE DETECTION USING DEEP LEARNING

Manick

Department of Computer Science
National Institute of Technology, Hamirpur
Hamirpur, India
17mi515@nith.ac.in

Jyoti Srivastava

Department of Computer Science
National Institute of Technology, Hamirpur
Hamirpur, India
jyoti.s@nith.ac.in

ABSTRACT

In this study, a clever plan to detect cassava leaves has been developed using a customized fine-tuned deep learning model. Five categories of diseases are used in this study: Cassava Brown Steak Disease (CBSD), Cassava Green Mite (CGM), Cassava Bacterial Blight (CBB), and Cassava Mosaic Disease (CMD) and Health. The results showed an accuracy on the test data obtained was over 77% on original problem using clever data augmentation without affecting the scope of this problem.

***Index Terms*— CNN; Convolutional Neural Network; Cassava Leaf**

I. INTRODUCTION

Cassava is a plant whose leaves are particularly rich in protein and vitamins[1]. Additionally, cassava is the key ingredient of rice. According to Bandar Lampung City's 2015, 2016 and 2017 Central Bureau of Statistics, cassava production totaled 5,323.00 tons, or second place with rice. Plants that produce 29,583.68 tons. However, output of cassava declined by 34.5 percent in 2018 compared to the previous year (Central Statistical Office of Bandar Lampung City, 2019). This is due to declining yields that occur due to disease outbreaks in cassava plants. Varieties The types of diseases in cassava leaves are caused by insects, viruses, bacteria and fungi [2]. This study investigated four common cassava leaf diseases, namely: Cassava Brown Steak Disease (CBSD), Cassava Green Mite (CGM), Cassava Bacterial Blight (CBB), and Cassava Mosaic Disease (CMD).

Infected leaves interfere with harvesting, as leaves are an important part of the plant that catalyzes photosynthesis, which is necessary process for plants, so that they can produce energy. Phloem tissues transport effects of photosynthesis in all other parts of plants. As long as the plant's leaves are in excellent condition and the photosynthetic process continues normally, the stems and clusters of the plant will reach maturity. However, when leaves become diseased and photosynthesis is disturbed, stem growth and root crops are also affected, reducing yields. Laboratory tests or plant help specialist is usually performed to diagnose cassava leaves. However, producers are unable of promptly and effectively combating cassava leaf disease. This is due to the fact that laboratory testing is often difficult due to cost constraints, and doctors cannot discover the sickness in time. Therefore, a clever solution to these problems is required as a pro-interactive python notebook pro, which can run on a browser using google

colab which is designed to work easily [3].

Previously, Adaptive Neural Fuzzy Inference System has been used in a number of research to detect flavonoid chemicals in guava leaves using artificial intelligence (ANFIS) [4]. Case-based reasoning[5], the Dempster-Shafer method[6], the separation of images using a soft computer[7], and comparable diagnostic techniques all utilise images. Alternatively, as these methods do not turn the training process into a database, they must be compatible with both the database training process and the network transformation in order to generate highly accurate results. Convolutional Neural Network is the name of the methodology proposed in this study (CNN)[8]. CNN's system has the most important effects on image perception. This is because the CNN method can process 2D data/images. A CNN's network has a special layer called a convolution layer. In this convolutional layer, the image input displays patterns in different parts of the image for easy separation. Based on this, the CNN method can read images easily and efficiently.

II. LITERATURE SURVEY

Authors	Description	Outcomes
Hany Elnashar et.al.	The author has used Resnet50, a pre-trained model to classify images into 5 given classes.	Using this pre-trained model, authors were able to achieve validation accuracy of around 64%[9].
H R Ayu, A Surtono, D K Apriyanto et.al.	The author has used MobileNetV2, a pre-trained model to classify images into 5 given classes.	Using this pre-trained model, authors were able to achieve validation accuracy of around 65.6%[10].
D.A. Bashish et.al.	The authors assumed that there should be a bad pixel in at least one of the clusters and used k-means clustering to divide the leaf image into 4 clusters.	Using this neural network, authors were able to achieve a classification accuracy of around 93%[11].
E.Kiani et.al.	Here, Two inputs were associated with iron deficiency and the other associated with fungal infections. If the leaves are sick, the output indicates two diseases.	Using this model, authors were able to achieve an overall system accuracy of 96% using the proposed algorithm[12].
M.Bhange et.al.	Clustering k-means is more efficient when applied to large data sets due to which they have used it in this research.	Using SVM model, authors were able to achieve an overall system accuracy of 82% using the proposed algorithm[13].

Authors	Description	Outcomes
G.Saradhambal, et.al.	Otsu's algorithm forms a bi-modal histogram with the foreground and background pixels. For feature extraction, shape and texture-oriented features were used.	The shape-oriented features used were area, color axis length etc whereas, contrast, homogeneity etc were the texture-oriented features.[14].
A Ramcharan, K Baranowski, et.al.	This article investigates transfer learning using a deep convolutional neural network (CNN) model for cassava picture datasets.	The Inception v3 model had the highest accuracies of 98% and 0.95% with the leaflet dataset for CBSD and GMD, respectively[2].
G Sambasivam, GD Opiyo et al.	This paper focuses on strategies like SMOTE (Synthetic Minority Over-sampling Technique) and focus losses with deep neural networks	This paper obtained more than 93% accuracy[1].
David Opeoluwa Oyewola, Emmanuel Gbenga Dada et al.	The research offers a unique deep residual convolution neural network (DRNN) for detecting CMD in photos of cassava leaf.	The DRNN model produced the best and achieved an accuracy of 96.75%[15].
Amanda Ramcharan, Peter McCloskey et al.	This paper proposes the "One-Vs-All" methodology for solving the multi-class classification problem of detecting disease in cassava plant's leaves.	The proposed model achieved an overall accuracy of around 85.64%[16].

III. DATASET AND EVALUATION METRIC

Pictures of cassava leaves found in the Kaggle competition, which includes five classes, namely: Cassava Bacterial Blight (CBB) (1087 photos), Cassava Brown Steak Disease (CBSD) (2189 photos), Cassava Green Mite (CGM) (2386) images, Cassava Mosaic Disease (CMD) (13158 images) and Healthy (2577 images). During the experiment, the database was divided into three parts: training data (80%) and testing data (20%). Fig 1-5. shows the examples of a cassava leaf images, which is a database by class. The main symptoms and pathogens that cause the appearance of cassava leaves are:

CBB is a disease caused by the bacteria *Xanthomonas manihotis*, which is typically found in humid environments[17]. The first symptoms of angular ulcers, tissue death at the site of infection shoots death. The whitefly virus *Bemisia tabaci* causes CBSD (Genn). There are two leaf symptoms that include yellow chlorosis in the bones of the upper extremities [2], another sign is chlorotic spots. The most common symptom of this disease is the development of dry sepia into brown, foam, and necrotic lesions in the nodular tissues. CGM is caused by the small leaf-eating insect *Mononychellus tanajoa*. This causes bleaching because chlorophyll enters the cells. As a result, the leaf turns into a dot, dies and becomes an abscissa. CMD is a virus caused by the begomovirus. Depending on the intricacy of the disease, the symptoms visible on the leaves contain a mixture of yellow and white sulphide particles. These particles can interfere with photosynthesis and plant growth. This causes a little decrease in the total yield[17].



Fig. 1. Cassava Bacterial Blight (CBB)



Fig. 2. Cassava Brown Streak Disease (CBSD)



Fig. 3. Cassava Green Mottle (CGM)



Fig. 4. Cassava Mosaic Disease (CMD)



Fig. 5. Healthy

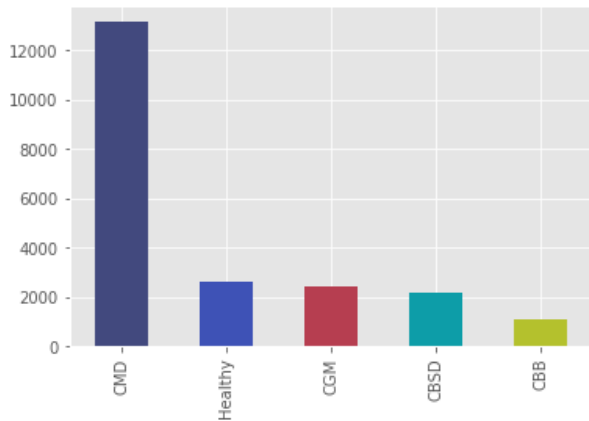


Fig. 6. Frequency of each class in the dataset

As we can see in Fig 6. input data is highly imbalanced, which can cause bias in the network. If we augment all the classes except Cassava Mosaic Disease (CMD) to around 13000, then we can get a dataset which will have nearly zero class-imbalance.

Numerous researchers have used accuracy as a measure of model performance[18][19]. Accuracy is the proportion of predictions that the model made correctly, i.e. the number of correct predictions in all predictions.

$$\text{Accuracy} = (T P + F N) / (T P + T F + F P + F N)$$

IV. METHOD

After deeply analyzing the frequency graph, this can be noticed that the class imbalance in the input dataset is at it's extreme especially due to Cassava Mosaic Disease (CMD). Previous techniques [9][10] used for augmentation are defective, as augmentation is applied over the whole dataset. Whereas, class Cassava Mosaic Disease (CMD) already contains 13,000 images in the original dataset. These techniques are actually contributing to increased class imbalance as internal data augmentation of tensorflow augments each image by a factor of around 6, which means CBB class will have around 3000 images after augmentation, whereas CMD class will have around 75000 images, which is accountable for huge class imbalance. This paper proposes a technique to resolve this issue by use of a custom image augmentation once with certain filters in the included images that can directly benefit the model such as:

- Gaussian Blur [20]
- Thresholding[21][22]
- Sharpening kernel[23][24]

Once we get the filtered data, we can generate the augmented data using the following pseudo-code:

```

1 traindf = load_dataset()
2 X_data = array()
3 Y_data = array()
4
5 for i in 1...size(traindf):
6     img = load_img(traindf['image_id'][i],target_size
7                 =(64,64))
8     X_data.append(img)
9     Y_data.append(label(img))
10
11     if current label is not Cassava Mosaic Disease:
12         X_data.append(rot90(img))
13         X_data.append(rot270(img))
14         X_data.append(flip_horizontal(img))
15         X_data.append(flip_vertical(img))
16         X_data.append(scale_in(img))
17         Y_data.append(label(img)*4)

```

Once we get the augmented data, we can then feed it into a custom CNN(Convolutional Neural Network), Fig 7. shows the architecture of the CNN used for the training. The augmented dataset generated has nearly 70,000 images of 5 classes, nearly 14,000 of each class.

To implement the model, we utilized Keras (<https://keras.io/>) environment with TensorFlow [25]. Python version 3.6 is used including standard libraries like Numpy, matplotlib, PIL, cv2, pandas etc. We implemented the model via Google Colaboratory (<https://colab.research.google.com/>) which included computer specifications as follows:

- GPU: 1xTesla K80, having 2496 CUDA cores, compute 3.7, 12GB (11.439GB Usable) GDDR5 VRAM
- CPU: 1xsingle core hyperthreaded, i.e., (1 core, 2 threads) Xeon Processors @2.3Ghz (No Turbo Boost), 45MB Cache
- RAM: 15.6 GB Available

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 64, 64, 256)	3328
average_pooling2d (AveragePooling2D)	(None, 32, 32, 256)	0
dropout (Dropout)	(None, 32, 32, 256)	0
conv2d_1 (Conv2D)	(None, 32, 32, 128)	131200
average_pooling2d_1 (AveragePooling2D)	(None, 16, 16, 128)	0
dropout_1 (Dropout)	(None, 16, 16, 128)	0
conv2d_2 (Conv2D)	(None, 16, 16, 64)	32832
average_pooling2d_2 (AveragePooling2D)	(None, 8, 8, 64)	0
dropout_2 (Dropout)	(None, 8, 8, 64)	0
flatten (Flatten)	(None, 4096)	0
dense (Dense)	(None, 512)	2097664
dropout_3 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 5)	2565

=====
 Total params: 2,267,589
 Trainable params: 2,267,589
 Non-trainable params: 0

Fig. 7. Architecture of CNN

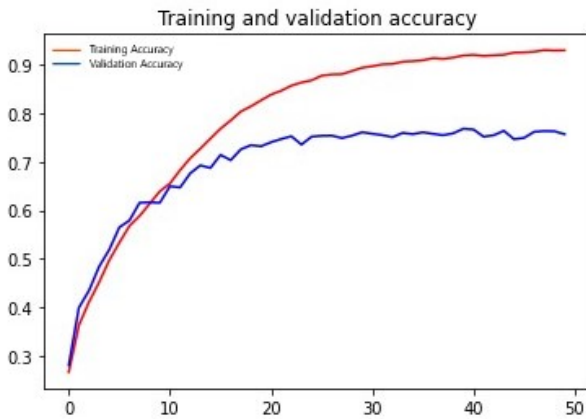


Fig. 8. Graph of Accuracy

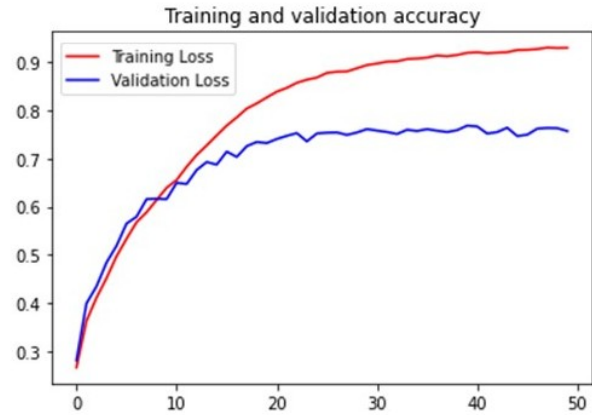


Fig. 9. Graph of Loss

V. ANALYZING THE PREDICTIONS

Many researchers have used the level of accuracy as a determinant of model performance [18][19]. Fig 9. shows the graph for Training and Validation Accuracy and Fig 10. shows the graph for Training and Loss. The performance of the model over augmented dataset is quiet good with training accuracy of over 95% and validation accuracy of over 77% respectively, which is a significant improvement over the original 64%[9] and 65.6% [10] validation accuracy.

Authors	Description	Accuracy
Hany Elnashar et.al[9]	The author has used Resnet50, a pre-trained model to classify images into 5 given classes.	64%.
H R Ayu, A Surtono, D K Apriyanto et.al[10]	The author has used MobileNetV2, a pre-trained model to classify images into 5 given classes.	65.6%.
Proposed paper	This paper proposes a better method of data augmentation as compared to what is used in above research papers.	77%

VI. CONCLUSION AND FUTURE WORK

As we can see in the Accuracy and Loss graph, this model is suffering from Overfitting [26]. Overfitting is a mathematical mistake that arises when a small number of data points are firmly connected to a small number of characteristics. As a result, this model is very useful for referencing other datasets too which contain high level of class imbalance. So, next step in this research is to prevent overfitting as training accuracy is already over 95%. Therefore, an improvement on validation accuracy can be achieved by overcoming this overfitting.

VII. REFERENCES

- [1] G Sambasivam and Geoffrey Duncan Opiyo, "A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks," *Egyptian Informatics Journal*, vol. 22, no. 1, pp. 27–34, 2021.
- [2] Amanda Ramcharan, Kelsee Baranowski, Peter McCloskey, Babuali Ahmed, James Legg, and David P Hughes, "Deep learning for image-based cassava disease detection," *Frontiers in plant science*, vol. 8, pp. 1852, 2017.
- [3] Teddy Surya Gunawan, Arselan Ashraf, Bob Subhan Riza, Edy Victor Haryanto, Rika Rosnelly, Mira Kartiwi, and Zuriati Janin, "Development of video-based emotion recognition using deep learning with google colab," *TELKOMNIKA*, vol. 18, no. 5, pp. 2463–2471, 2020.
- [4] Humairoh Ratu Ayu, Suryono Suryono, Jatmiko Endro Suseno, and Ratna Kurniawati, "Determination of the ultrasound power effects on flavonoid compounds from psidium guajava l. using anfis," in *Journal of Physics: Conference Series*. IOP Publishing, 2018, vol. 1025, p. 012024.
- [5] Olaide N Oyelade and Absalom E Ezugwu, "A case-based reasoning framework for early detection and diagnosis of novel coronavirus," *Informatics in Medicine Unlocked*, vol. 20, pp. 100395, 2020.
- [6] Sergio Peñafiel, Nelson Baloian, Horacio Sanson, and José A Pino, "Applying dempster–shafer theory for developing a flexible, accurate and interpretable classifier," *Expert Systems with Applications*, vol. 148, pp. 113262, 2020.
- [7] Vijai Singh and Ak K Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Information processing in Agriculture*, vol. 4, no. 1, pp. 41–49, 2017.
- [8] Nadia Jmour, Sehla Zayen, and Afef Abdelkrim, "Convolutional neural networks for image classification," in *2018 international conference on advanced systems and electric technologies (IC_ASET)*. IEEE, 2018, pp. 397–402.
- [9] Hany Elnashar, "Intelligent corp's leaf disease diagnosis for cassava using computer vision and deep learning approach," .
- [10] HR Ayu, A Surtono, and DK Apriyanto, "Deep learning for detection cassava leaf disease," in *Journal of Physics: Conference Series*. IOP Publishing, 2021, vol. 1751, p. 012072.
- [11] Dheeb Al Bashish, Malik Braik, and Sulieman Bani-Ahmad, "A framework for detection and classification of plant leaf and stem diseases," in *2010 international conference on signal and image processing*. IEEE, 2010, pp. 113–118.
- [12] Gurleen Kaur Sandhu and Rajbir Kaur, "Plant disease detection techniques: a review," in *2019 international conference on automation, computational and technology management (ICACTM)*. IEEE, 2019, pp. 34–38.
- [13] Manisha Bhangе and HA Hingoliwala, "Smart farming: Pomegranate disease detection using image processing," *Procedia computer science*, vol. 58, pp. 280–288, 2015.
- [14] G Saradhambal, R Dhivya, S Latha, and R Rajesh, "Plant disease detection and its solution using image classification," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 14, pp. 879–884, 2018.
- [15] Rafi Surya and Elliana Gautama, "Cassava leaf disease detection using convolutional neural networks," in *2020 6th International Conference on Science in Information Technology (ICSITech)*. IEEE, 2020, pp. 97–102.
- [16] Aryan Methil, Harsh Agrawal, and Varadh Kaushik, "One-vs-all methodology based cassava leaf disease detection," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 1–7.
- [17] Ernest Mwebaze, Timnit Gebru, Andrea Frome, Solomon Nsumba, and Jeremy Tsubira, "icassava 2019 fine-grained visual categorization challenge," *arXiv preprint arXiv:1908.02900*, 2019.
- [18] Sharada P Mohanty, David P Hughes, and Marcel Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in plant science*, vol. 7, pp. 1419, 2016.
- [19] Parul Sharma, Yash Paul Singh Berwal, and Wiqas Ghai, "Performance analysis of deep learning cnn models for disease detection in plants using image segmentation," *Information Processing in Agriculture*,

- vol. 7, no. 4, pp. 566–574, 2020.
- [20] Robert A Hummel, B Kimia, and Steven W Zucker, “Deblurring gaussian blur,” *Computer Vision, Graphics, and Image Processing*, vol. 38, no. 1, pp. 66–80, 1987.
 - [21] Derek Bradley and Gerhard Roth, “Adaptive thresholding using the integral image,” *Journal of graphics tools*, vol. 12, no. 2, pp. 13–21, 2007.
 - [22] David L Donoho, “De-noising by soft-thresholding,” *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995.
 - [23] Wenfeng Zhan, Yunhao Chen, Ji Zhou, Jing Li, and Wenyu Liu, “Sharpening thermal imageries: A generalized theoretical framework from an assimilation perspective,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 773–789, 2010.
 - [24] J Yuhendra, Hiroake Kuze, and J Sri Sumantyo, “Performance analyzing of high resolution pan-sharpening techniques: increasing image quality for classification using supervised kernel support vector machine,” *Research Journal of Information Technology*, vol. 3, no. 1, pp. 12–23, 2011.
 - [25] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., “{TensorFlow}: A system for {Large-Scale} machine learning,” in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.
 - [26] Tom Dietterich, “Overfitting and undercomputing in machine learning,” *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 326–327, 1995.

IoT-based DDoS on Cyber Physical Systems: Research Challenges, Datasets and Future Prospects

Manish Snehi

Department of Computer Science and Engineering,
Punjabi University,
Patiala, Punjab, India
snehi.manish@outlook.com

Abhinav Bhandari

Department of Computer Science and Engineering,
Punjabi University,
Patiala, Punjab, India
abhinavbhandari@pbi.ac.in

Abstract— The fusion of widespread IoT devices, highly adopted cloud services, and advanced network technologies has laid the foundation of Cyber-Physical Systems. However, the architecture is highly vulnerable due to weak security bound to IoT devices. The IoT devices have played a significant role in launching DDoS attacks, thereby making the state of affairs catastrophic. The attacks are novel and traditional systems are not efficient enough to address high volume and diverse attacks. The roadblock to researchers is the unavailability of an adequate number of datasets that apprehends the diversity of attacks. The paper describes the specifications of cyber-physical systems, description of IoT-DDoS attacks, and highlights the research challenges. The paper further collates and elucidates the IoT-DDoS datasets with a comparative analysis of the test environment, features offered, and statistical details of datasets. At the end of the paper, based on the dataset study, recommendations are proposed for dataset standardization and extracting most out of the datasets.

Keywords— *Internet of Things; IoT-DDoS; Distributed Denial of Service; Dataset; Cyber Physical System*

I. INTRODUCTION

Internet of Things (IoT) and Cloud Computing are the keys in the electronic cosmos of connected living. The IoT devices are consistently ideated, innovated, and created for novel disciplines. The companion cloud computing paradigm offers software, platform, and infrastructure as services and is under high adoption because of its ability to offer on-demand services, broad network coverage, resource pooling, rapid elasticity, and measured services. It is predicted that predicted number of connected IoT devices will grow to 75.44 billion by 2025 from 42.62 billion in 2022 [1]. Gartner, a leading research company, has forecasted the cloud market to grow to 3.64 billion US dollars by 2022 [2] and COVID-19 has proven catalyst to the cloud growth.

The unprecedented growth of Internet of Things, transformation of communication layer, and cloud computing services amalgamated on a time-series to coined the term Cyber-Physical Systems (CPS).

In a cyber-physical system, a swarm of IoT devices and cloud services are deeply intertwined and are supported by a networking backbone to operate on different temporal and spatial scales. Fig. 1 shows Smart City as an illustration of the Cyber-Physical System. The US National Science Foundation

has perceived cyber-physical systems as a substantive realm of research. The key soup ingredients of the contemporary cyber-physical system are the physical sensors and actuators that are fabricated onto communicable and intelligent embedded devices, the dimensions of automation, integration at Spatiotemporal scales, dynamic configuration, and networking at divergent planes.

The proliferation of IoT systems and innovative cyber-physical systems have made society so much dependent on the digital epoch that even a dearth of any of the operational capability may lead to severe consequences. Naturally, any technological innovation that is widely embraced, allures the cyber attackers to exploit the implementation domain by employing advanced tools and techniques such as Botnets. The lack of uniformity in low-cost, low-power, and lightweight IoT devices has exacerbated the problem. There are shreds of evidence that attackers have exploited the IoT system fragility for botnet propagation to launch an unparalleled IoT-DDoS attack of magnitude 2.3 Tbps in 2020, which is the peak recorded in the last 10 years [3]. As manifested, the attackers can compromise IoT devices, snap out of it, orchestrate a zombie army, launch an IoT-based DDoS attack, and cause substantial damage to the cyber-physical systems or its components (such as IoT network, communication backbone, and cloud network) in atomicity. The IoT-DDoS attacks are geared towards disrupting the services of the server by aiming at the availability vertex of CIA (Confidentiality, Integrity, and Availability) triad.

The researchers have been investing substantial efforts in envisioning and devising an effective defense mechanism against IoT-DDoS attacks that can analyses such an enormous amount of data in real-time. In our previous research [4], [5], [6] we have presented the vulnerability retrospection of IoT-based DDoS attacks on cyber-physical systems. Though contemporary mathematical model-based machine learning and deep learning techniques are applied for an intelligent framework/solution, the major roadblock is the availability of benchmark datasets for research purposes. The IoT-DDoS attacks are novel, and the research is in inception.

A. The relevance of IoT-based DDoS datasets to the Cyber-Physical Systems

IoT devices are the key ingredients of a cyber-physical system. IoT devices form the perception layer and are the only layer where standardization has not happened. The reasons are perceptible that the devices are diverse, heterogeneous, fragile

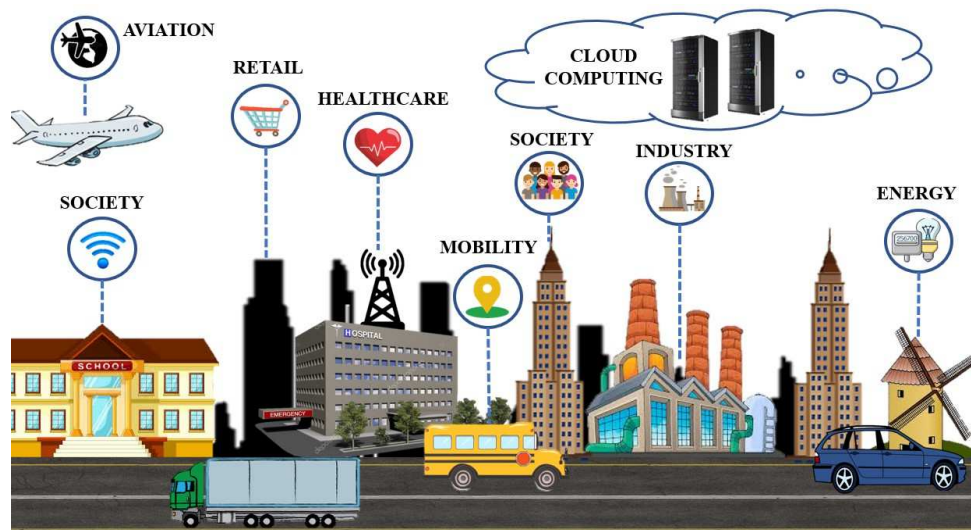


Fig. 1. Smart City - An example of Cyber-Physical System

to security issues. Though there are no direct datasets available for cyber-physical systems, IoT-based datasets make the closest relevance to the cyber-physical systems. Numerous datasets are available for the cloud layer with the detailed available literature, but no such literature is available for IoT-based attack datasets. Hence, the discussed IoT-based attack datasets are the missing piece to the jigsaw puzzle.

B. Our Contribution

The article has the following prime contributions to the research community: To our perception and awareness, the research community lacks the IoT-DDoS datasets because IoT-DDoS are at the inception stage. The paper offers comprehensive details of IoT-DDoS attacks datasets. The paper also presents the comprehensive challenges of IoT-DDoS attacks on the cyber-physical system. Furthermore, the paper proposes recommendations for dataset standardization.

C. Paper Organization

Rest of the paper is structured as follows: Section II puts forward the details of cyber-physical system, security issues with cyber-physical systems, most devastating IoT-DDoS attacks, and research challenges. Section III discusses about the details of publicly available IoT-DDoS datasets, dataset features for the purpose of preparing defense solution, and a comparative analysis of various datasets under discussion. Section IV highlights future prospects of the work. Section V concludes the study.

II. CYBER-PHYSICAL SYSTEM SECURITY

The Cyber-Physical System overview, security issues, and research challenges are described in the following sections.

A. Cyber-Physical Systems

The unprecedented growth of physical objects linked to the Internet has proven a boon for implementing the Internet of Things (IoT) concept. While developments such as the Internet of Things can be traced in the early 80s by connecting the Carnegie Mellon University Coca Cola vending machine to the internet. Likewise, cloud computing had been around

since 2000. A correlation between the real world (e.g. automobiles, wearable technology, medical instruments, sensor elements, etc.) and digital world has proven to be a boon to innovation in network domains. Interdependency and union of the Internet of Things (that represents the physical space) and cloud computing (that represents cyberspace) have given birth to an era of Cyber-Physical Systems (CPS).

However, the popularization of IoT, Cloud, and CPS has proven a battlefield for a new and more devastating form of DDoS attacks, i.e. IoT-DDoS [7]. The following section discusses more about IoT-DDoS attacks.

B. IoT-DDoS Attacks

The IoT-DDoS attacks frequently require millions of compromised IoT devices to exploit huge resources against a target. The Fig. 2 shows the modus operandi of IoT-based DDoS attacks. The recent high-volume DDoS attacks (recently reported 2.3 Tbps in Mar-2020 on U.S. Health Agency) from the millions of diverse and heterogeneous IoT devices has turned out that the traditional defense solutions are not sufficient enough to handle the modern-day IoT traffic [8]. The best known Mirai form of IoT-DDoS has evolved from compromising the MIPS based Linux devices to MIPS, ARM, PPC, and many-more CPU architectures. In its simplest form, IoT-based DDoS attacks are composed to have four primary elements. The core of the botnet is Command and Control Server (C&C Server), which provides the centralized management, tools, and commands to drive the IoT-DDoS attack. The supporting component is the loader system that diffuses the bot binary (Based on target CPU architecture) on the victimized device. Report Server keeps the specifics of the botnet army in the database. The Scanner and the Bot Dyad have the dual objective of spreading the infection to feebly configured devices and attacking the server once the attack command is received from the C&C server.

C. Research Challenges

The IoT devices were introduced only a decade ago and have been adopted by public in large in a short span of time.



Fig. 2. Modus operandi - IoT-DDoS

This popularity has, no doubt, attracted researcher's attention. The flip side of the coin is, attackers have also shown great interest in IoT devices and have leveraged IoT devices to prepare a zombie device army due to weak security of IoT device. The IoT devices have greatly been used to launch the most devastating security attack, IoT-based DDoS. Although, a very significant amount has been invested by research community, there are a number of research challenges faced by the community. The research challenges are highlighted as below:

1) *Unavailability of IoT-based Datasets*

The major roadblock for the researchers is unavailability of IoT-based datasets. As IoT devices are still under evolution, only a limited dataset is available for the research community. To the best of our knowledge, the publicly available IoT-dataset is still in scattered form, and no documentation is available for comparative analysis for the research purpose.

2) *High Volume Network Traffic*

With the dawn and exponential increase in IoT devices, there is a sky-scraping increase in network traffic. The traditional security solutions are not efficient enough to address the modern-day IoT traffic [9]. There is a strong need for the researchers to pay the primary attention to the defense solutions than spending significant efforts in locating an appropriate dataset to test with.

3) *Diversity in datasets*

As IoT devices are heterogeneous in nature, no dataset alone is enough to support a defense solution because of unavailability to capture the diverse scenarios. Hence, there is a strong need to collate the datasets available in IoT domain.

4) *Dataset Outlook*

The rapidly advancing technologies demand a shift in the way datasets are created. The available datasets don't address the surfacing technologies such as Big Data, Blockchain, Software-defined Networking (SDN), Network Function Virtualization (NFV), and the associated threats.

5) *Zero-day Attacks*

The attacks are evolving at a higher pace than the datasets are being generated. Hence, new techniques are required for dataset generation and benchmarking the datasets.

6) *Security-as-a-Service*

The cloud computing services have introduced the concept of availability of Everything-As-A-Service. For example, infrastructure, platform, software, and even machine learning and IoT platforms are available as services. With the technology focus on making everything realized as a service, the work towards realizing security-as-a-service is still in inception. Security-as-a-service can be an exceptional platform to test the available datasets thoroughly [10].

D. *Research Motivation*

To our acquaintance, there has been no research article that details contemporary traffic and datasets. The lack of information on modern-day IoT datasets poses a significant roadblock for researchers. Hence, we have summed up the article that offers in-depth details of several current datasets, the list of attacks covered, a description of the benign and attacks instances, etc. The benchmark datasets details will clear the roadblocks so that researchers can focus on the security solutions rather than simulate the scenarios.

III. BACKGROUND AND RELATED WORK

There is a decent effort in attack detection and mitigation in CPS and IoT systems. Karthick et al. [11] manifested a Fog layer-based solution which includes data processing, storage, and security, in a proposal for a Medical Cyber System (MCS) and uses IoT technologies for cardiovascular disease patients.

There are several benchmark datasets (such as NSL-KDD, CICIDS2017, CICDDoS2019, etc.) and detailed literature available for traditional attack scenarios. However, only limited datasets are available for IoT-DDoS attacks. Most of the datasets captures the real time traffic and are explained in the following section.

A. *IoT-DDoS Dataset Specifics*

The details of publicly available IoT-DDoS datasets are described as follows. Table 1 summarizes the statistical attributes of the contemporary IoT-DDoS datasets.

1) *IoT-23 Datasets*

IoT-23 dataset [12] dataset is a benchmark labelled IoT-dataset created by Stratosphere Laboratory, Avast AIC in association with Czech Technical University (CTU), Czech Republic with an aim to contribute to research community. The dataset is a recently published dataset in 2020, with captures extending from 2018 to 2019. The dataset consists of 23 scenarios (hence, named IoT-23) i.e., 20 malware captures from infected IoT devices, and 3 captures from benign IoT-devices. The scenarios were created after compromising the devices by Mirai, Torii, Trojan, Gagfyt, Kenjiro, Okiru, Hakai, IRCBot, Muhstik, and Hide & Seek malwares. The network traffic flows in the datasets have further been labelled to describe the possible malicious activity. The authors have categorized the traffic into attack traffic, benign traffic, C&C traffic, and file download traffic. When the traffic flows from the infected device to the target host, the traffic is attack traffic. The traffic is benign if no suspicious activity exists in the flow. The network communication between C&C and the infected device is the

C&C traffic. Furthermore, when the file is downloaded to the infected device, it produces the file download traffic. Moreover, the packets sent on a specific connection to check KeepAlive of the infected device by the C&C server are classified as HeartBeat traffic. The goal of the dataset is to prepare a dataset for malicious traffic and another for benign network traffic.

2) LITNET-2020 Dataset

The LITNET-2020 Netflow dataset [13] is a benchmark dataset created from real world academic research environment containing the attack as well as benign traffic. The dataset presents 12 attack types. The dataset was created

at Kaunas University of Technology (KTU), Lithuania and have captures ranging from 2019 to 2020. The network traffic was captured using nfcap tool and feature set was generated using Python Scripts, Nfsen, and MeSequel tools. The network topology used for collecting LITNET traffic consisted of senders and receivers. The CISCO routers and Fortige firewalls (FG-1500D) have been used as senders and high configuration server is used as collector. The malicious traffic is further classified into 12 classes based on the attack type.

TABLE 1. IoT DDoS DATASETS COMPARISON

Dataset	Year	Num Instances	Malicious Instances	Benign Instances	Num Attributes	Num IoT Devices	Attack Types / Protocols for Attack	Size of Dataset (GB)
IoT-23	2020	32,53,07,990	29,44,49,225	3,08,58,765	--	--	HTTP, DNS, DHCP, SSL, SSH, and IRC	21
LITNET-2020	2020	4,53,30,333	53,28,934	4,00,01,399	85	--	Smurf Attacks, Flooding ICMP, UDP, SYN, HTTP, W32.Blaster, SPAM, Reaper Worm, Scan, Code Red, Fragmentation, and LAND.	1.15
BOT-IoT	2019	73,360,900	7,33,51,357	9,543	--	5 (Simulated)	Port Scanning, OS Fingerprinting, Dos/DDoS (HTTP, TCP, UDP), Metasploit (Data Theft, Key Logging) .	69.3
Sydney UNSW IoT	2019	Raw Data (.pcap files)	Raw Data (.pcap files)	Raw Data (.pcap files)	Raw Data (.pcap files)	28	DDoS	80.3
N-BaIoT	2018	70,62,606	65,06,674	5,55,932	115	9	BASHLITE (Scan, UDP, TCP, COMBO), Mirai (Scan, ACK, SYN, UDP, UDPPlain)	34
CTU-13	2011	2,06,43,076	4,32,755	3,69,806	22	Simulated	IRC, SPAM, ClickFraud, Port Scan, DDoS, Fast Flux, HTTP.	696.9

3) BOT-IoT

Unlike other traditional available benchmark datasets, Bot-IoT dataset [14] was created specifically for IoT-Botnet scenarios. The dataset was generated in Cyber Range Lab of UNSW Canberra Cyber, Australia. The authors performed feature selection is performed using Correlation and Entropy [15] methods. The dataset is classified into normal network traffic, probing attacks (Port Scanning, OS Fingerprint), DoS attacks (HTTP, TCP, UDP), DDOS attacks (HTTP, TCP, UDP), and Information Theft attacks (Key logging, Data theft).

4) Sydney UNSW IoT Dataset

The researchers from University of New South Wales (UNSW), Sydney [16] instrumented 28 unique IoT devices to collect the traffic traces spanning 26 weeks. The stage is set for further analysis and defense solutions against IoT-DDoS attacks.

5) N-BaIoT Dataset

The N-BaIoT [17] fills the gap in the publicly available botnet datasets, specifically for IoT. The data was collated from 9 commercially produced IoT devices that have been infected with Mirai and BASHLITE. in the controlled experimental setup and is multi-classified into 10 classes of

benign data and 1 category of benign data. The authors of the dataset originally used anomaly detection technique. They applied Deep Autoencoders and concluded with a 100% TPR rate.

6) CTU-13 Dataset

The CTU-13 dataset [18] was captured and published at Czech Technical University (CTU), Czech Republic in 2011. The dataset consisted of 13 scenarios, each one executed with different malware. A variety of bots such as neris, Rbot, Virut, Menti, Sogou, Murlo, and NSIS.ay were used to generate the attack traffic. CTU-13 is as old as hills and the most recent version, IOT-23, has been published by the laboratory.

B. IoT-DDoS Dataset Specifics

IoT traffic behavior is the key to distinguish the IoT devices contrast from non-IoT devices [19]. The IoT-device specific features are flow volume, flow duration, sleep time (time duration for which IoT device doesn't have active flow), etc. [16].

The characteristics possessed by an individual IoT device factors in concluding the association of generated traffic to specifics of network communication. IoT devices, despite

their heterogeneous purposes of existence, generate a predefined traffic pattern that can be used to generate clusters of IoT devices based on network traffic characteristics. Such features are used for IoT device identification [16], type of service being used by IoT devices [20], unauthorized device detection in the network [21], and detection of traffic anomalies [17][22]. The network traffic features generated by IoT devices are classified into following two categories:

1) Network traffic features at packet level

Notably, the features used in IoT traffic study, network intrusion detection systems, DDoS defense solution, or any

network security solution is comprised of network packet level features of the traffic transaction. The most commonly used network traffic features of IoT devices at network packet level are described in Table 2. The first five attributes form the primary tuple of a flow, i.e. $\langle IP.src, IP.dst, IP.proto, port.src, port.dst \rangle$. Other least commonly used packet level features are: *pkSeqID* (Packet Sequence Number), *Stime* (Flow start time), *flgs* (Flow state flags in transaction), *bytes* (Total number of bytes in a transaction) [23]. The network packet features are the prime keys to DDoS detection.

TABLE 2. NETWORK TRAFFIC PACKET LEVEL FEATURES OF IOT-TRAFFIC

Feature Label	Feature	Feature Description	Research References
IP.src	Source IP address	The IP address of the device originating the network traffic.	[23] [24] [17] [25] [26] [27]
IP.dst	Destination IP Address	The IP address of the destination to which the network traffic is targeted.	[23] [24] [25] [26][27]
IP.proto	Protocol for communication	The protocol used for communication between two entities.	[23] [24] [17] [25] [26] [27][28]
port.src	Source port	The port number of the source device.	[23] [17] [25]
port.dst	Destination Port	The port number of destination where the sought service is available.	[23] [24] [25] [29]
Ttl	Time to live	Time to live for the packet for preventing the packet from circulating in the network indefinitely.	[21]
psize	Packet Size	Size of each packet in the flow.	[23] [17] [25] [26]
Num_Pkts	Total Packets exchanged	Total number of packets exchanged during a flow	[23] [17]

TABLE 3. NETWORK TRAFFIC FLOW LEVEL FEATURES OF IOT-TRAFFIC

Feature Label	Feature	Feature Description	Research References
intensity	Traffic Intensity	Sum total of transmitted and received packets during a flow.	[24] [16]
S_dur	Session Duration	The length of duration during which session is established between two end points during a flow.	[24] [16]
Sleep_time	Device inactive time	Time duration between first and last packet during a flow	[24] [16][6]
Flow_vol	Flow Volume	Sum total of size of all incoming and outgoing packets during a flow	[24] [16]
Avg_flow_rate	Average Flow rate	Ratio of Flow Volume and Flow Duration	[24] [16]
Pkt_size_S	Additional Statistical Packet Size related Feature	Additional features added as a result of performing statistical analysis, such as mean, standard deviation and maximum values of traffic flow	[17] [30] [31]
IP.proto	Protocol for communication	The protocol used for communication between two entities	[23] [24] [16][17] [25] [26][27] [28]
Num_packets	Number of packets	Number of packets during a flow	[23][17]
Int	Packet inter arrival time a.k.a. packet jitter	Time difference between two consecutive packets arrival in a flow	[17]
DNS Query	DNS Query by IoT	DNS Query by IoT devices	[24] [16]
NTP Queries	NTP Query by IoT	NTP Query by IoT devices	[24] [16]
Cipher Suite	Cipher Suite for data encryption	Cipher suite used by IoT during secure communication	[24] [16]

2) Network traffic features at flow level

The selection of features is hand-picked based on the purpose of the study. For example, The authors of [23] inferred from the experimental results that the IoT devices exchange low volume of data per flow. They empirically

proved that 95% of the IoT devices generate less than 1000 bytes of data, with most devices transferring as low as 120 bytes per flow. The aforesaid researchers used other flow level attributes such as flow duration, average flow rate, and

inactive (sleep) pattern of IoT devices. The features identified at network flow level are described in Table 3.

3) IoT Dataset Comparison

The Fig. 3 summarizes the dataset and the type of attack covered in the dataset. The datasets from various researchers and laboratories have been published and are available for further study. In this section, we discuss the contrast between

the testbeds used for generating the datasets in the literature. A testbed is an experimental setup used for generating the network traffic. The comparative analysis of experimental testbeds used for generating IoT traces are shown in Table 4 as follows:

TABLE 4. COMPARATIVE ANALYSIS OF ENVIRONMENT FEATURES OF DATASETS

Dataset	Testbed Configuration (R: Realistic/ S: Simulated)	Traffic Type (R: Realistic/ S: Simulated)	Labelled Dataset (Y/N)	IoT Trace (Y/N)	Diverse Attacks Scenarios (Y/N)	Full Packet Network Capture (Y/N)
IoT-23	R	R	Y	Y	Y	Y
LITNET-2020	R	R	Y	N	Y	Y
BOT-IoT	S	S	Y	Y	Y	N
Sydney UNSW IoT	R	R	N	Y	N	Y
N-BaIoT	R	R	Y	Y	Y	Y
CTU-13	S	S	Y	Y	Y	Y

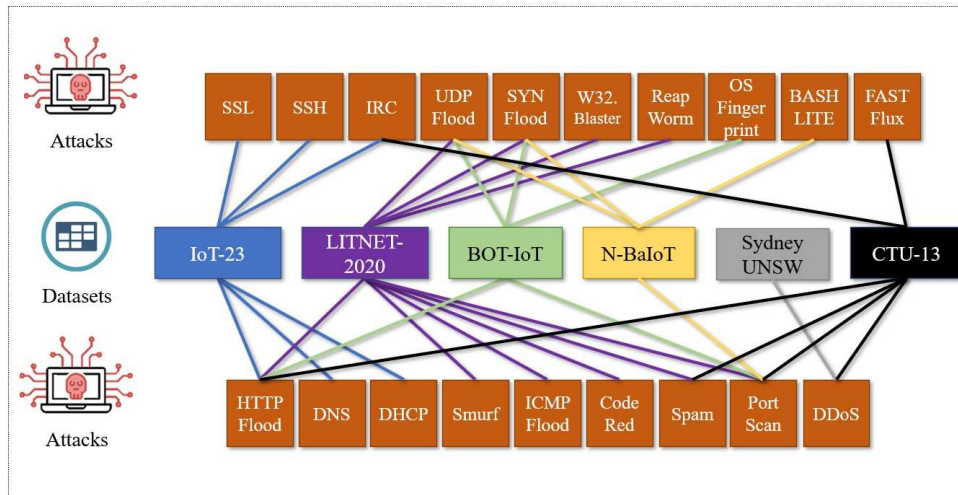


Fig. 3. Datasets and Attacks mapping

IV. FUTURE PROSPECTS

The following suggestions would help develop the next-generation cyber security framework for intelligent devices, driven by the strong reliance of datasets on the development of IoT-DDoS defense solutions. The outcome would be real-life deployments.

A. Generation of extensible and modular datasets

As aforesaid, diverse datasets are highly required to cover complex network architectures or to put forward zero-day attacks. To benefit the most from the datasets, they should be flexible to extend or to integrate with other datasets. This would make datasets agile and ready for frequently changing scenarios. That way, anomaly-based DDoS defense solutions could be trained for advanced machine and deep learning techniques to identify zero-day attacks.

B. Standardization of dataset generation, collection, evaluation and validation processes

One of the prime challenges faced by researchers is the lack of standard procedures for dataset creation and unavailability of detailed documentation associated with newly available datasets. Furthermore, the lack of benchmark tools and document templates makes the researchers use their own tailored methods. The datasets should be validated on the standard parameters and performance metrics. Large degree of researchers focuses on showcasing the accuracy of the defense solution/model/framework. However, the performance metrics, other than accuracy, should be considered to reflect the precise performance of the proposed model

C. Datasets resilience

To ensure dataset resilience, the datasets variations should be available. This may include variations in attack scenarios, diversity in traffic loads (daytime and inactive time). The datasets availability in raw format along with the standardized template format would make researchers apply their own tailored methods and have a choice between stateless or stateful features.

D. Close Collaboration

The study shows that there is no common platform used for dataset generation. Each of the generated datasets exist in isolation, and the community is continuously investing efforts in developing the modern-day datasets. The shared platform will also help in dataset standardization.

E. Dataset Publication Platform

We also recommend using a common standardized platform for datasets and the associated metadata for analysis. The research community can leverage the platform to focus on solution-oriented approaches rather than setting up the data generation environment.

V. FUTURE PROSPECTS AND CONCLUSION

This paper described the cyber-physical system and investigated the security issues associated with contemporary IoT-devices that form the perception layer of cyber-physical system. The paper also discusses IoT-DDoS as the most devastating version of DDoS attacks and then highlights the unavailability of IoT-datasets as a challenge to the research community. The available IoT-DDoS datasets are well collated in the paper which will serve as the base document for the researchers. Finally, the paper concludes that the lack of standardization of datasets are the roadblock to study and proposes the recommendation to the dataset standardization process.

ACKNOWLEDGMENT

The authors would like to pass-on gratitude to anonymous reviewers for their constructive feedback and valuable suggestions.

REFERENCES

[1] S. Rose, "Top 14 IoT Trends to Expect in 2020!" <https://towardsdatascience.com/top-14-iot-trends-to-expect-in-2020-fa81a56e8653> (accessed Oct. 29, 2020).

[2] STAMFORD, "Gartner Forecasts Worldwide Public Cloud Revenue to Grow 6.3% in 2020." <https://www.gartner.com/en/newsroom/press-releases/2020-07-23-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-6point3-percent-in-2020> (accessed Oct. 29, 2020).

[3] E. TARGETT, "Record DDoS Attack Hits AWS: 2.3 Tbps Assault Lasted Days." <https://www.cbronline.com/news/record-ddos-attacks> (accessed Oct. 27, 2020).

[4] M. Snehi and A. Bhandari, "Vulnerability retrospection of security solutions for software-defined Cyber-Physical System against DDoS and IoT-DDoS attacks," *Computer Science Review*, vol. 40, p. 100371, 2021, doi: 10.1016/j.cosrev.2021.100371.

[5] M. Snehi and A. Bhandari, "Apprehending Mirai Botnet Philosophy and Smart Learning Models for IoT-DDoS Detection," in 2021 8th International Conference on Computing for Sustainable Global

Development (INDIACom), 2021, pp. 501--505, doi: 10.1109/INDIACom51348.2021.00089.

[6] M. Snehi and A. Bhandari, "A Novel Distributed Stack Ensembled Meta-Learning-Based Optimized Classification Framework for Real-time Prolific IoT Traffic Streams," *Arabian Journal for Science and Engineering*, 2022, doi: 10.1007/s13369-021-06472-z.

[7] J. Snehi, A. Bhandari, M. Snehi, and V. Baggan, "Diverse Methods for Signature based Intrusion Detection Schemes Adopted," *International Journal of Recent Technology and Engineering*, vol. 9, no. 2, pp. 44-49, Jul. 2020, doi: 10.35940/ijrte.A2791.079220.

[8] J. Verma, A. Bhandari, and G. Singh, "A Meta-analysis of Role of Network Intrusion Detection Systems in Confronting Network Attacks," in 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 2021, pp. 506-511, doi: 10.1109/INDIACom51348.2021.00090.

[9] J. Snehi, M. Snehi, A. Bhandari, V. Baggan, and R. Ahuja, "Introspecting Intrusion Detection Systems in Dealing with Security Concerns in Cloud Environment," in 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), Dec. 2021, pp. 345-349, doi: 10.1109/SMART52563.2021.9676258.

[10] M. Snehi and A. Bhandari, "An SDN/NFV based Intelligent Fog Architecture for DDoS Defense in Cyber Physical Systems," in 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), Dec. 2021, pp. 229-234, doi: 10.1109/SMART52563.2021.9676241.

[11] T. Karthick and M. Manikandan, "Fog assisted IoT based medical cyber system for cardiovascular diseases affected patients," *Concurrency Computation*, vol. 31, no. 12, pp. 1-9, 2019, doi: 10.1002/cpe.4861.

[12] Stratosphere, "IoT-23 Dataset: A labeled dataset of Malware and Benign IoT Traffic - Stratosphere Laboratory Datasets." <https://www.stratosphereips.org/datasets-iot23> (accessed Oct. 31, 2020).

[13] "LITNET-2020: an annotated real-world network flows dataset for network intrusion detection," 2020. <https://dataset.litnet.lt/> (accessed Oct. 31, 2020).

[14] M. A. Lawal, R. A. Shaikh, and S. R. Hassan, "An anomaly mitigation framework for iot using fog computing," *Electronics (Switzerland)*, vol. 9, no. 10, pp. 1-24, 2020, doi: 10.3390/electronics9101565.

[15] A. Lesne and H. Etudes, "crossroads between probability , information theory , Shannon entropy: a rigorous mathematical notion at the crossroads between probability , information theory , dynamical systems and statistical physics," *Science*, 2011.

[16] A. Sivanathan et al., "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1745-1759, Aug. 2019, doi: 10.1109/TMC.2018.2866249.

[17] Y. Meidan et al., "N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12-22, 2018, doi: 10.1109/MPRV.2018.03367731.

[18] S. Garcia, M. Grill, J. Stiborek, and A. Zunino, "The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background traffic - Stratosphere IPS," 2011. <https://www.stratosphereips.org/datasets-ctu13> (accessed Nov. 01, 2020).

[19] B. Finley, J. Benseny, A. Vesselkov, and J. Walia, "How does enterprise IoT traffic evolve? Real-world evidence from a Finnish operator," *Internet of Things*, vol. 12, p. 100294, 2020, doi: 10.1016/j.iot.2020.100294.

[20] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network Traffic Classifier with Convolutional and Recurrent Neural Networks for Internet of Things," *IEEE Access*, vol. 5, pp. 18042-18050, 2017, doi: 10.1109/ACCESS.2017.2747560.

[21] Y. Meidan et al., "Detection of Unauthorized IoT Devices Using Machine Learning Techniques," 2017, [Online]. Available: <http://arxiv.org/abs/1709.04647>.

[22] I. Cvitić, D. Peraković, M. Periša, and M. Botica, "Novel approach for detection of IoT generated DDoS traffic," *Wireless Networks*, vol. 1, 2019, doi: 10.1007/s11276-019-02043-1.

- [23] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019, doi: 10.1016/j.future.2019.05.041.
- [24] A. Sivanathan et al., "Characterizing and classifying IoT traffic in smart cities and campuses," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPs)*, May 2017, pp. 559–564, doi: 10.1109/INFOCOMW.2017.8116438.
- [25] A. Kumar, M. Shridhar, S. Swaminathan, and T. J. Lim, "Machine Learning-Based Early Detection of IoT Botnets Using Network-Edge Traffic," 2019.
- [26] R. Doshi, N. Apthorpe, and N. Feamster, "Machine learning DDoS detection for consumer internet of things devices," *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, no. M1, pp. 29–35, 2018, doi: 10.1109/SPW.2018.00013.
- [27] M. S. Elsayed, N.-A. Le-Khac, and A. D. Jurcut, "InSDN: A Novel SDN Intrusion Dataset," *IEEE Access*, vol. 8, pp. 165263–165284, 2020, doi: 10.1109/access.2020.3022633.
- [28] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, and I. Ray, "Behavioral fingerprinting of IoT devices," *Proceedings of the ACM Conference on Computer and Communications Security*, no. October, pp. 41–50, 2018, doi: 10.1145/3266444.3266452.
- [29] A. Sivanathan, H. H. Gharakheili, and V. Sivaraman, "Can We Classify an IoT Device using TCP Port Scan?," *2018 IEEE 9th International Conference on Information and Automation for Sustainability, ICIAfS 2018*, 2018, doi: 10.1109/ICIAfS.2018.8913346.
- [30] L. Bai, L. Yao, S. S. Kanhere, X. Wang, and Z. Yang, "Automatic Device Classification from Network Traffic Streams of Internet of Things," *Proceedings - Conference on Local Computer Networks, LCN*, vol. 2018-Octob, pp. 597–605, 2019, doi: 10.1109/LCN.2018.8638232.
- [31] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A. R. Sadeghi, and S. Tarkoma, "IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT," *Proceedings - International Conference on Distributed Computing Systems*, pp. 2177–2184, 2017, doi: 10.1109/ICDCS.2017.283.

Resilience Evaluation of Cyber Risks in Industrial Internet of Things

1st Mayer Silva
 Technology Managent Department
 SENAI CIMATEC
 Salvador, Brazil
 mayer_eq@hotmail.com

2nd Herman Lepikson
 Technology Management Department
 SENAI CIMATEC
 Salvador, Brazil
 herman.lepikson@fieb.org.br

Abstract—Wireless communication is one of the most implemented solutions, allowing a high level of interconnectivity in different networks, including industrial critical systems responsible for production control and safety. Performance improvements with the increase in wireless connectivity between "things", however, are followed by the increase in vulnerabilities due to the risks related to the security of wireless networks. Thus, this study aims to provide a view of the wireless communication vulnerabilities in the context of applications for industrial environments, assess maturity of current technologies in critical systems and provide a map of resilience measures to be adopted to anticipate and mitigate risks.

Keywords—Industrial Internet of Things, Cyber Risks, Resilience, Digital Manufacturing, Wireless Hart, ISA100

I. INTRODUCTION

In a digital manufacturing era, new communication and control infrastructures are operating in cooperation and collaboration for optimized performance of process, equipment and devices [1]. The increasing in the systems interoperability with integration of IoT, cloud computing and other emerging technologies generated a hyperconnection of systems where it is imperative to define a role for cybersecurity area where risks and vulnerabilities should be considered and addressed throughout systems lifecycle [2]. In December 2010, a new worm (type of malware) called Stuxnet targeted a highly specialized industrial system with unprecedented sophistication and impacts, making this event a key milestone for cybersecurity in industrial systems. Due to the potential impacts, this event was considered the wake-up call for the industrial cybersecurity awareness [3].

II. INDUSTRIAL INTERNET OF THINGS SECURITY

A. IIoT Standards and Protocols

Wireless devices needs low power consumption to be integrated into IoT and WSN (wireless sensor networks) with hundreds or thousands of other devices. Typically, a wireless communication solution is defined based on the data transfer rate, required signal range, security and reliability required for the network [4]. As part of the IEEE 802 communication standards, the IEEE802.15 group of short-distance Wireless Personal Area Networks (WPAN) stands out. The WPAN working group was initially created as IEEE 802.15.1 standards for Physical Access and Medium Access Control (MAC) Layers based on Bluetooth technology. In 1999, the IEEE 802.15.3 WPAN group was included and in 2000 the IEEE 802.15.4 WPAN was introduced for low transmission rates [5].

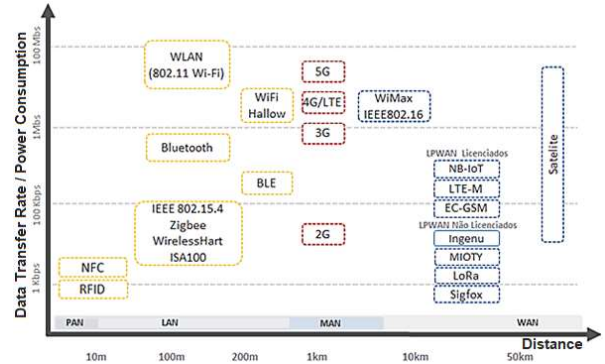


Fig. 1. Wireless Protocols Characteristics

IIoT networks might be affected by a variety of threats and vulnerabilities where attackers may take advantage of different mechanisms to cause interference or to intercept a communication. Industrial environments are also composed by a diversity of physical areas that adds important functional challenges for IIoT devices. As described in Figure 2, according to [11] wireless networks can be characterized as layered communication protocols following the Open System Interconnection (OSI) model. The communication stack consists of main 5 layers: Physical, MAC (Data-Link), Network Layer, Transport and Application.

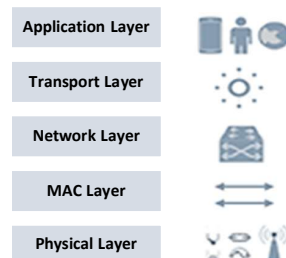


Fig. 2. Wireless Network OSI Stack Layers

B. Systems Resilience

Although widely discussed in other areas of computer science, in the context of CPS (Cyber-Physical System) there are opportunities to discuss resilience mechanisms that can be applied for industrial systems, contributing to adapt to challenging scenarios and maintaining long-term reliability in interoperable and diverse systems. Considering the heterogeneity of equipment in these networks, reliability and security characteristics are essential for building an overview

of the state of the art of resilience of IoT in industrial environments [7].

For industrial services it is expected that components of a system are reliable and safe throughout system lifecycle, ensuring robustness for continuous operation even in the presence of failures of any of its components. [6] defines the dependability property of a system as the combination of some attributes such as availability (readiness for correct service), reliability (continuity of correct service), safety (absence of severe incidents), integrity (absence of undue changes) and maintainability (modifications and repairs with minimal effort). Availability, integrity and confidentiality (absence of unauthorized disclosures) are treated as features included in the sense of security.

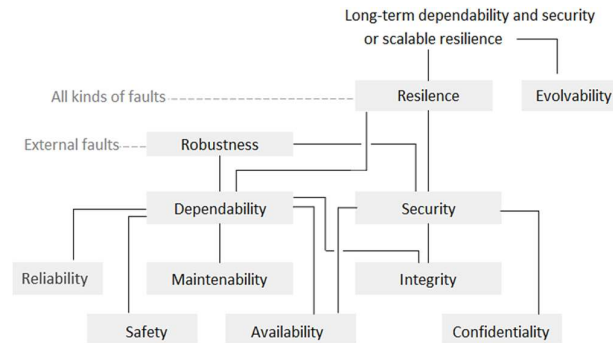


Fig. 3. System Attributes Relations in a Resilience Context [7]

The term resilience is often used by the security community to describe resistance to malicious attacks and crashes. In this context, two important properties contribute to cyber resilience and address security in different ways: safety and security [7]. [8] refers to safety as the property of a system to maintain the physical integrity of people and structures, in order to avoid damage, dangers and various risks. Resilient systems are characterized by having safeguards and security measures as part of the system architecture, with the ability to resist to attacks and failures, maintaining their operation even in a degraded or weakened state [9].

A system with a high level of security must comply with security standards and have failsafe mechanisms. The reliability and robustness of a system is also linked to its property relative to fault tolerance, that is, its ability to maintain its service continuously without significant impacts even in a scenario of device and components failures. The term resilience is often used by the security community to describe resistance to malicious attacks and crashes [7].

III. METHODOLOGY

This study was carried out by identifying security parameters on IIoT networks, including communication protocol technologies and risk management strategies important for a secure operation. This was initially performed through a systematic literature review in four main areas: (i) IoT Communication Protocols, (ii) IIoT vulnerabilities and attacks, (iii) IIoT Security and resilience measures, (iv) cyber risk management.

Literature findings were qualitatively explored and analyzed to create a map of vulnerabilities and attacks mechanisms to IoT networks. These vulnerabilities and challenges found were individually identified, classified,

prioritized, and treated to find actions to prevent or mitigate incidents in production systems in a digital manufacturing environment. These prior steps supported the mapping of risks involving the IoT on a huge diversity of scenarios, including industry environment.

Through an in-depth analysis of IoT technologies, applications and literature review, wireless sensor networks were assessed regarding its resilience properties considering all parts of the system, as well as the life cycle. The purpose is to provide a view of vulnerabilities and assess current maturity of IoT application in the industrial environment.

IV. IIoT RESILIENCE EVALUATION

A. Challenges and Cyber Attacks in IoT

The implementation of IoT technologies makes it possible to optimize operations through the application of previously unfeasible monitoring and controls, however it comes with several challenges in keeping thousands of devices coexisting and connected efficiently and securely. Some of the main technical challenges listed in the literature for the implementation of IoT networks are listed below [11]:

- Heterogeneity – IoT consists of a variety of devices belonging to the same network with gateways, switches, sensors, actuators, intelligent applications and mobile systems.
- Scalability – challenge in generating addresses, names, management and services from thousands of connected devices.
- Communication – Various technologies are used by IoT devices and networks, such as wired and wireless communications in different protocols.
- Energy Consumption – is one of the biggest challenges for IoT networks. Any kind of algorithm or mechanism running on IoT devices needs to be designed to operate with low processing load in order to extend life of its battery.
- Data Privacy – User data privacy when using IoT can be an issue in some specific cases. For example, in regular mode IoT devices can provide location information to system administrators or neighboring devices, but when in private mode, information such as this must be kept secret.
- Self-Awareness – IoT smart objects must organize themselves autonomously to perform some predetermined tasks in response to the real environment in which they are inserted, minimizing the need for human intervention.
- Interoperability – standardization of communication between different IoT devices, allowing heterogeneous objects and networks to communicate.

Other challenges related to IoT devices availability [12]:

- Fault Tolerance – The system must prevent an infected or problem node from affecting the entire system. This network must also be able to avoid the resource exhaustion attack against resource-limited devices.

- Availability – IoT devices and networks must be functionally available when needed, maintaining system operational continuity.

[13] demonstrated other challenging attributes for IoT communication:

- Trust Management – Objects and devices cannot communicate without at least having an established trust level. Building trust between devices coming from different entities is a challenge, especially in such a large-scale network.
- Latency and time constraint – objects in IIoT networks need to respond in real time to events and messages. For example, the SCADA system (Supervisory Control and Data Acquisition) must respond in real time to any variation in current, voltage or frequency values of electricity, in addition to other environmental parameter that influence the operation of the equipment.
- Mobility – the use of mobile devices, such as e-cars and mobile operations in the field, demand a continuous need for authentication and secure communication with a changing environment.
- Resource limitation – many IoT devices have restricted resources. Special care must be taken when developing security solutions to make sure that resources will meet the solutions as intended. This makes applying classic security solutions a challenge, especially those based on public-key cryptography.
- Bootstrapping – efficiently boot the thousands of IoT devices with resources required for encryption (cryptographic keys, parameters, functions and cryptographic algorithms, etc.).

Vulnerabilities and threats are important agents to be considered in the management of security risks for manufacturing systems. Security challenges change overtime and requires up-to-date information as a key part for addressing system vulnerabilities [10]. Figure 4 shows main attacks possible in wireless networks that should be considered to evaluate risks for IoT systems.

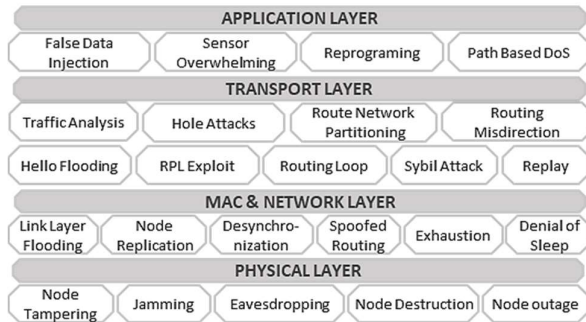


Fig. 4. Wireless Network Attacks [11]

B. IoT Resilience in the Industry Environment

The risk of cyberattacks to the wireless network is one of the most critical technical challenges of implementing this technology on the industrial environment. Vulnerabilities can be exploited by third parties and may cause critical security or safety issues. Passive attacks, which the attacker stays hidden

with no identification of intrusion in the system, normally target sensitive data, silently inserting himself into the network. In active attacks, the target is not only confidential data but also system integrity, causing disturbances to services [11].

In an industrial environment, wireless sensors need to operate securely and safely for a long period embedded in current process control systems. From this perspective, WirelessHART and ISA100 are considered reliable wireless protocols and provide means of path diversity, channel hopping and high reliability in message delivery. Zigbee protocol does not provide same tools to avoid issues with interference and obstacles, usually inherent in industrial environments, which is an issue for this protocol. For this reason, ZigBee is not suitable for critical industrial process applications as it does not meet requirements for industrial-grade network reliability and robustness. On the other hand, safety requirements and energy consumption are satisfactory in this protocol. Table I illustrate industrial WSN resilience for different cyberattack types.

TABLE I. WSN PROTOCOLS RESILIENCE TO ATTACKS

Attack	Layer	W. Hart	ISA100	ZigBee	Source
Eavesdropping	Physical	✓	✓	✗	[16]
Node outage	Physical	✓	✓	✗	[14], [16]
Traffic Analysis	Physical	✓	✓	✓	[17]
Jamming	Physical	✓	✓	✓	[16], [17]
Node Tampering	Physical	✓	✓	✓	[17]
Exhaustion	MAC	✓	✓	✗	[15], [16]
BlackHole	Network	✓	✓	✓	[17]
SinkHole	Network	✓	✗	✗	[14], [19]
Wormhole	Network	✓	✓	✗	[17], [19]
Sybil Attack	Network	✓	✗	✓	[17], [19]
DoS	Application	✓	✓	✗	[18], [19]
Spoofing	Application	✓	✗	✗	[16], [18]
Replay	Network	✓	✗	✗	[16]

✗ - Vulnerability exploitability cited in the literature.
 ✓ - Found demonstration of security for the attack type in the literature.

WirelessHART and ISA100 protocols define a set of security keys that are used to ensure reliable communication. The strong encryption relies on both communication endpoints using the same key to communicate securely. Attackers who do not share keys cannot modify messages undetected and cannot decrypt encrypted payload information. Common to both standards is that a new device is provisioned with a connection key before attempting to connect to a network.

The connection key is used to authenticate the device to a specific network. Once the device successfully connects to the network, the security manager provide keys for further communication. While it is intrinsic in WirelessHART, the use of the join key is optional in ISA100.11a. A global key, a well-known key with no security guarantees, can also be used in the joining process for devices that do not support symmetric keys. ISA100.11a allows optionally encrypting

messages while WirelessHART does not allow security to be optional, which avoids errors that can compromise the system [12]. From general resilience perspective and considering native security in different WSN, Table II shows some resilience attributes for main protocols applied.

TABLE II. WSN PROTOCOLS RESILIENCE ATTRIBUTES

Attribute Resilience	WSN Protocol			Source
	ZigBee	ISA 100.11A	WirelessHart	
Security	High	Very High	Very High	[20]
Reliability	Low	Very High	Very High	[20]
Power Consumption	Medium	Low	Low	[20]
Scalability	Medium	High	High	[20]
Coexistence	Low	High	High	[21]
Channel Hopping	No	Yes	Yes	[21]

At the field level, ISA100.11a devices are not required to support the router function. For this reason, it is possible to see ISA100.11a networks that have a star topology while wirelessHART networks are inherently mesh. The choice for ISA100.11a devices still has a significant influence on the topology that can be deployed. If the user is forced to use a star topology, it is very likely that it will be necessary to perform an on-site test to ensure that the network will be created as designed and performance ensured.

Due to the wide diversity of applications, resilience and maturity of wireless networks, it is necessary to evaluate the types of native protections, additional security devices and usage policies that enable these protocols for safe and reliable use according to the system criticality. In order to identify possible safeguards for ensuring resilience of wireless networks against attacks, the literature findings were compiled and results illustrated in Figure 5.

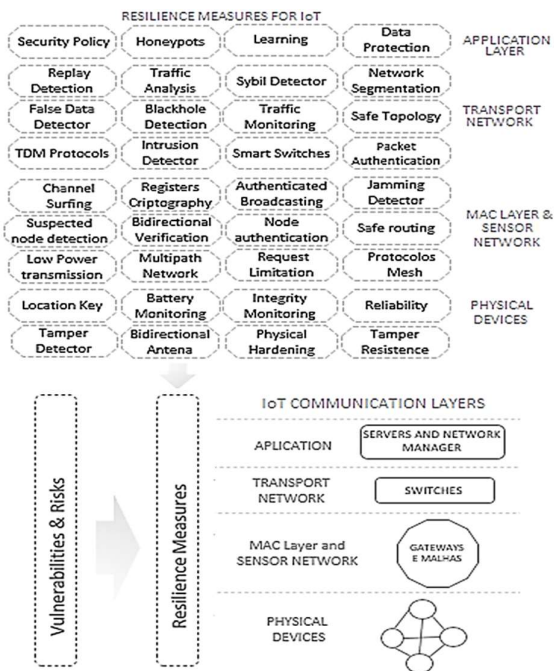


Fig. 5. Wireless Network Resilience Measures

The security and availability of IoT networks is directly linked to the treatment of vulnerabilities and challenges identified during the life cycle of these networks. In industrial environments, there are different types of networks with different criticality. Wireless sensing networks in ICS (Industrial Control Systems) zones must coexist with other technologies and systems in a corporate environment (outside ICS) that also use radio waves to connect to their databases and applications. ICS connected systems are normally critical and, for this reason, they must be implemented using a classification criterion to support resilience analysis required for these networks during their life cycle (accuracy, latency, integrity, and others).

[22] portray that as a vehicle for promoting discussion in the cyber resilience space, the American organization MITRE developed a framework for cyber resilience engineering (Cyber Resilience Engineering framework - CREF). This group is based on resilience engineering disciplines, network resilience, fault-tolerant systems, intrusion resilience in critical infrastructures with the objective of acting in the cyber space also including not cyber events like natural disasters, errors, and failures. Figure 6 illustrates cyber resilience goals, objectives and techniques proposed by MITRE organization.

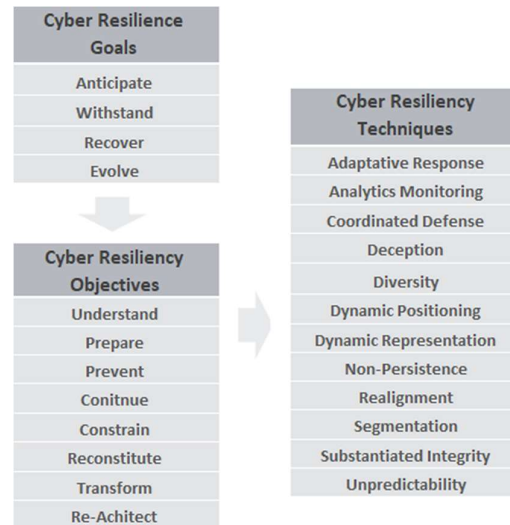


Fig. 6. Cyber Resiliency Engineering Framework [22]

Each wireless communication protocol has its own security features and maturity in terms of reliability and resilience. As shown in Figure 6 in the field of cyber resilience it is needed to apply a series of techniques and plans to ensure proper operation. A cyber risk management is not performed only by system features but as well as security policies, procedures, and culture. However, applying a system with a native secure protocol is still one of the main points for a resilient operation. Unfortunately, although there are protocols with considered high level of security, most of system features are not properly used or even disable during the operation. From the customer and vendors perspective, most of times the efforts during system applications are directed to implement, enable and operate the new wireless system from the functionality perspective and all performance benefits of this system, not specifically about security.

C. Cyber Risk Management

It is recommended that all the security elements for IoT (challenges, threats, vulnerabilities, and resilience techniques) discussed so far should be considered during the system lifecycle to address many types of risks such as ethic risks, technical risks, security risks and privacy risks [23]. To address proper treatment for risks present in system operations it was created important Cyber Security Risks Frameworks (CSRF) with different approach and considerations. Although the main frameworks widely used were not created specifically for IoT application security analysis, these can be used with an approach adjustment for IoT. Table III describes some of main frameworks used for system risk analysis.

TABLE III. MAIN CYBER SECURITY RISK FRAMEWORKS

CSRF		Description
CVSS	Common Vulnerability Score System	Standard for assessing the severity of cyber vulnerabilities on a scale of 1 to 10
FAIR	Factor Analysis of Information Risk	Methodology to quantify the factors related to your risk
OCTAVE	Operationally Critical Threat, Asset, and Vulnerability Evaluation	Comprehensive risk assessment and four-step assessment procedure, from selecting risk measurement criteria to identifying mitigation approach
CMMI	Capability Maturity Model Integration	5-level maturity model with guidance through the best case studies for the continuous improvement of production methods, especially in software development
TARA	Threat Agent Risk Assessment	Threat identification/classification and filtering methodology that prioritizes mitigations over threat importance
NIST SP800	National Institute of Standards and Technology, Special Publication	Cybersecurity risk framework. NIST SP800 is a comprehensive list of countermeasures/controls for cyber threats

Cyber risk can be considered as a relation with vulnerabilities, threats and consequences from intentional or unintentional events. It is necessary a comprehension of this three variables on an organizational environment to understand network resilience and drive measures. Tangible elements of risk in an information system can be listed like as assets, vulnerabilities and threats. The risk analysis process and definition of metrics to measure security maturity involves a sequence of steps like: (1) identification of vulnerabilities, (2) identification of threats, (3) determination of probability of occurrence, (4) determination of magnitude of the impact, (5) risk determination.

The implementation of IoT in industrial environments with different operation zones require wireless solutions with different levels of reliability and maturity. Information Systems (IT) and Operation Systems (OT) with different requirements must coexist safely in the airspace, requiring care during the implementation of new projects. To operate in automation and control networks, wireless protocols need high levels of security, reliability and maturity that can be achieved, for example, by the WirelessHART and ISA100 standards, which are used and approved by the main manufacturers of control and automation systems.

D. Wireless Sensor Networks for Critical Systems

Industrial systems are designed prioritizing reliability, safety of people and response time when defining requirements for the communication between sensors and systems. At this environment system communication must have mechanisms to ensure that data packets are exchanged in a determined time frame. Normally, safety systems use a periodic sensor data measure model to collect data to the safety controller [24].

Performance requisites for industrial sensors depends on the objective and criticality of the application to which it will be implemented. A user association for automation technologies in the process industries, NAMUR, defines 3 classes of industry wireless sensor applications in its standard NAMUR NE 124 for “Wireless Automation Requirements”. The International Society of Automation (ISA) also defined classes for the using of wireless industrial sensors in the specification ISA100.11a for field wireless devices. Table IV demonstrates types of industry application and classes defined by ISA and NAMUR.

TABLE IV. APPLICATION CLASSES OF INDUSTRIAL SENSORS

Application	Industrial Application Classes		
	NAMUR	ISA	Wireless Suitability
Safety	Class A Functional Safety	Class 0 Emergency Action	Lack of Maturity for Suitability
Control	Class B Process Management and Control	Class 1 Closed-Loop Regulatory Control	Suitable with considerations
		Class 2 Closed-Loop Supervisory Control	Suitable with considerations
		Class 3 Open-Loop Control	Suitable
Monitoring	Class C Display and Monitoring	Class 4 Alerting and Flagging	Suitable
		Class 5 Logging	Suitable

Process and machine safety systems shall comply with certain Safety Integrity Level (SIL) in order to ensure proper behavior when on safety demand. IEC 61508 standard defines the SIL from a set of safety and integrity requisites carried out by devices, software and the whole system. There are 4 levels of integrity for a safety loop (SIL1 to SIL4), where the level SIL4 is the most reliable. Neither WirelessHART nor the ISA100 directly support safety required mechanisms for SIL certification.

Then one alternative for using wireless communication for safety is combining these widely used WSN with safety certified protocols. There are few cases of SIL application with wireless networks like [16] mention that a SIL2 point to point communication between a machine safety controller and a wireless sensor was developed using ISA100.11a tunneled by PROFIsafe implemented for a hydrocarbon gas detection. WirelessHART current command limitation in the application layer still limits the implementation of a tunneling need to apply PROFIsafe with this protocol. Until a potential

modification in HART communication protocol specification is done it will not be possible to use PROFIsafe over wirelessHART.

The review of literature, vendors information and industry cases of WSN demonstrated that it is not possible to define a valid maturity level for the using of wireless networks for critical applications as functional safety. This lack of suitability for safety systems makes visible an opportunity to develop studies at this field. It is a call for manufacturers and for researchers to direct efforts to develop new approach and enhance technologies to enable the use of wireless devices in the field of people and machine safety, as well as SIL certification.

The benefit of expand wireless measurements in many fields of the industry is well known and brings elements to enhance operations performance. However, wide application of WSN is still recent and very dense networks are not a reality yet. Studies involving functional tests with application of hundreds, or thousands of devices are expensive and difficult to be accessible for independent researchers. Due to the lack of studies on real and dense scalability for WSN exists an opportunity for future studies to emulate conditions of scalability to high density networks to evaluate real performance of different protocols in expansion scenarios. Understand behavior of wide WSN scalability and scenario of future coexistence with dense networks is an important step to enhance resilience maturity that can enable secure application of this technology for more diversity of cases.

The review of vulnerabilities and threats in IoT networks when applied to critical systems demonstrated that it is still needed to enhance the level of maturity of these networks at industrial environment to deliver the secure and reliable operation required. Current WSN technologies like ISA100 and wirelessHART are mature for certain industrial applications however these networks should be designed and operated in accordance with security policies as defined by manufacturers and main wireless standards. These security requirements are not always native by the protocol and requires efforts to develop security culture and procedures for system maintenance.

V. CONCLUSION

The management of risks on wireless sensor networks in industrial environments depends on a deep comprehension of vulnerabilities during system lifecycle. Each network layer is differently affected by a huge number of threats which creates complex security management scenarios. The mapping of vulnerabilities and measures previously discussed as well as the application of resilience maturity models, are important tools to ensure proper level of network resilience during its lifecycle.

With the exploration of specific vulnerabilities, attack types, solution proposals described in the literature and specific issues when applying wireless sensor networks, it was possible to establish a path to increase maturity and resilience for wireless networks in critical systems. ISA100 and wirelessHART protocols demonstrated high level of resilience when using all security features available, but still need customized configuration to ensure high level of resilience.

With this article it was also possible to identify important gaps on the literature and with technologies to enable wireless

networks to be used for machinery and process safety carrying integrity level certification. Other literature gaps like performance analysis for high density wireless network in a real industrial environment was visualized, which enables opportunity for future studies.

REFERENCES

- [1] I. Jamaï, L. B. Azzouz, L. A. Saidane. Security Issues In Industry 4.0. IEEE, 2020. p. 481-488.
- [2] M. Dawson. Cyber Security in Industry 4.0: The Pitfalls of Having Hyperconnected Systems. v. 10, n. 1, p. 19-28. 2018.
- [3] S. Karmouskos. Stuxnet Worm Impact on Industrial Cyber-Physical System Security. IEEE, 2011. p. 4490-4494.
- [4] P. Varga, P. Sandor and G. Soos. Security Threats and Issues in Automation IoT. IEEE, 2017. p. 1-6.
- [5] K. N. Qureshi, A. H. Abdullah. Adaptation of Wireless Sensor network Industries and Their Architecture, Standards and Application. 2014.
- [6] A. Avizienis, J. C. Laprie, B. Randell, C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. V. 1, n. 1, p. 11-33, 2004.
- [7] D. Ratasich, F. Khalid, F. Geissler, R. Grosu, M. Shafique, E. Bartocci. A roadmap toward the resilient internet of things for cyber-physical systems. IEEE Access, v. 7, p. 13260-13283, 2019.
- [8] V. Gunes, S. Peter, T. Givargis, F. Vahid. A survey on concepts, applications, and challenges in cyber-physical systems. v. 8, n. 12, p. 4242-4268, 2014.
- [9] R. Ross, V. Pillipetri, R. Graubart, D. Bodeau, R. McQuaid. Developing cyber resilient systems: a systems security engineering approach. 2019.
- [10] J. Hudgens, T. Meers. Sources for Vulnerability and Threat Information. 2021. <https://pratun.com/blog/328-sources-for-vulnerability-and-threat-information>
- [11] I. Butun, P. Osterberg, H. Song. Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures. V. 22, n. 1, p. 2019.
- [12] A. Shukla, S. Tripathi. Security Challenges and Issues of Internet of Things: Possible Solutions. 2018. p. 26-27.
- [13] C. Bekara. Security issues and challenges for the IoT-based smart grid. V. 34, p. 532-537. 2014.
- [14] EMERSON. WirelessHART® and Wi-Fi® Security. Emerson Wireless Security Technical Note. 2017.
- [15] S. Raza. Secure Communication in WirelessHART and its Integration with Legacy HART. 2010.
- [16] K. Kitano, S. Yamamoto. Strong Security Measures Implemented in ISA100.11a Wireless System. Yokogawa, p. 57, 2014.
- [17] C. Alcaraz, J. Lopez. A security analysis for wireless sensor mesh networks in highly critical systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), v. 40, n. 4, p. 419-428. 2010.
- [18] P. Radmand, M. Domingo, J. Singh, J. Arnedo, A. Talevski, S. Petersen, S. Carlsen. ZigBee/ZigBee PRO security assessment based on compromised cryptographic keys. In: 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. IEEE, 2010. p. 465-470.
- [19] J. Tournier, F. Lesueur, F. Le Mouël, L. Guyon, H. & Ben-Hassine. A survey of IoT protocols and their security issues through the lens of a generic IoT stack. Internet of Things, 16, 100264. 2021.
- [20] J. H. Sanchez. WirelessHART Network Manager - Software Design and Architecture. 2011.
- [21] P. Anand, Y. Singh, A. Selwal, M. Alazab, S. Tanwar, N. Kumar. IoT vulnerability assessment for sustainable computing: threats, current solutions, and open challenges. IEEE Access, v. 8, p. 168825-168853.
- [22] D. Bodeau, R. Graubart. Cyber resiliency and nist special publication 800-53 rev. 2013.
- [23] K. Kandasamy, S. Srinivas, K. Achuthan, V. Rangan. IoT cyber risk: a holistic analysis of cyber risk assessment framework, risk vectors, and risk ranking process. V. 2020, n. 1, p. 1-18, 2020.
- [24] S. Petersen, N. Aakvaag. Wireless Instrumentation for Safety Critical Systems. 2015.

Detecting Various Chemical Samples and Cancer Cells With a Bio-Chemical Sensor By Using LNOI Based Optical Micro Ring Resonator (OMRR)

Md Ashif Uddin, Uzzwal Kumar Dey, Moriomece Akter

Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Bangladesh
ashifuddin9836@gmail.com, deyuazzwal.eceku, moriomece13ku@gmail.com

Abstract—In this work, we design a bio-chemical sensor based on Optical Micro Ring Resonator (OMRR) by using the feature of Lithium Niobate ($LiNbO_3$) on Insulator (LNOI) to sense the various concentrations of Sodium Chloride ($NaCl$), sucrose, and glucose in water as well as different cancer cells. Some conventional methods that are used in laboratories and industries to detect the different bio and chemical samples are expensive, slow, and waste a large volume of samples. Recently, OMRR has become very popular for sensing applications to overcome those limitations of conventional sensing methods as it is compact, robust, and cheaper. Furthermore, LNOI has some particular properties; these are photorefractive, favorable optics, piezoelectric, and photoelastic properties. In the early stages, we implemented different geometric structures of LNOI based OMRR in COMSOL to perform the mode of analysis. After that, we designed ring resonators in Opti-FDTD using the associated information of geometric structures from COMSOL to evaluate the achievement of this bio-chemical sensor. Our proposed design is found for the rib height, rib width, thickness of $LiNbO_3$, ring radius, and space between waveguide and ring of $0.56 \mu\text{m}$, $0.5 \mu\text{m}$, $0.16 \mu\text{m}$, $15 \mu\text{m}$, and 80 nm , respectively. Simulation results prove that this bio-chemical sensor design shows better Quality (Q) factor, high sensitivity, sharp peak at resonance wavelength, and low transmission output.

Index Terms—Optical Micro Ring Resonator, Effective Refractive Index, Resonance Wavelength, Bio-Chemical Sensor, and Quality Factor.

I. INTRODUCTION

The sensing technology and mechanism based on photonics has great possibilities to enable the new measurement of different concentration of samples more precisely in medical, laboratory and other sensing fields. Photonic based sensors are the subject of effective research to detect a wide and variety of chemical and biological samples. This sensor is involved in different application fields; these are biomedical, chemical, electrical, temperature, magnetic, pressure, rotation, position, vibration, and acoustic sensors [1]. Nowadays the Mystery challenge is to detect and analyze different genres of cancer cells such as leukemia, brain, cervical, and breast cancer. Cancer creates one of the most leading causes of death now. The American society of cancer association showed a report on cancer which expresses that more than 50% of

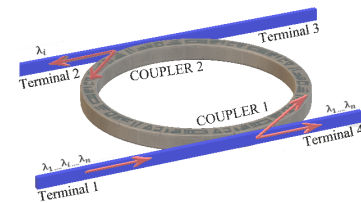


Fig. 1. Single ring resonator coupling procedure

death is involved by cancer. This assertion indicates that these deaths can be reduced by detection of cancer at the primary stages [2]. Currently, the traditional techniques, which can detect different cancers at the early period, are very limited where these are not only low resolution but also less sensitive. In recent times, optical sensing mechanisms have greatly enhanced sensor performance, particularly in the field of light source, fluidic design, and analyte interaction. It also improves in bio-chemical sensor sensitivity, fewer sample consumption, detection limit, faster time of detection, and lower cost in each measurement which illustrate that the economic interest in many laboratories have grown significantly to optical sensors [1]. Therefore nanostructures with their unique sensing properties are capable of replacing the traditional sensing mechanism.

Though OMRR was widely used as wavelength frequency converters, filters, circuits, optical communication, and switches in the last decade. In recent times, most researchers emphasize developing this resonator in different sensing fields. A OMRR is a set of ring and straight waveguides among which the ring shape waveguide acts as the resonant cavity and the straight waveguide acts as different ports [3]. The basic coupling procedure of OMRR is depicted in Fig. 1. In terminal 1, multiple wavelengths are given as inputs among which those that satisfy the resonant condition are coupled with the ring through a coupler. From this figure, we can observe that wavelength λ_i can only satisfy the resonant condition and thus, coupled with the ring and remaining input wavelengths are suppressed. Therefore, only the wavelength λ_i will be dropped from terminal 2, while

other wavelengths will pass through terminal 4 which is through port [4]. When resonance wavelength takes place, Only these specific wavelengths that satisfy the resonance condition will be capable of passing through the ring resonator fully and traveling wave dielectric material mode of micro cavity is formed in ring resonator structure [5]. The sensor mechanism of the ring resonator depends on light confinement of the waveguides.

The innovative contribution of our work is to design an optimized LNOI based OMRR to achieve high Q factor, sharp resonance wavelength, high sensitivity, and low transmission output. Moreover, several other designs related to our work are shown in Table I. However, compared with these works, in terms of designing simplicity and sensing capacity, our optimized sensor shows better performance compared to the sensing applications. We have designed 5 different models by changing the value of different size parameters to find out the optimized structure. We have shown that our optimized model gives best sensing performance by comparing with our other models. We also take different bio-chemical samples such as various concentrations of NaCl, sucrose, and glucose in water and different cancer cells to indicate that our sensor can sense different samples more precisely.

Additionally, we have proposed LNOI as a base material as it shows high sensing performance and sensitivity only for the tiny change in refractive indices and others. LNOI has a high refractive index and widespread transparent window. It also capables the detection of extremely little analyte quantities with fast sample preparation. This enhanced the interaction of light-matter in a cavity that increases the sensitivity. In addition, it exhibits a very strong bulk photovoltaic and significant photorefractive effect which are very attractive properties of LNOI. The main advantage of photorefractive properties of LNOI is to induce the change in refractive index linearly and it can be very useful properties in sensing application [6]. Due to the difficulties of dry etching process of LNOI, we may apply different techniques like wet etching, diamond dicing, chemical mechanical polishing (CMP), and femtosecond laser direct writing to the nano-structuring of LNOI [7]. For the fabrication process, if crystal ion slicing technique is used, the process of radius bending of $LiNbO_3$ waveguides becomes more perfect. This can help to achieve the high quality sidewalls on the LNOI nano-photonics structures [8].

The rest of this research paper is organized as follows. We explain some of our related works and show an effective comparison table on different ring resonators and sensing applications in Section II. In Section III, we focus on the basic principle of bio-chemical sensor using OMRR. Section IV represents the description of different models, material, method, and our working procedure using different software. In Section V, We show the sensing results of our proposed design to detect various bio-chemical samples, we further choose our proposed design by observing the performance of our other structure. We also conclude our paper by providing some further research plan in Section VI.

TABLE I
COMPARISON OF RELATED WORK FROM DIFFERENT PERSPECTIVES

Reference No.	Ring Qty	Ring Radius μm	Q Factor	Material Used
[8] naznin et al.	2	10, 5	413-1600	LNOI
[9] sharma et al.	1	0.2	75-1967	Silicon
[10] butt et al.	2	2, 1.8	321, 230	SOI
[11] deshours et al.	1	0.026	250	Copper
[12] tu at al.	1	5	NA	SOI
Our Proposed	1	15	2450-3706	LNOI

II. LITERATUR REVIEW

In recent times, to detect various constituents of blood for biomedical applications using a photonic crystal based ring resonator (PCRR) was successfully demonstrated in [9]. A serially cascaded micro ring resonator based on a hybrid plasmonic waveguide to detect multiple analysts simultaneously is presented in [10]. Here, they used two ring waveguides with separate ring radiuses which are coupled serially through a silicon strip waveguide. This model shows best sensitivity when it is observed that ring radiuses are 2.0 μm and 1.8 μm , respectively. A planar sensor based on the prototype of microwave ring resonator is presented in [11] for sensing different biological tissues by observing high sensitivity. A dual and label free optical micro ring resonator biosensor based on LNOI technology is presented for bio sensing applications in [8]. In [12], a high-sensitivity complex refractive index sensing SWG micro-ring resonator is proposed. In [13], A chemical sensor is designed to achieve low confinement losses and high sensitivity for sensing various concentration of ethanol-water and benzene-toluene mixer at the wavelength range of 800 to 1600 nm using photonic crystal fiber (PCF).

From Table I it can be stated that the ring radius of our proposed glucose sensing resonator model is comparatively larger compared to other resonator models and only one ring. Therefore, our resonator can be easily fabricated and less complex for bio-chemical sensing applications. Furthermore our proposed model also shows the highest Q factor compared to all related works in Table I. Thus, to our knowledge, our proposed OMRR model for sensing different bio-chemical (NaCl, sucrose, glucose, and cancer cells) is less complex and easy to fabricate, and provides best sensitivity compared to the state-of-the-art works.

III. BIO-CHEMICAL SENSING

Sensing of different bio-chemical samples is one of the most effective and important sectors in laboratory, medical, and industry to detect various types of liquid concentrations. The structure of our OMRR sensor is designed in a way such that it confines light within a cavity by creating resonance. When a specific wavelength of incoming light or laser satisfies the condition of resonance, it couples into the ring waveguide from the input straight waveguide and continuously recirculates among it. By the same process, resonant light escapes from the ring waveguide to the output waveguide where

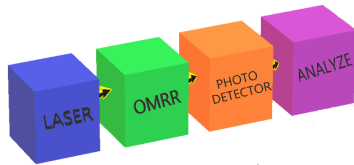


Fig. 2. Block diagram for bio-chemical sensing with micro ring resonator

the rest of all wavelengths of input light propagate into the through port. We can observe the resonance wavelength from the through port. By persuading these properties, OMRR for bio-chemical sensing has become more exoteric and popular. In this work, we take input light as 1550 nm and output is analyzed between 1250 to 1300 nm wavelength. Then, this input light is propagated through the in port of the OMRR. The different samples are deposited on the top of the straight and ring waveguides where we set up a photodiode detector on the through port for analyzing the sensing performance. Various samples contain different refractive indices which are specific and fixed for that material. And when we change the sample with various concentrations on the top of the ring resonator, their refractive index as well as effective refractive index will be changed. Therefore, the mode of propagation of this structure will be changed and these create a resonance wavelength shift. So more difference between two consecutive effective refractive indices, more resonance wavelength shift which is the attractive property for sensing. A block diagram of bio-chemical sensing procedure is presented in Fig. 2.

IV. METHODS AND MATERIALS

The main theme of our research work is to sense various samples by designing and simulating an LNOI based OMRR. At first, we designed the geometric structure of LNOI at COMSOL to get effective refractive indices of different solutions. Then, using these effective refractive indices as waveguide materials, we drew specific ring resonators in Opti-FDTD. We observed and analyzed all the sensing parameters from the wavelength vs. transmission curve to detect different concentrations of NaCl, sucrose, and glucose in water and various cancer cells. Therefore two software are needed for this observation, design, and simulation process i.e. COMSOL and Opti-FDTD. We have used COMSOL RF solver and transverse electric (TE) mode for resonance as TE shows better power confinement. In Opti-FDTD, we also use Gaussian continuous modulated waves as a propagation mode.

A. Comsol design

COMSOL multiphysics is a simulation based software that uses finite element analysis to calculate power confinement and effective refractive index. To draw a geometry structure of the ring resonator, we used COMSOL RF module-Perpendicular wave-hybrid analysis. Then from physics properties, we changed the frequency into wavelength. Then, we drew a LNOI based geometry structure which is a small portion of our ring resonator as shown in Fig.

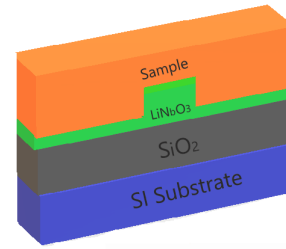


Fig. 3. Layout of micro-ring resonator geometry

3. LNOI has three tiers where the base tier is made with Silicon ($n_{Si}=3.479$) [14]. The buffer tier is made with Silica ($n_{SiO_2}=1.444$) [15]. The upper tier is formed with Lithium Niobate ($n_{LiNbO_3}=2.211$) [15]. The rib structure with specific height and width is formed of $LiNbO_3$. The input light is set at 1550 nm wavelength [15]. Then, we set a specific refractive index of different concentrations of samples from the physics subdomain section. After that, we set up a boundary condition of our geometry structure. Next, we analyzed the mesh view which is a smaller triangle in the active region. Then, we selected the solve-solver parameter for getting mode solution and effective refractive index. All the refractive and effective refractive indices of different concentrations of samples using our optimized structure are shown in Table IV, V, VI, and VII accordingly.

B. Opti-FDTD design

Opti-Finite Difference Time-Domain (FDTD) is a very effective method for modeling Computational Electromagnetic (CEM) processes. We used FDTD software for drawing a ring resonator with a specific light source and analyzing all sensing parameters from the wavelength vs. transmission curve. To design a ring resonator in Opti-FDTD software, firstly we selected a "waveguide layout designer" for setting up the material properties from "profile and material". Now from the channel section, we find our proposed material to be selected with specific name, height, width, and thickness. By using the similar process, we also selected other waveguides and Si as a base wafer material. Then, we drew two linear waveguides and a ring waveguide from the draw section and selected a specific profile which is already created in the "profile and material" section. We inserted an input light as a vertical input plane from the draw section. We used a Gaussian modulated continuous wave with positive Z direction. The mode of one of the waveguides is a full vector along the Y axis where input light will be propagated for sensing operation. We added an observation point for analysis at the through port. After completing design, we simulated and observed wavelength vs. transmission curve for sensing application. The ring resonator design using Opti-FDTD software is presented in Fig. 4.

C. Performance parameters

1) *Quality factor*: It is defined as the ratio between the resonance wavelength and the half of the width of the curve

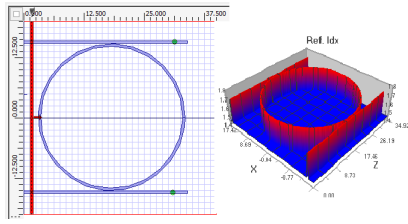


Fig. 4. Optical micro ring resonator layout design using Opti-FDTD

for which resonance wavelength is produced [16].

$$Q = \lambda_r / \Delta\lambda \quad (1)$$

Here, λ_r is called resonance wavelength where $\Delta\lambda$ is known as half of the width of the curve for which resonance wavelength is produced.

2) *Sensitivity*: It is defined as the ratio between the distance of resonance wavelength and difference between refractive indices [17].

$$S = \Delta\lambda_r / \Delta n_c \quad (2)$$

Here, $\Delta\lambda_r = \lambda_r - \lambda_0$ and $\Delta n_c = n_c - n_0$, where λ_0 and λ_r are initial and respective resonance wavelength, respectively.

D. Different designs

We have designed different LNOI based sensing models with several parameters and compared those designs to find out the optimized design. At first, we designed a LNOI based geometric layout using COMSOL software. The base tier of this structure is about 1 μm thick Silicon. The buffer tier is settled with 1 μm thickness of SiO_2 . The upper tier is made of 0.16 μm thick LiNbO_3 . The rib structure with specific width and height is formed of LiNbO_3 are 0.5 μm and 0.56 μm , respectively. The different samples with various concentrations are deposited 0.8 μm from the upper tier. In Opti-FDTD, the thickness and width of the straight and ring waveguides used are 0.5 μm , and 0.5 μm . The ring radius is calculated as 15 μm and the proper space between the ring and straight waveguide is about 80 nm and the length of each one straight waveguide is approximately 34 μm . We use these parameter's values in Design 1. By changing the size of LiNbO_3 thickness and ring radius, we designed 5 OMRR structures which are presented in Table II. We already fixed the rib height and width by analyzing the performance parameter with increasing and decreasing the size of rib height and width in [16].

V. RESULT AND DISCUSSION

A. Determining proposed design by performance analysis of different designs

Here, we determine the optimized design by observing the impact of LiNbO_3 thickness and ring radius. It is clear that if we increase or decrease the size of the parameters such as ring radius and LiNbO_3 thickness from the Design 1 of Table

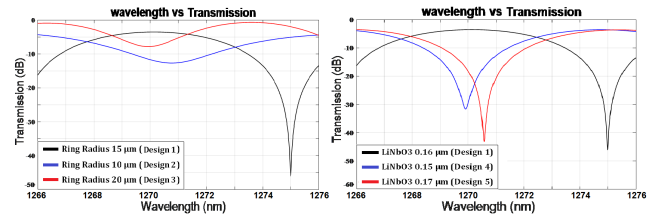


Fig. 5. (a) Comparison of Design 1, 2, and 3 by varying ring radius (b) Comparison of Design 1, 4, and 5 by varying thickness of LiNbO_3

TABLE II
DIFFERENT DESIGNS BY VARYING VARIOUS SHAPE OF PARAMETERS

Design Name	LiNbO_3 Layer μm	Rib H μm	Rib W μm	Ring Radius μm	Distance between Waveguide to Ring (nm)
Design 1	0.16	0.56	0.5	15	80
Design 2	0.16	0.56	0.5	10	80
Design 3	0.16	0.56	0.5	20	80
Design 4	0.15	0.56	0.5	15	80
Design 5	0.17	0.56	0.5	15	80

TABLE III
THE OBSERVED RESONANCE WAVELENGTH, Q FACTOR, AND TRANSMISSION OUTPUT OF DIFFERENT DESIGNS

Design Name	Resonance Wavelength (nm)	Quality factor	Transmission (Output dB)
Design 1	1274.9	2450	-46.1946
Design 2	1270.75	290	-12.69
Design 3	1269.94	465	-7.83
Design 4	1269.92	1097.1	-31.6703
Design 5	1270.57	2173	-40.8231

II, none of the wavelength vs. transmission curves of these figures show sharp resonance and better Q factor than Design 1. Fig. 5(a) shows the comparison among Design 1, 2, and 3 only varying the ring radius while keeping other parameters permanent. By same process, in Fig. 5(b), we varied only the thickness of LiNbO_3 in Design 1, 4, and 5. In every design, we only change one specific size of parameter while keeping all other sizes of parameters permanent as well as fixed. From Table III, we can observe that Design 1 shows highest Q factor, lowest transmission output, and sharpest resonance peak among these designs. So we can state that the Design 1 is our proposed structure to sense different biochemical samples.

B. Bio-chemical samples (NaCl, sucrose, glucose, and cancer cell) sensing with proposed design

Here, we analyze and explore the capability of the optimized sensor by utilizing wavelength vs. transmission curves of Fig. 6 and 7 and Table IV, V, VI, and VII. From Fig. 6(a) and Table IV, when NaCl concentration in water is about 5%, a green curve is shown whose Q factor is 1438. Then we increase the NaCl concentration at 10%, after that, our

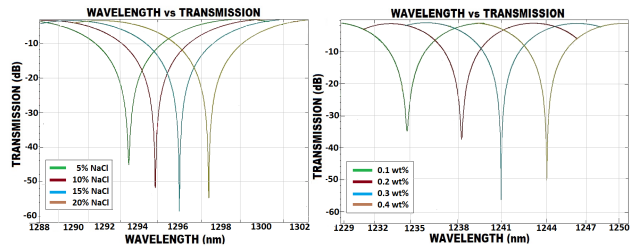


Fig. 6. Wavelength vs. transmission curve of our optimized design at various concentration of NaCl (a) and Sucrose (b) in water

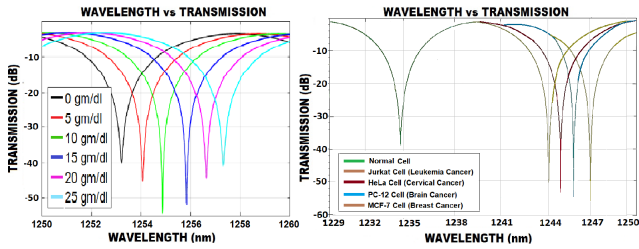


Fig. 7. Wavelength vs. transmission curve of our optimized design at various concentration of glucose in water (a) and different cancer cell (b)

optimized design indicates the red curve of Fig. 6(a). The Q factor and sensitivity are 2345 and 133.4, respectively. All information on sensing using different NaCl concentration in water as shown in Fig. 6(a) are tabulated in Table IV. By similar process, the sucrose and glucose concentration sensing in water are depicted in Fig. 6(b) and 7(a), respectively and the related information is summarized in Table V and VI. For different cancer analysis, we observed normal cells and various cancer cells such as jurkat, HeLa, PC-12, and MCF-7 cells. From Fig. 7(b) and Table VII, when the cell is normal with refractive index 1.350, green curve is shown where effective refractive index and Q factor is 1.7150 and 1732, respectively. Then when we analyzed different cancer cell with refractive index (1.39 to 1.40), different curves are displayed where range of effective refractive index, Q factor, and Sensitivity are (1.7585 to 1.7617), (2252 to 4370), and (237 to 188), respectively. It can be clearly stated that the difference of resonance wavelengths of normal cell and cancer cells is 9.1nm. All sensing information using different cancer cells as shown in Fig. 7(b) are tabulated in Table VII. So, we can easily detect different cancer cells by using our proposed OMRR based bio-chemical sensor.

C. An effective analysis of our optimized bio-chemical sensor based on resonance wavelength, spectral Shift, quality Factor, and sensitivity

Fig. 8 shows the concentration vs specific resonance wavelength and spectral shift for different bio-chemical samples such as NaCl, sucrose, and glucose as well as various cancer cells. From this figure, due to the increment of NaCl, sucrose, glucose concentration in water, spectral shift increases and thus refractive and effective refractive indices of solution are also increased which are tabulated in Table IV, V, VI, and

TABLE IV
VARIOUS CONCENTRATION OF NaCl IN WATER WITH SENSING PERFORMANCE PARAMETERS

Concentration (%) [18]	Refractive Index [18]	Effective Refractive Index	Quality Factor	Sensitivity S
5	1.342	1.7130	1438	–
10	1.351	1.7157	2345	133.4
15	1.359	1.7170	4210	141.2
20	1.369	1.7193	2526	137.1

TABLE V
VARIOUS CONCENTRATION OF SUCROSE IN WATER WITH SENSING PERFORMANCE PARAMETERS

Concentration (wt%) [19]	Refractive Index [19]	Effective Refractive Index	Quality Factor	Sensitivity S
0.1	1.346	1.7114	1196	–
0.2	1.363	1.7780	1252	200
0.3	1.378	1.7214	3650	184.4
0.4	1.398	1.7262	2560	171.3

TABLE VI
VARIOUS CONCENTRATION OF GLUCOSE IN WATER WITH SENSING PERFORMANCE PARAMETERS

Concentration (gm/dl) [20]	Refractive Index [20]	Effective Refractive Index	Quality Factor	Sensitivity S
0	1.3155	1.7071	1564	–
5	1.3222	1.7086	1788	134.4
10	1.3279	1.7100	3795	137.1
15	1.3357	1.7117	2090	131.1
20	1.3421	1.7131	1780	132.2
25	1.3474	1.7143	1568	128.5

TABLE VII
VARIOUS CANCER CELL WITH SENSING PERFORMANCE PARAMETERS

Name of The Cell	Different Cancer Name	Refractive Index	Effective Refractive Index	Quality Factor	Sensitivity (S)
Normal	No	1.350	1.715	1732	–
Jurkat	Leukemia	1.390	1.759	2252	237
HeLa	Cervical	1.392	1.760	3350	205
PC-12	Brain	1.395	1.761	3840	194
MCF-7	Breast	1.401	1.762	4370	188

VII. The increment in the effective refractive index changes the mode of propagation and thus causes the shift of resonance wavelength through the positive direction. This indicates that the spectral resonance shift becomes wider with increasing

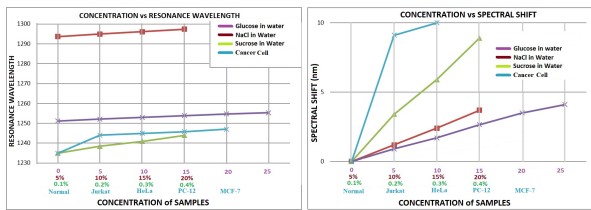


Fig. 8. Resonance wavelength (a) and Spectral shift (b) for different bio-chemical samples

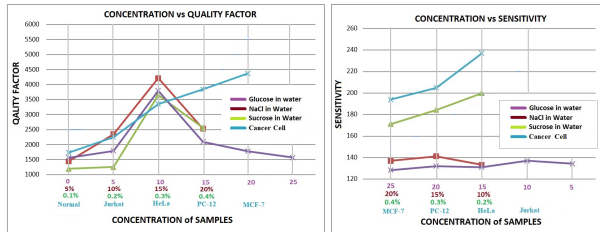


Fig. 9. Quality factor (a) and Sensitivity (b) for different bio-chemical samples

concentration in samples. In Fig. 9, we observe the Q factor and sensitivity of different NaCl, sucrose, and glucose in water and various cancer cells. From this figure it is clearly stated that Q factors are increased in terms of sharp resonance wavelength by identifying the different curves of Fig. 6 and 7. The sharpest is the resonance peak and the highest is the Q factor. We can also observe the sensitivity of different samples. From this figure, we can state that the relative sensitivity increases with the decrease of concentration in water. These Fig. 8 and 9 give a clear indication that our proposed device has comparatively great capability to sense various concentrations.

VI. CONCLUSION AND FUTURE WORK

Now-a-days, OMRR is a very attractive field for research and implementation. Through this work, we showed 5 designs of OMRR by varying different ring sizes and $LiNbO_3$ thickness. We also compared our optimized model with some related works. Among all these, we proposed our optimized design which is Design 1 that can detect the various concentrations of NaCl, sucrose, and glucose in water as well as different cancer cells most precisely with respect to others. The optimized design contains ring radius and bus waveguides of $15 \mu\text{m}$ and $34 \mu\text{m}$, respectively. The height and width is $0.5 \mu\text{m}$ and $0.5 \mu\text{m}$ for both ring and bus waveguides. We also analyzed a small portion of this ring resonator as a geometry structure which shows wider change in effective refractive index. Simulation results also indicate that our optimized design is capable of sensing different samples more perfectly. In the future, we have a robust motive to fabricate our optimized LNOI based OMRR design and compare the practical result with the simulated one.

REFERENCES

- [1] V. M. Passaro, B. Troia, M. La Notte, and F. De Leonardis, *Chemical sensors based on photonic structures*. IntechOpen, 2012.
- [2] L. Ali, M. U. Mohammed, M. Khan, A. H. B. Yousuf, and M. H. Chowdhury, "High-quality optical ring resonator-based biosensor for cancer detection," *IEEE Sensors Journal*, vol. 20, no. 4, pp. 1867–1875, 2019.
- [3] J. Scheuer, Y. Yadin, and M. Margalit, "Micro-ring resonator," Apr. 26 2005, uS Patent 6,885,794.
- [4] S. Sahai, A. D. Varshney, and S. Varshney, "Analysis, modeling and simulation of single all-silicon ring resonators and their comparison on the basis of their coupling lengths," in *AIP Conference Proceedings*, vol. 2136, no. 1. AIP Publishing LLC, 2019, p. 050011.
- [5] "Optical microcavity," accessed: 2020-08-17. [Online]. Available: https://en.wikipedia.org/wiki/Optical_microcavity
- [6] A. Boes, B. Corcoran, L. Chang, J. Bowers, and A. Mitchell, "Status and potential of lithium niobate on insulator (Inoi) for photonic integrated circuits," *Laser & Photonics Reviews*, vol. 12, no. 4, p. 1700256, 2018.
- [7] Y. Qi and Y. Li, "Integrated lithium niobate photonics," *Nanophotonics*, vol. 1, no. ahead-of-print, 2020.
- [8] S. Naznin and M. S. M. Sher, "Design of a lithium niobate-on-insulator-based optical microring resonator for biosensing applications," *Optical Engineering*, vol. 55, no. 8, p. 087108, 2016.
- [9] P. Sharma and P. Sharan, "Design of photonic crystal based ring resonator for detection of different blood constituents," *Optics Communications*, vol. 348, pp. 19–23, 2015.
- [10] M. Butt, S. Khonina, and N. Kazanskiy, "A serially cascaded microring resonator for simultaneous detection of multiple analytes," *Laser Physics*, vol. 29, no. 4, p. 046208, 2019.
- [11] F. Deshours, G. Alquié, H. Kokabi, K. Rachedi, M. Thili, S. Hardinata, and F. Koskas, "Improved microwave biosensor for non-invasive dielectric characterization of biological tissues," *Microelectronics Journal*, vol. 88, pp. 137–144, 2019.
- [12] Z. Tu, D. Gao, M. Zhang, and D. Zhang, "High-sensitivity complex refractive index sensing based on fano resonance in the subwavelength grating waveguide micro-ring resonator," *Optics express*, vol. 25, no. 17, pp. 20911–20922, 2017.
- [13] M. M. S. Maswood, M. A. Uddin, U. K. Dey, M. M. I. Mamun, M. Akter, S. S. Sonia, and A. G. Alharbi, "A novel sensor design to sense liquid chemical mixtures using photonic crystal fiber to achieve high sensitivity and low confinement losses," in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2020, pp. 0686–0691.
- [14] T. Claes, J. G. Molera, K. De Vos, E. Schacht, R. Baets, and P. Bienstman, "Label-free biosensing with a slot-waveguide-based ring resonator in silicon on insulator," *IEEE Photonics journal*, vol. 1, no. 3, pp. 197–204, 2009.
- [15] "Refractive index database," accessed: 2018-02-01. [Online]. Available: <https://refractiveindex.info/>
- [16] M. A. Uddin, M. M. S. Maswood, U. K. Dey, A. G. Alharbi, and M. Akter, "A novel optical micro ring resonator biosensor design using lithium niobate on insulator (Inoi) to detect the concentration of glucose," in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, 2020, pp. 350–354.
- [17] L. Hasanah, H. S. Nugroho, C. Wulandari, B. Mulyanti, D. D. Berhanuddin, M. H. Haron, P. S. Menon, A. R. Md Zain, I. Hamidah, K. Khairurrijal *et al.*, "Enhanced sensitivity of microring resonator-based sensors using the finite difference time domain method to detect glucose levels for diabetes monitoring," *Applied Sciences*, vol. 10, no. 12, p. 4191, 2020.
- [18] D. Z. Stupar, J. S. Bajić, A. V. Joža, B. M. Dakić, M. P. Slankamenac, M. B. Živanov, and E. Cibula, "Remote monitoring of water salinity by using side-polished fiber-optic u-shaped sensor," in *2012 15th International Power Electronics and Motion Control Conference (EPE/PEMC)*. IEEE, 2012, pp. LS4c–4.
- [19] C.-H. Chen, T.-C. Tsao, J.-L. Tang, and W.-T. Wu, "A multi-d-shaped optical fiber for refractive index sensing," *Sensors*, vol. 10, no. 5, pp. 4794–4804, 2010.
- [20] Y. Ong, W. Kam, S. Harun, R. Zakaria, and W. S. Mohammed, "Low-cost transducer based on surface scattering using side-polished d-shaped optical fibers," *IEEE Photonics Journal*, vol. 7, no. 5, pp. 1–10, 2015.

Proposing A Cloud and Edge Computing Based Decision Supportive Consolidated Farming System By Sensing Various Effective Parameters Using IoT

Md Ashif Uddin, Uzzwal Kumar Dey, Moriom Akter

Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Bangladesh
ashifuddin9836@gmail.com, deyuzzwal.eceku@gmail.com, moriomece13ku@gmail.com

Abstract—Now-a-days, we are facing significant challenges as a result of food shortages and a lack of large-scale food production utilizing traditional methods. As a result, global food supply rivalry has become the most pressing issue of the world. Cloud computing based on Internet of Things (IoT) and Wireless Sensor Networks (WSN) have become a comforting part of the solution to overcome this situation by replacing traditional methods in food production and monitoring systems. In this work, we represent a consolidated farming system with LED, SMS, E-mail, and website-based monitoring using WSN and IoT. To monitor a smart farm, several effective sensors are utilized in various industries, such as agriculture, aquaculture, livestock, air quality, weather monitoring, and health monitoring. All sensor's data is reserved on the IoT cloud platform as well as our built website named Website for Consolidated Smart Farms Monitoring (WCSFM), allowing a concerned person to readily access all relevant data from anywhere on the planet. Our model notifies the end user of expected information based on predefined sensor values by sending SMS, E-mail notifications, and turning on LEDs. Moreover In remote places, our IoT and WSN-based approaches may be more useful.

Index Terms—Internet of Things, Cloud Computing, Wireless Sensor Networks, Consolidated Farms Monitor, and Sensors

I. INTRODUCTION

Agriculture is a vital sector in the economic development of improving countries like Bangladesh. It is impossible for a farmer to physically irrigate and fertilize the entire field while making the best use of these. As a result, a sophisticated agriculture infrastructure system is needed to take necessary steps across the entire field in order to increase crop yield [1]. Aquaculture will generate about 62% of all seafood produced for human consumption by 2030. Our ocean's and other natural resources overfishing is increasing year after year. Aquaculture is the answer to future generation's access to healthy, sustainable protein sources. The need for animal protein will increase about 52% by 2050 when the population of the world will reach 10 billion [2]. The livestock industry is an important part of the global food system that helps to alleviate poverty, ensure food security, and promote agricultural development. Livestock generates 40% of worldwide agricultural output and supports the livelihoods. Simultaneously,

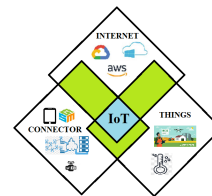


Fig. 1. The fundamental IoT model for connecting different things.

there is a need to enhance livestock sector practices in order to make them more sustainable, equitable, and less hazardous to animal and human health [3]. Furthermore, in areas where the weather changes frequently, it is critical to managing the crop in accordance with the changing weather. We are expected to boost crop yields by employing the probability of climate change. There is substantial evidence that poor air quality can harm the respiratory health of numerous livestock species. Farm employees may be affected by long-term exposure to particles in barns. Livestock farming emits a wide range of pollutants, ammonia, methane, nitrous oxide, and particles such as dust and microbes are all examples of airborne emissions [4].

Due to lack of a proper monitoring system, a variety of obstacles to farming situations arise. Farming growth is inhibited by several problems like lack of environmental issues and effective decision making, resulting in the loss of certain food while they are still in the juvenile stage [5]. As the world's population expands, so does the amount of food needed. Traditional technologies have been used to address these issues, but they require a substantial amount of labor, so the investment is always significant. As a result, an automated method based on IoT and WSN will be a cost-effective and efficient way to deal with the problems. An IoT is a network which is connected with computing devices, digital and mechanical gear, and objects that can perceive, collect, and transmit data over the internet without human interaction. It has changed the society where we live now. Fig. 1 presents various things connected with the internet which support IoT. Smart farming is necessary to address the aforementioned is-

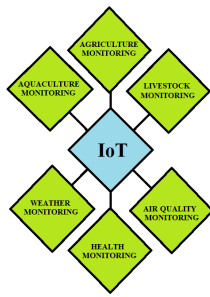


Fig. 2. IoT connecting different farming fields in an intelligent monitor.

sue. Weather, traditional agri-aquaculture techniques, climate change, and other environmental factors are all influenced by farming techniques. Nowadays, technology is a boon to the modern world. Internet-based technologies have recently made our lives easier and more comfortable in terms of power, size, and cost consumption. Those technologies could also be used in agriculture, communication, medicine, and other industries. Using IoT and WSN, it is feasible to improve and monitor food production more effectively. Fig. 2 depicted the IoT connecting different farming fields with an intelligent monitor.

By persuading these problems and necessity, we have developed an IoT, cloud computing, and website based intelligent consolidated farming technique that can effectively notify users or farmers of smart farming, air quality, weather, and health monitoring by using wireless sensor networks and environmental data. Our work contributes to the field of agriculture, aquaculture, air quality, weather monitoring, and livestock husbandry by developing an intelligent framework based on IoT and edge computing that delivers automated notifications. The unique contribution of our work is to develop a website which can be accessed from anywhere from the globe and get relevant all sensing information from farms for making proper decisions. In this manner, a person will be notified via not only SMS and E-mail, but also LED status. These combined notifications improve our farm monitoring system more effectively. To our knowledge, no research has looked at using IoT and cloud computing technology to create an intelligent consolidated agricultural technique.

The rest of the paper is laid out as follows: Section II examines similar studies in several fields. Section III shows an intelligent farming system based on IoT and Cloud Computing. The framework of our designed intelligent farming process is described in Section IV. Section V contains our created circuit schematic for sensing different farming parameters, as well as the seven-days notification results. Finally, Section VI brings the paper to a close.

II. RELATED WORK

The researchers developed a novel method for using sensors to monitor weather and climate so that the technology makes real-time data about temperatures, humidity, and wind velocity via the IoT [6]. Parra et al. suggested a smart irrigation

system which is made with a smart irrigation tube that uses humidity and water content sensors to work. [7]. Intensive aquaculture has been condemned for its negative impact on the environment and the usage of antibiotics. Integrated Multi-Trophic Aquaculture (IMTA) is the practice of cultivating a variety of species in such a way that uneaten food and waste can be captured [8]. Sensors, a processing unit, a gateway, and a cloud platform are the three main components of an IoT system. Sensors collect data and deliver it to a processing unit, which processes it according to the requirements and then sends the processed data to a cloud platform for air quality monitoring protocol as proposed by Gupta et al [9].

For successful cattle farming, Artificial Intelligence (AI) is a data-driven technology that can process and represent huge volumes of data acquired by sensors and IoT [10]. Other sensors, such as RFID sensors, body position sensors, respiration sensors, and CO2 sensors, are employed in some health monitoring systems. They also discuss smartphone-based and microcontroller-based health monitoring systems [11]. Using an IoT-based website and a smart umbrella, created a system for sending automated weather notifications. A person is notified by the smart umbrella turning on a light and sending an SMS to their phone [12]. Finally, Uddin et al. proposed an intelligent rice-fish farming process that can effectively notify users of water and fertilizer management events where different sensors are distributed to monitor environmental and physical factors including pressure and temperature [13].

III. SENSOR DEPLOYED CONSOLIDATED FARMING SYSTEM BASED ON IoT AND CLOUD COMPUTING

Different sensors are implemented in various sectors like agriculture, aquaculture, livestock, air quality, weather, and health monitoring. All sensing data is stored in a gateway node, which is passed to the end user through the internet. Different protocols are used to transmit data from node to node. The UDP protocol is widely used and we use this protocol for sending those sensing information from node to user end via the internet. Different fields that are monitored by end users via the internet are presented in Fig. 3. In fig. 4 depicts the block diagram of field deployed sensors connected to the smart IoT network. In this figure, various sensors are connected by a microcontroller, which is the LPC1768 [14]. All sensing information from different sensors are collected by this microcontroller and, based on the predefined values that are set for comparison with sensing parameters, information is passed to mobile SMS via GSM and E-mail via the internet. This network is designed based on full IP that creates it more easily to end users and minimizes the cost of maintenance and better monitoring in rural, suburban, and urban areas [15]. In rural areas, where the network is weak, the femtocell is more effective to transmit data from the internet to user [16].

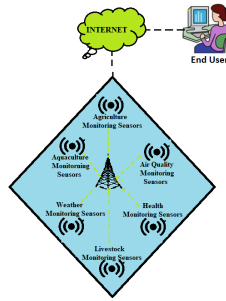


Fig. 3. Data from field-deployed sensors is linked to the Internet for users.

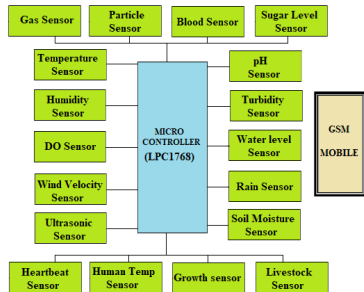


Fig. 4. Field deployed sensors are connected to the smart IoT network represented in the block diagram [13].

IV. THE FRAMEWORK OF OUR DESIGNED CONSOLIDATED FARMING SYSTEM

The main theme of our work is to propose an IoT and cloud-based consolidated farm monitoring system by using different sensors that are deployed in various fields. All sensing data is stored on our developed website, named WCSFM. Besides this, our model also sends SMS, emails, and turns on LEDs by comparing predefined values with sensing values, which makes this model more effective and feasible. For hardware design, we take different things such as Arduino UNO R3, Node MCU ESP8266, AC-DC 5V 2A power supply, LPC1768, SIM900A GSM, gateway node, access node, various sensors, LEDs, and jumpers. For website design, we use different software such as javascript, CSS, HTML, and python.

A. Establishing our proposed model with the intelligent consolidated farming model

The basic working process of our proposed model is shown in Fig. 5. In this figure, there are 3 main layers, where the first layer is called the collection layer. In this sector, all sensing information of different sensors from farming fields is collected by gateway nodes. This sensor information is placed in layer 2, which is the decision-making layer. In this layer, some specific values for different types of sensing information are predefined as threshold values. At a specific time, this layer checks all sensing information by utilizing threshold values and makes decisions by turning on different LEDs. At

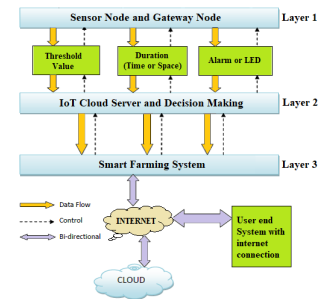


Fig. 5. Overall system architecture, from raw data sensing to intelligent decision making [17].

the final layer, which is called the application layer, a person can be notified by LEDs, SMS, E-mail, proposed website [17].

B. Presenting the work step-by-step to show the overall working process

The overall working process of our proposed model is depicted in Fig. 6. In this figure, different sensors are deployed in different monitoring places. All sensing information from different fields are collected and stored at the edge point by using WSN gateway nodes. To monitor the agriculture field, we deployed different sensors like temperature, humidity, soil moisture, water level, and pH sensors. To get the proper information about aquaculture, we inserted temperature, humidity, turbidity, water level, DO, and pH sensors. For the livestock farm monitor, we deployed gas, water level, count, temperature, and humidity sensors. To know the air quality of these farming areas, we also take temperature, pH, gas, humidity, and particle sensors. To monitor this, we connect temperature, humidity, wind velocity, and wind direction sensors. To track our health information, we use BP, sugar level, pulse rate, and body temperature sensors. IoT cloud servers collect all relative information from access points and check it by comparing it with predefined values that are stored in the decision making sector. Those associated decisions for individual fields are transmitted to the end user through SMS and E-mail. The user can also access all the sensing information by using our developed website named WCSFM from anywhere on the globe and may take effective decision for farm monitoring.

C. Developing a cloud-based website model that is compatible with the deployed sensor data store and distribution

WCSFM is a website developed using javascript, CSS, HTML, and python. On this website, all the sensing information is placed automatically by the gateway and node access point. We developed WCSFM where we can see all the sensing information from agriculture, aquaculture, livestock, air quality, weather, and health monitoring at a glance. Using this website, a concerned person can get all the sensing information from any place, which may boost our productivity. The proposed experimental homepages of different field-sensing data viewing website are shown in Fig. 7.

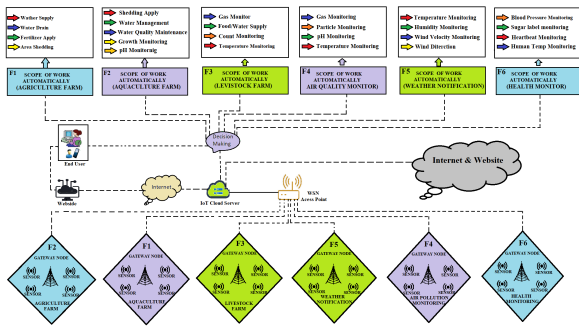


Fig. 6. Intelligent system models ranging from sensing to decision making, with a demonstration of the scope of work.

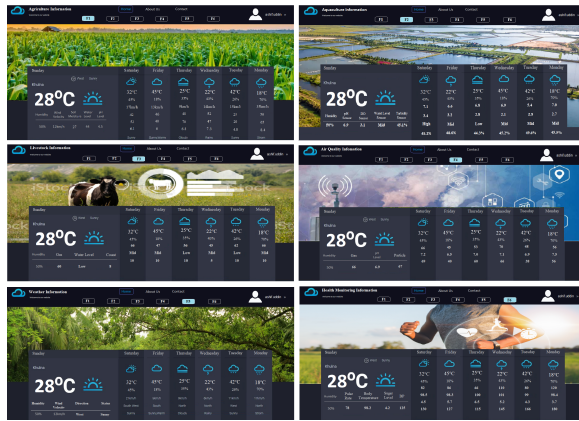


Fig. 7. A sample website view when system monitoring information is visible in a consolidated farming manner.

D. Proposed algorithm techniques for different farming field monitoring intelligently

The proposed algorithm for consolidated farm monitoring systems is presented in Fig. 8. Based on this figure for the agriculture monitoring, when the soil moisture sensed data is less than 40, turning on the red LED is satisfied and the white LED turns OFF. This indicates that water is needed at the farm. On the other hand, when this data is higher than 40, the red light is turned OFF and the white LED is turned ON, which indicates that there is no need to supply water at the farm. Analogously, when the water level is greater than 60 ppm, the blue LED turns ON and the associated white LED turns OFF. This LED status indicates that we need to drain water from the farm. In contrast, when the water level is less than 60 ppm, the blue LED turns OFF and the associated white LED turns ON, which represents that there is no need to drain water from the farm. Similarly, based on various sensor information in different fields and with individual algorithms, they are depicted in Fig. 8.

V. RESULTS AND DISCUSSION

The schematic circuit diagram of our intelligent consolidated farming model hardware connections is shown in Fig. 9. Our implemented sensor's sensed values, which are stored at

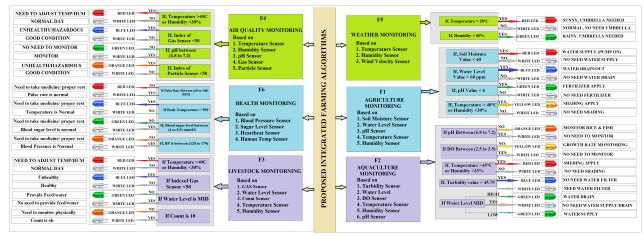


Fig. 8. A proposed algorithm for a consolidated brainy farming model ranging from monitoring to physical view.

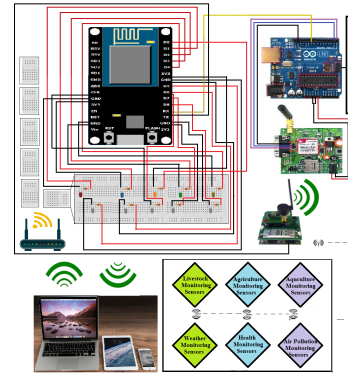


Fig. 9. The system circuit diagram for an consolidated farming model displays hardware connections.

the website, are tabulated in Table I and II. For observing agricultural information, From the first row of the first subtable of the Table I, the first temperature, humidity, soil moisture, water level, and pH level read from the website and IoT cloud server are 28°C, 50%, 27, 44, and 4.5. These are depicted in the first subtable of Table I. As long as the value of soil moisture and pH level are below the threshold, the condition for turning on the red and green LEDs is satisfied and they remain ON while the white LEDs associated with the red and green LEDs remain OFF. On the same day, the temperature is below, the humidity is higher, and the water level is also below the predefined value. Therefore, the blue and orange LEDs remain OFF and both white LEDs associated with the blue and yellow LEDs remain ON. This assembled LED status can clearly disclose to the user that it is needed to apply water and fertilizer. On the 4th day of the second subtable, the temperature, humidity, soil moisture, water level, and pH level are observed. In this case, the temperature is lower where the humidity, soil moisture, and pH are higher than the threshold value, so the condition of turning ON the red, green, and orange LEDs is not satisfied where the water level is higher than the threshold value, so the blue LED is satisfied. Therefore, the red, green, and orange LEDs are turned off while the blue LED remains on. This notifies the person that it needs to be drained of water. Other information for 7-days monitoring of the intelligent agriculture process is tabulated in the first subtable of Table I. Similarly, other information about aquaculture, livestock, air quality, weather, and health

IEMTRONICS 2022 (International IOT, Electronics and Mechatronics Conference)

TABLE I

ALL SENSING INFORMATION FOR 7 DAYS IN THE FIELD OF AGRICULTURE, AQUACULTURE, LIVESTOCK, AIR-QUALITY, AND WEATHER-MONITORING MANNER BY USING OUR PROPOSED MODEL

F1: AGRICULTURE MONITORING (Subtable 1)												
Day	Temp °C	Humidity %	Soil Moisture	Water Level	pH Level	RED LED	BLU LED	GRE LED	ORG LED	SMS & Email (F1)		
SUN	28	50	27	44	4.5	ON	OFF	ON	OFF	F1-Water and Fertilizer Supply		
SAT	32	45	42	52	6.1	OFF	OFF	OFF	OFF	F1-Good Day		
FRI	45	18	46	40	6	OFF	OFF	OFF	ON	F1-Adjust Temperature		
THU	25	35	48	70	6.8	OFF	ON	OFF	OFF	F1-Water Drain		
WED	22	43	52	47	7.3	OFF	OFF	OFF	OFF	F1-Good Day		
TUE	42	26	25	20	4.8	ON	OFF	ON	ON	F1-Water & Fertilizer apply, and Adjust temp		
MON	18	70	56	65	8.4	OFF	ON	OFF	OFF	F1-Water Drain		
F2: AQUACULTURE MONITORING (Subtable 2)												
Day	Temp °C	Humidity %	pH Sensor	DO Sensor	Water Level	Turbidity %	RED LED	ORG LED	YEL LED	GRE LED	BLU LED	SMS & Email (F2)
SUN	28	50	6.9	3.1	MID	45.1	OFF	OFF	OFF	OFF	OFF	F2-Good Day
SAT	32	45	7.1	3.4	HIGH	45.2	OFF	OFF	OFF	ON	OFF	F2-Water drain
FRI	22	43	6.6	3.2	MID	48.6	OFF	ON	OFF	OFF	ON	F2-Adjust pH & Water Filter
THU	25	35	6.8	2.8	LOW	44.3	OFF	OFF	OFF	ON	OFF	F2-Water Supply
WED	45	18	6.9	2.1	MID	45.2	ON	OFF	ON	OFF	OFF	F2-Shedding & Adjust Growth
TUE	42	26	5.4	2.9	MID	49.6	OFF	ON	OFF	OFF	ON	F2-Adjust pH & Water Filter
MON	18	70	7.0	2.7	MID	45.5	OFF	OFF	OFF	OFF	OFF	F2-Good day
F3: LIVESTOCK MONITORING (Subtable 3)												
Day	Temp °C	Humidity %	Gas Index	Water Level	Count	RED LED	BLU LED	GRE LED	ORG LED	SMS & Email (F3)		
SUN	28	50	60	LOW	8	OFF	OFF	ON	ON	F3-Adjust Water and Count		
SAT	32	45	66	MID	10	OFF	OFF	OFF	OFF	F3-No Need Monitor		
FRI	45	18	47	MID	10	ON	ON	OFF	OFF	F3-Adjust Temperature and Gas		
THU	25	35	56	LOW	10	OFF	OFF	ON	OFF	F3-Adjust Water		
WED	22	43	45	MID	5	OFF	ON	OFF	ON	F3-Adjust Gas and Count		
TUE	42	26	62	LOW	10	ON	OFF	ON	OFF	F3-Adjust Temperature and Water		
MON	18	70	66	MID	10	OFF	OFF	OFF	OFF	F3-Good day		
F4: AIR QUALITY MONITORING (Subtable 4)												
Day	Temp °C	Humidity %	Gas Index	pH Level	Particle Index	RRD LED	BLU LED	GRE LED	ORG LED	SMS & Email (F4)		
SUN	28	50	66	6.9	67	OFF	OFF	OFF	OFF	F4-Perfect Air Quality		
SAT	32	45	66	7.2	69	OFF	OFF	OFF	OFF	F4-Perfect Air Quality		
FRI	45	18	45	6.5	40	ON	ON	ON	ON	F4-Adjust Temp, Gas, pH, and Particle		
THU	25	35	53	7.0	60	OFF	OFF	OFF	OFF	F4-Perfect Air Quality		
WED	22	43	70	7.1	66	OFF	OFF	OFF	OFF	F4-Perfect Air Quality		
TUE	42	26	48	6.9	38	ON	ON	OFF	ON	F4-Adjust Temp, Gas, and Particle		
MON	18	70	56	7.5	56	OFF	OFF	ON	OFF	F4-Adjust pH		
F5: WEATHER MONITORING (Subtable 5)												
Day	Temp °C	Humidity %	Wind Velocity	Wind (D)	Weather Status	RED LED	GRE LED	SMS & Email (F5)				
SUN	28	50	12km/h	W	SUNNY	OFF	OFF	F5-Normal Day				
SAT	32	45	21km/h	SW	SUNNY	ON	OFF	F5-Sunny Day, Umbrella Needed				
FRI	45	18	5km/h	S	WARM	ON	OFF	F5-Sunny Day, Umbrella Needed				
THU	25	35	9km/h	N	CLOUDY	OFF	OFF	F5-Normal Day				
WED	22	43	6km/h	N	RAINY	OFF	OFF	F5-Normal Day				
TUE	42	26	11km/h	W	SUNNY	ON	OFF	F5-Sunny Day, Umbrella Needed				
MON	18	70	17km/h	N	STORM	OFF	ON	F5-Rainy Day, Umbrella Needed				

TABLE II
ALL SENSING INFORMATION FOR 7 DAYS IN THE FIELD OF HEALTH-MONITORING MANNER BY USING OUR PROPOSED MODEL

F6: HEALTH MONITORING (Subtable 6)											
Day	Temp °C	Humidity %	Pulse Rate	Body Temp	Blood Sugar	Blood Pressure	RED LED	BLU LED	GRE LED	ORG LED	SMS & Email (F6)
SUN	28	50	78	98.2	4.2	125	OFF	OFF	OFF	OFF	F6-Good Health
SAT	32	45	82	98.5	4.5	130	OFF	OFF	OFF	OFF	F6-Good Health
FRI	45	18	86	98.3	5.7	127	OFF	OFF	ON	OFF	F6-Adjust Blood Sugar
THU	25	35	66	100	6.5	115	OFF	ON	ON	ON	F6-Adjust Body Temp, Blood Sugar, and BP
WED	22	43	110	101	5.2	145	ON	ON	OFF	OFF	F6-Adjust Pulse rate and Body Temp
TUE	42	26	80	99	4.3	166	OFF	OFF	OFF	OFF	F6-Good Health
MON	18	70	120	98.4	3.7	180	ON	OFF	ON	ON	F6-Adjust Pulse rate, Blood sugar, and BP

monitoring systems are depicted in the second, third, fourth, fifth subtable of Table I, and Table II accordingly.

VI. CONCLUSION AND FUTURE WORK

Different field-deployed sensors are used in our IoT-based consolidated farming framework to measure diverse environmental and meteorological information on farms. For agriculture, aquaculture, livestock, air quality monitoring, weather monitoring, and health monitoring, we deploy temperature, humidity, wind velocity, soil moisture, pH, turbidity, DO, gas, particles, blood pressure, sugar level, BP, and human temperature sensors. A gateway called WSN connects all of the sensors from various fields. IoT cloud servers accept all sensing data from this gateway and transmit it on to a website as well as the decision-making sector, which is overseen by node MCU. Decisions are made in the decision-making sector based on specified values that are sent to the end user via GSM, and E-mail. The user can simply monitor all important information from any location on the planet and carry out physical tasks in smart and interconnected farming areas. This IoT-based consolidated combination farm system with access to data visualization via websites named WCSFM may help us increase production by requiring less labor and allowing us to operate more efficiently. It reduces the cost of crop production while speeding up the process. We intend to use Machine Learning (ML) to analyze vast amounts of farm field monitoring data in the future.

REFERENCES

[1] V. Kumar *et al.*, "Importance of weather prediction for sustainable agriculture in bihar, india," *Archives of Agriculture and Environmental Science*, 2017.

[2] "Global seafood," accessed: 2022-04-15. [Online]. Available: <https://www.globalseafood.org/blog/what-is-aquaculture-why-do-we-need-it/>

[3] "The livestock sector and the world bank," accessed: 2022-04-15. [Online]. Available: <https://www.worldbank.org/en/topic/agriculture/brief/moving-towards-sustainability-the-livestock-sector-and-the-world-bank>

[4] T. Godish and J. S. Fu, *Air quality*. CRC Press, 2019.

[5] "Agriculture growth reduces poverty," accessed: 2022-04-15. [Online]. Available: <https://www.worldbank.org/en/news/feature/2016/05/17/bangladeshs-agriculture-a-poverty-reducer-in-need-of-modernization>

[6] Y. Rahut, R. Afreen, D. Kamini, and S. S. Gnanamalar, "Smart weather monitoring and real time alert system using iot," *International Research Journal of Engineering and Technology*, vol. 5, no. 10, pp. 848–854, 2018.

[7] L. Parra, J. Rocher, S. Sendra, and J. Lloret, "Smart irrigation tube: a wireless system for water detection in precision agriculture."

[8] C. Dupont, P. Cousin, and S. Dupont, "Iot for aquaculture 4.0 smart and easy-to-deploy real-time water monitoring with iot," in *2018 Global Internet of Things Summit (GIoTS)*. IEEE, 2018, pp. 1–5.

[9] H. Gupta, D. Bhardwaj, H. Agrawal, V. A. Tikkiwal, and A. Kumar, "An iot based air pollution monitoring system for smart cities," in *2019 IEEE International Conference on Sustainable Energy Technologies and Systems (ICSETS)*. IEEE, 2019, pp. 173–177.

[10] L. O. Tedeschi, P. L. Greenwood, and I. Halachmi, "Advancements in sensor technology and decision support intelligent tools to assist smart livestock farming," *Journal of Animal Science*, vol. 99, no. 2, p. skab038, 2021.

[11] A. Rahaman, M. M. Islam, M. R. Islam, M. S. Sadi, and S. Nooruddin, "Developing iot based smart health monitoring systems: A review." *Rev. d'Intelligence Artif.*, vol. 33, no. 6, pp. 435–440, 2019.

[12] M. M. S. Maswood, U. K. Dey, M. A. Uddin, M. M. I. Mamun, S. S. Sonia, M. Akter, and A. G. Alharbi, "A novel website development for weather notification system using smart umbrella based on internet of things," in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2020, pp. 0129–0133.

[13] M. A. Uddin, U. K. Dey, S. A. Tonima, and T. I. Tusher, "An iot-based cloud solution for intelligent integrated rice-fish farming using wireless sensor networks and sensing meteorological parameters," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2022, pp. 0568–0573.

[14] P. Susmitha and G. S. Bala, "Design and implementation of weather monitoring and controlling system," *International journal of Computer applications*, vol. 97, no. 3, 2014.

[15] M. M. S. Maswood, U. K. Dey, M. A. Uddin, A. K. Ghosh, M. M. I. Mamun, S. S. Sonia, and A. G. Alharbi, "A novel approach to design the cell of lte cellular network to improve the call quality and coverage area in rural and suburban area," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 1442–1447.

[16] M. M. S. Maswood, U. K. Dey, S. Akter, M. A. Uddin, M. M. I. Mamun, S. S. Sonia, and A. G. Alharbi, "Improving system performance in indoor environment by designing femtocell considering interference and mobility management," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 1204–1209.

[17] B. S. Rao, K. S. Rao, and N. Ome, "Internet of things (iot) based weather monitoring system," *international journal of advanced research in computer and communication engineering*, vol. 5, no. 9, pp. 312–319, 2016.

Artificial Magnetic Conductor Unit Cell Design Using Machine Learning Algorithms

Tasfia Nuzhat

Department of Computer Science and Engineering
Chittagong Independent University
Chittagong, Bangladesh

Md Nazmul Hasan

Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, Canada
Email: mnazmulh@ieee.org

Abstract—Commercial electromagnetic (EM) simulator tools solve complicated Maxwell’s equations to design and optimize electromagnetic devices, which is computationally expensive and time consuming. There is a dire need to solve complex electromagnetic problems with least amount of computational resources in a short time. This work proposes the application of machine learning techniques in design process of electromagnetic problem. For the proof of concept, we demonstrated an optimum design process of an artificial magnetic conductor, which is a metasurface unit cell, by applying machine learning algorithms namely, artificial neural network (ANN), k-nearest neighbor (KNN), support vector machine (SVM), extreme gradient boosting (XGBoost), and least absolute shrinkage and selection operator (LASSO). The performances of these machine learning optimization models were evaluated on the test data set based on root mean squared error (RMSE) values. To the best of our knowledge, this is the first work that yields an excellent match with the original EM results from a commercial simulator tool with very small training dataset. Thus, it obviates the need of using computationally expensive and time-consuming electromagnetic simulators and massive training datasets for data-driven design approach of complex electromagnetic problems.

Index Terms—AMC, metasurface, XGboost, machine learning, gradient boosting

I. INTRODUCTION

Artificial magnetic conductor (AMC) belongs to the category of metasurfaces [1]. Metasurfaces are printed 2D electromagnetic surfaces consisting of periodically arranged unit cells. AMC can be used as the ground plane of microstrip patch antennas [2]. The advantage of using AMC as a ground plane is that it generates an image current in-phase of the radiating current, unlike a ground plane. Image current in antenna ground plane is 180° out of phase with respect to radiating current which can create ripples in the overall radiation pattern of the antenna, leading to bad design. Additionally, AMC can be used as a superstrate above the antenna to increase gain and reduce radar cross-section. Optimizing AMC unit cells by using a conventional CAD-based EM simulator requires time-consuming and computationally expensive efforts. Recently,

machine learning and deep learning techniques have been applied to solve different EM problems such as designing, and optimizing EM problems such as in designing metasurfaces [3] [4]. In [5], deep learning techniques have been applied to compute code in programmable metasurfaces. In [6], an AMC surface was designed by using particle swarm optimization (PSO) but the authors did not compare the performance between conventional EM simulation and PSO approach. In [7], a deep neural network (DNN) was proposed that predicts the corresponding scattering parameters given a geometrical feature of metasurface. However, the DNN model proposed in [7] requires considerable large amount of training dataset from prior computationally intensive EM simulations. At present, the solution to the design of complex EM problems with a small amount of training dataset is still challenging which requires careful selection of the appropriate machine learning approach. Often, trade-offs among the training dataset size, speed and accuracy are found in the existing literature.

In this work, we propose machine learning (ML) techniques for predicting the resonance frequency of an AMC unit cells given its geometrical features as input with very little amount of training dataset with significantly fast speed. We have considered artificial neural network (ANN), k-nearest neighbor (KNN), support vector machine (SVM), extreme gradient boosting (XGBoost), and least absolute shrinkage and selection operator (LASSO) algorithm in this work. We have chosen ANN, KNN, SVM and LASSO as these are some common ML algorithms that have been applied in the field of electromagnetic problems found in literature [1] - [7]. We have also decided to experiment with XGBoost as this is one of the most popular and powerful ensemble techniques in recent days and also one of the most successful algorithms in Kaggle competitions.

Moreover, we have compared the performance of these algorithms and chose the one that gives the most accurate prediction. The results have shown good agreement with the original EM simulation results. The main goal of our work is to reduce the time and computational expense in electromagnetic problem such as metasurface design. To the best of our

knowledge, this is the first demonstrative work that can predict the resonance frequency of artificial magnetic conductor unit cell, even with a smaller training datasets, thereby reducing the computational burden for electromagnetic researchers. Section II describes the necessary theoretical background. The ML algorithms used in this works are described in section III and the results are discussed and analyzed in section IV followed by conclusion in Section V.

II. IMPLEMENTATION

A. Theoretical Background

To implement the idea, we proposed an AMC unit cell as shown in Fig.1. AMC unit cell consists of a dielectric substrate, a metal ground plane beneath and a metal patch radiator with a rectangular outer ring on top.

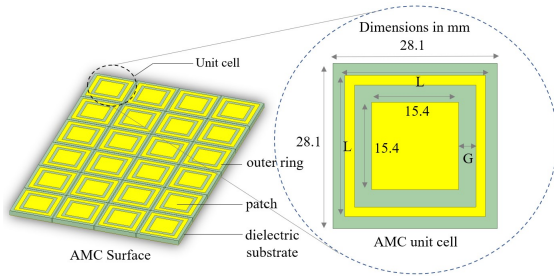


Fig. 1. Artificial magnetic conductor(AMC) and a unit cell

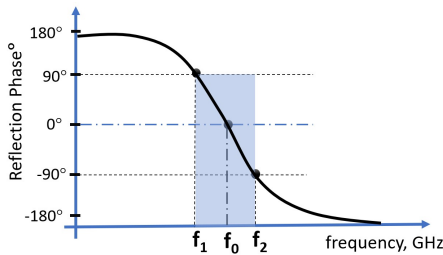


Fig. 2. Reflection phase curve of an AMC unit cell

Together, this unit cell acts as a resonator like an electrical RLC resonator circuit. This implies that the unit cell has its own resonance frequency around 5.4 GHz at which it behaves as a metasurface i.e. artificial magnetic conductor, to say more precisely. Moreover, when a free-space electromagnetic wave having a frequency equal to the resonance frequency of the AMC unit cell impinges on its surface, the AMC unit cell reflects the wave with a characteristic phase attributed to its functionality. This characteristic phase of the reflected wave is called reflection phase. At its resonance, AMC surface always reflects the electromagnetic wave with 0°, unlike a typical metal sheet which inverts the reflected wave phase by 180°. The resonance frequency of the AMC unit cell is determined by simulating the reflection phase with respect to a range of frequency as shown in Fig.2. The frequency where the reflection phase of the AMC unit cell drops at 0° is called its resonance frequency (f_0) as shown in Fig.2. The reflection

phase plot in Fig.2 also helps to determine the bandwidth ($f_1 - f_2$) of AMC unit cell ranging from +90° to -90°. Within the bandwidth, it behaves as an artificial magnetic conductor and for all other frequencies, it behaves like a typical metal conductor.

B. Interfacing of Electromagnetic simulation and Machine Learning

The capacitive coupling between the metal patch and the outer ring is controlled by the gap width, G . Additionally, the outer ring length, L has an effect on the frequency bandwidth of AMC surface. Our main focus is to predict the resonance frequency of AMC by taking the design parameters L and G as inputs. AMC unit cell with different values of design parameters was first simulated using commercial EM simulator ANSYS HFSS software [8]. The results of the EM simulation with design parameters were then recorded in a .csv file. Initially, the .csv file contains hundreds of entries of EM simulations which were then filtered out to get the main dataset. Since our main interest was in resonance frequency, we retained the dataset relevant to resonance frequency and discarded all other unnecessary frequency bands from EM simulation records. Our main dataset contains total of 129 data points of resonance frequencies for different values of design parameters, L and G . Afterwards, ML algorithms were trained on this dataset to predict the desired resonance frequency of AMC unit cell. We used python programming language on Jupyter notebook to apply the ML algorithms. 80% of data points were used for training purpose and the remaining 20% of data points were used in testing. The performance of these ML models were evaluated on the test set based on root mean squared error (RMSE) values. The following section contains the main ideas of five ML algorithms that are used in this work.

III. MACHINE LEARNING ALGORITHMS

A. ANN

Artificial neural network (ANN) is a part of machine learning that works on the principle of biological neurons and is perhaps the most used technique in electromagnetic optimization problems. ANN model consists of one or more layers with a group of artificial neurons in each and can be trained using various algorithms like gradient descent, Levenberg-Marquardt algorithm etc. An output of a neuron i in a hidden layer is,

$$\sum_{i=1}^n (w_i \cdot x_i + b) \tag{1}$$

where w_i is the weight associated with the neuron, x_i is the input to the neuron and b is the bias of the neuron [9]. The final output is produced after applying an activation function. In the case of hidden layers, the most suitable activation function is ReLU. For the output layer, the linear activation function is used in regression problems. Our ANN model has an input layer, two hidden layers with 44 neurons in each, and an output layer with a single neuron. The numbers of layers

and neurons are chosen via *Kerastuner* which is a framework of tuning parameters of ANN in python. It is important to choose optimal parameters of the ANN model based on the dataset. *Adam optimizer* is used with a learning rate = 0.001, which is also chosen based on the results of the *kerastuner*. Our model is trained using 50 epochs and later the model is verified on the test dataset.

B. KNN

K-nearest neighbor (KNN) is one of the simplest supervised ML algorithms that have been applied successfully in the applications of electromagnetic fields such as antenna design and optimization. In KNN algorithm, we consider the nearest points to predict a new data point and the nearest points are chosen using a distance measure like Euclidean distance, and Manhattan distance. For a new data point, based on the value of K or nearest neighbors, the mean or weighted average of the outputs of all nearest points is computed in regression problem which is [10],

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i \tag{2}$$

It is an instance-based algorithm with a simple model structure. It is necessary to choose the optimal value of *k* based on the dataset. By choosing a proper *k* value, the problem of higher variance or bias can be avoided. An effective method of choosing the proper value of *k* is cross-validation and in our analysis, we chose the value of *k* based on the results of 5 fold cross-validation. The *k* value with minimum error rate is chosen. From Fig.3, it can be seen that *k* = 3 has the smallest error rate. In order to fit the KNN model in data, the function *KNeighborsRegressor()* from *sklearn* library of python is used.

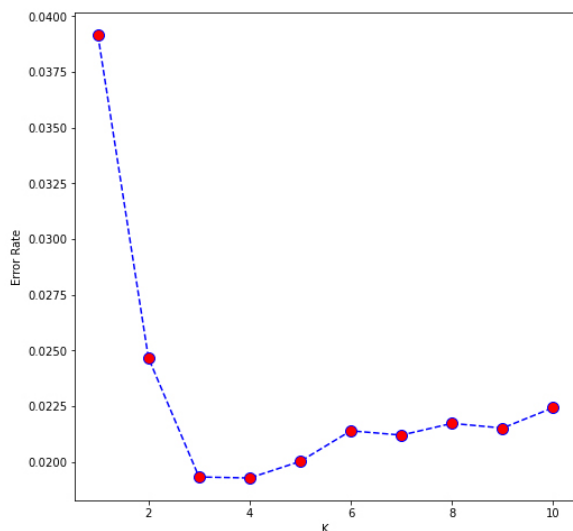


Fig. 3. k value vs. error rate

C. SVM

The support vector machine (SVM) is also an instance based algorithm that has gained popularity in designing and optimizing EM problems because of its great generalization property and high prediction accuracy. The data is transformed into higher dimensional feature space by using a non-linear Gaussian kernel function as following [11],

$$G(m, n) = e^{-\|m-n\|^2/2\sigma^2} \tag{3}$$

In real-world problems, it is often quite difficult to find an optimal hyperplane in a higher dimension data. The support vector regression algorithm applies the kernel function to choose an optimal hyperplane which contains maximum number of points in higher dimension. This hyperplane is used to predict the numerical output in the regression problem. Based on this hyperplane, a boundary line is chosen keeping a margin of tolerance as shown in Fig.4. The loss that is used by the algorithm is basically a uniform loss function which equally penalizes high and low misestimates [12]. In our analysis, we used the *GridSearchCV* method in python to choose the best hyper-parameters for the SVR model. We need to choose a proper kernel function, and the value of *C* which is a regularization parameter upon which the generalization power of the model depends and, the value of γ which is a parameter of Gaussian kernel function. Based on the result of *GridSearchCV* method the values of *C* and γ are decided to be 1 and 0.1, respectively.

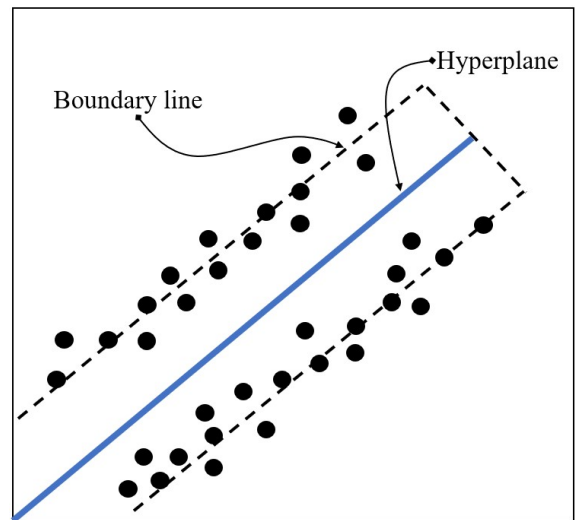


Fig. 4. Support vector machine architecture

D. XGBoost

Extreme gradient boosting (XGBoost) is a decision tree-based boosting ensemble technique well known for its execution speed and performance. It is a powerful method for prediction tasks because it gradually makes a weak learner into strong learner by correcting the prediction error using a differentiable loss function and gradient descent optimization

algorithm. Electromagnetic researchers always try to find algorithms in optimization problem that can be trained using a small dataset in less training time with better accuracy. XGBoost can be more effective than other techniques in that respect because it uses less computational expense and the training time is significantly less than that of the ANN and SVR model. Additionally, XGBoost algorithm helps to determine the most important features in a dataset and prevents over-fitting which is one of the most common problems in machine learning based models. In the case of larger datasets, XGBoost still can be a better choice because it can use all the cores of CPU by parallel processing [13] which makes it a more efficient and fast algorithm. For regression, primarily in XGboost, an average prediction is given by a base model and the residuals are calculated to form the decision trees. After calculating all the residuals, a similarity score is computed which is [14],

$$S_{m_{score}} = \frac{S_{residual}^2}{num_{residual} + \lambda} \quad (4)$$

where $S_{m_{score}}$ means similarity score, $S_{residual}$ is the sum of residuals, $num_{residual}$ is the number of residuals and λ is the regularization parameter. Based on the similarity score and the residuals, a decision tree is formed keeping a splitting criteria. Then, again the similarity scores are calculated for the leaves of both left and right hand side of the tree. After that, the gain of the tree is calculated which is [14],

$$G = (L_{score} + R_{score} - S_{m_{score}}) \quad (5)$$

where G is the gain, L_{score} is the left similarity score, R_{score} is the right similarity score, and $S_{m_{score}}$ is the previously calculated similarity score. The gain score is required for auto-pruning the decision tree which is essential to maintain the growth of the tree. The gain score is compared with a value called γ which is set at earlier. If the γ value is greater than the gain, the auto-pruning will take place. An output value is computed for the leaves of the tree which is, [14],

$$output = \frac{S_{residual}}{num_{residual} + \lambda} \quad (6)$$

The prediction of the tree for a new value will be [14],

$$N_{pred} = I_{pred} + output * l_r \quad (7)$$

where, N_{pred} is the new prediction, I_{pred} is the initial prediction and l_r is the learning rate. To implement XGBoost technique in our work, learning rate, $l_r = 0.2$, number of decision trees = 300 and maximum depth of the trees = 1 are chosen as optimal values. These parameters are also chosen via *GridSearchCV* method.

E. LASSO

Another regression analysis method that has been used in this work is least absolute shrinkage and selection (LASSO). LASSO algorithm is mainly well known for its regularization

and variable selection capability. A new term is added with the standard residuals sum squares in LASSO regression which is shrinkage penalty,

$$\frac{1}{n} \left(\sum_{i=1}^n (y_i - \sum_{j=1}^m w_j x_j + w_0)^2 \right) + \alpha \sum_{j=1}^m |w_j| \quad (8)$$

where α is the parameter in which the shrinkage depends [15]. The model sensitivity to data can be reduced by shrinkage, resulting in a robust model and the lasso algorithm tries to shrink the co-efficient (feature weight) to exactly zero which helps to remove features sequentially that are not important in predicting the value of target variables or which are unnecessary for the prediction task. In this method, only the important features whose co-efficients are not zero are chosen and the error can be significantly reduced. As tuning hyper-parameters for a ML model requires significant amount of time, in our analysis, we choose to implement *LassoCV* which has the capability to tune parameters and perform cross-validation for a lasso model without external *GridSearch* process. In our analysis, we found no coefficient becomes zero which means the two input variables have significant importance in predicting the output variable.

IV. RESULTS AND DISCUSSION

Using the combinations of optimal hyper-parameters, each ML model is trained to predict the resonance frequency of AMC unit cell for different design parameters, G and L . Table I contains some data points of test data predicted by different ML algorithms and also the simulation results from ANSYS HFSS. The first column of the table contains some random values of design parameters, G and L in mm. The combination of design parameters are taken in the range of $G = 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0$ and $L = 21, 21.5, 22, 22.5, 23, 23.5, 24$. The next column of the table contains the values of resonance frequency obtained from various ML algorithms and original ANSYS HFSS simulation.

TABLE I
COMPARISON OF PREDICTED VALUES OF ML ALGORITHMS AND HFSS

Parameter G, L(mm)	Resonance Frequency of AMC unit cell (GHz)					
	ANN	KNN	SVM	XGboost	LASSO	HFSS
1.0, 21.0	4.56	5.11	5.20	5.22	5.15	5.28
1.5, 21.5	4.73	5.12	5.23	5.24	5.19	5.13
2.0, 22.0	4.90	5.29	5.27	5.24	5.23	5.38
2.5, 22.5	5.06	5.33	5.30	5.36	5.28	5.41
3.0, 23.0	5.23	5.38	5.32	5.37	5.32	5.43
3.5, 23.5	5.40	5.41	5.33	5.41	5.36	5.38
4.0, 24.0	5.56	5.34	5.32	5.38	5.40	5.38

As observed from table I, for $G = 1.0$ mm and $L = 21.0$ mm, the resonance frequency of original HFSS simulation is 5.28 GHz. The values predicted by ANN, KNN, SVM, XGBoost and LASSO model are 4.56, 5.11, 5.20, 5.22, 5.15 GHz, respectively. The value that is closest to the 5.28 GHz is 5.22 GHz which is predicted by XGBoost model.

For, $G = 1.5$ mm and $L = 21.5$ mm, the ML predicted resonance frequencies are 4.73, 5.12, 5.23, 5.24, 5.19 GHz

for ANN, KNN, SVM, XGBoost and LASSO model, respectively. The targeted frequency is 5.13 GHz obtained from HFSS EM simulation. The point nearest to the targeted frequency is 5.12 GHz, which is achieved by KNN model.

For 5.38 GHz, the design parameter values are $G=2$ mm and $L=22$ mm. For this specified set of input parameter values and frequency, the KNN model predicted the closest value which is 5.29 GHz. Other predicted frequencies are 4.90, 5.27, 5.24, 5.23 GHz, given by ANN, SVM, XGBoost and LASSO model, respectively.

The next combination of design parameter values is $G= 2.5$ mm and $L= 22.5$ mm. For this set of values, the resonance frequency from HFSS simulation is 5.41 GHz. The closest value to this frequency is 5.36 GHz predicted by XGBoost model. The values obtained from ANN, KNN, SVM and LASSO model are 5.06, 5.33, 5.30, 5.28 GHz which are also not very far from the targeted value.

For $G = 3$ mm and $L=23$ mm, the original value of resonance frequency from HFSS simulation is 5.43 GHz. The KNN predicted value is 5.38 GHz which is again quite near to the frequency obtained from HFSS simulation. The values obtained from ANN, SVM, XGBoost and LASSO model are relatively close, which are 5.23, 5.32, 5.37, 5.32 GHz respectively.

For the next targeted frequency 5.38 GHz, the set of design parameter is $G = 3.5$ mm and $L=23.5$ mm. For this particular set of design parameters, the prediction of the LASSO model is quite accurate which is 5.36 GHz. The corresponding predicted frequencies obtained from ANN, KNN, SVM and XGBoost model are 5.40, 5.41, 5.33, 5.41 GHz, respectively.

The last combination of design parameter values is $G = 4$ mm and $L=24$ mm. The resonance frequency from HFSS simulation for these design points is again 5.38 GHz. In this case, the value predicted by XGBoost is the most precise which is exactly 5.38 GHz. The frequencies estimated by ANN, KNN, SVM and LASSO model are 5.56, 5.34, 5.32, 5.40 GHz, respectively.

In order to better compare the performance of these ML models with HFSS simulation, the predictions of the ML algorithms on the entire test data which contains 30 data points including those on table I were visualized as individual plots from Fig. 5a-e. The horizontal and vertical axis represents the total number of data points and resonance frequency (GHz), respectively. The values from the HFSS simulation have been indicated with the blue curve. The curves of the predicted values by ANN, KNN, SVM, XGBoost and LASSO models are represented by black, magenta, cyan, red and green color, respectively.

From Fig.5, the overall performance of each ML technique can be distinguished. Fig.5a, it is seen that the predictions of ANN model are diverged from the actual simulated values. As our dataset has only 129 data points, this may be a possible reason of poor performance of ANN model compared to other ML techniques because often more data is required for training a neural network.

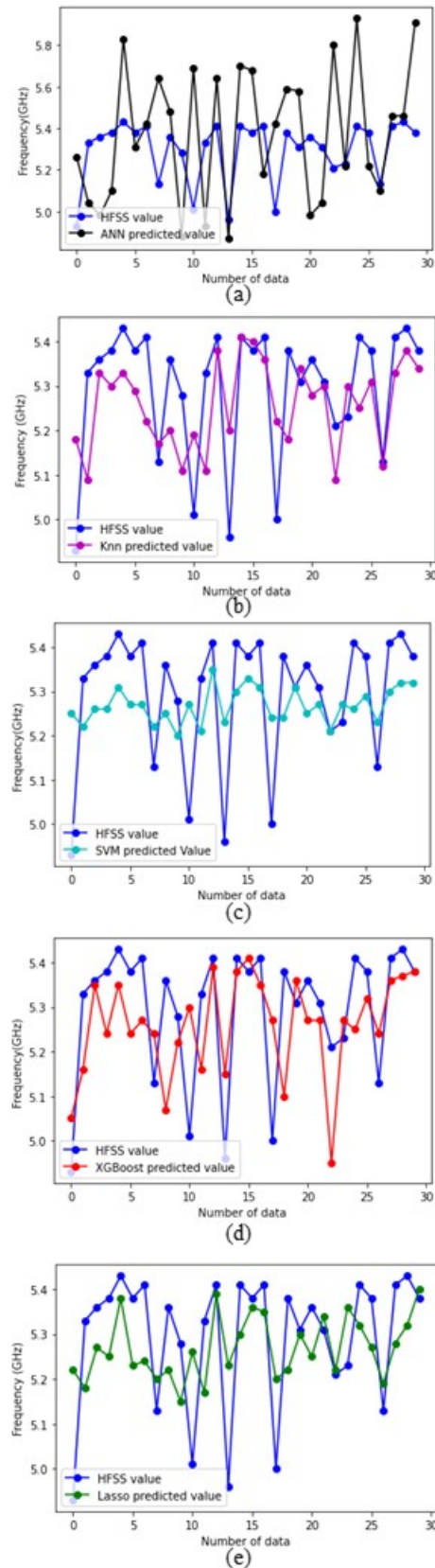


Fig. 5. Simulated and predicted resonance frequency (a) ANN vs. HFSS (b) KNN vs. HFSS (c) SVM vs. HFSS (d) XGBoost vs. HFSS (e) LASSO vs. HFSS

The results obtained from the KNN model in Fig.5b is quite satisfactory as the blue curve of HFSS is closely followed by the magenta curve of KNN model. As mentioned earlier, $k = 3$ is chosen as the optimal value due to the minimum error. The results of KNN is more precise than ANN in our analysis because KNN does not require large datasets compared to ANN model. Despite having a simple model structure, KNN showed better performance in predicting the resonance frequency of AMC unit cell.

As stated before, we tuned the parameters of SVM model with the help of *GridSearch* method and the optimal values of C and γ are chosen as 1 and 0.1, respectively. It is observed from the Fig.5c that, the overall predictions of SVM is not that much specific compared to other ML techniques such as KNN or XGBoost as the cyan curve of SVM does not follow the curve of HFSS simulated values exactly. As our dataset only contains two design parameters, namely G and L , we can say it has low dimension with a greater value of observations. SVM usually provides better result than KNN when the dimension of data is higher than the number of observations.

The XGBoost model also shows satisfactory results like KNN as shown in Fig.5d. Additionally, the performance of XGBoost is better than ANN and SVM model. XGBoost can work well on small datasets, unlike ANN. As XGBoost is an ensemble method, it is more robust and hence it has shown better prediction capability in our analysis.

From Fig.5e it can be seen that the LASSO model also provides acceptable results comparing to ANN and SVM model. For our small simulation dataset, LASSO outperforms ANN model because to get a generalized ANN model larger dataset is needed but lasso can work well even on small non-linear dataset for its regularization capability.

It can be seen from the above discussion that machine learning techniques can be successfully applied in optimization of EM problem that requires computationally expensive resources. However, performance of different machine learning techniques mainly depends on the dataset and in our experiment, all of the techniques that we applied is able to predict the resonance frequency of an AMC unit cell though there are slight differences in performance among these techniques. Also, the time taken by each ML technique to predict the resonance frequencies is significantly less than the conventional CAD-based EM simulation method e.g. in HFSS. For each sweep in HFSS simulation, the time took to find the resonance frequency was approximately 15 minutes. The HFSS simulation process was run in an Intel core-i7 processor with 32 GB RAM. On the other hand, the prediction time of each machine learning technique on the whole test data is just a few seconds on the same computing system. Table II compares the prediction time for each ML technique for the test data with HFSS which confirms that conventional CAD-based EM simulation is much slower than the proposed ML methods. It can also be seen from Table II that KNN and XGBoost model took less time in prediction than other three ML methods. Thus, the ensemble techniques like XGBoost

can be more efficient in solving complex electromagnetic problems. More generalized model can be obtained in future by collecting larger datasets for training.

TABLE II
COMPARISON OF PREDICTION TIME AMONG VARIOUS METHODS

Method Name	Time
ANN	84.36 seconds
KNN	62.36 seconds
SVM	118.23 seconds
XGboost	77.68 seconds
LASSO	86.76 seconds
EM simulation (HFSS)	15 minutes

V. CONCLUSION

In this work, we have applied machine learning algorithms to automatically predict resonance frequency of artificial magnetic conductor while taking its design parameters as input. We have used five ML algorithms which are ANN, KNN, SVM, XGBoost and Lasso and the performance of these ML models is superior compared to conventional commercially available EM simulator such as HFSS. All these ML techniques take considerably less time and computational power than traditional electromagnetic simulation methods. This shows the efficiency of ML techniques in complex electromagnetic problem.

ACKNOWLEDGMENT

This research work was supported by a grant from *Tensorbundle Lab*.

REFERENCES

- [1] D. Feng, H. Zhai, L. Xi, S. Yang, K. Zhang, and D. Yang, "A Broad-band Low-Profile Circular-Polarized Antenna on an AMC Reflector," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, 2017, doi: 10.1109/LAWP.2017.2749246.
- [2] J. Liu, J. Y. Li, J. J. Yang, Y. X. Qi, and R. Xu, "AMC-Loaded Low-Profile Circularly Polarized Reconfigurable Antenna Array," *IEEE Antennas and Wireless Propagation Letters*, vol. 19, no. 7, 2020, doi: 10.1109/LAWP.2020.2998493.
- [3] Y. Deng, S. Ren, K. Fan, J. M. Malof, and W. J. Padilla, "Machine Learning for Exotic Metasurfaces," in *International Conference on Infrared, Millimeter, and Terahertz Waves, IRMMW-THz, 2020*, vol. 2020-November. doi: 10.1109/IRMMW-THz46771.2020.9370973.
- [4] C. Liu, Q. Zhang, and T. J. Cui, "Deep Learning of Reflection Phase Prediction for Arbitrary Coding Metasurface Atoms," 2019. doi: 10.1109/COMPEN.2019.8778904.
- [5] T. Shan, X. Pan, M. Li, S. Xu, and F. Yang, "Coding Programmable Metasurfaces Based on Deep Learning Techniques," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 1, 2020, doi: 10.1109/JETCAS.2020.2972764.
- [6] A. Lalbakhsh, M. U. Afzal, K. P. Esselle, and S. Smith, "Design of an artificial magnetic conductor surface using an evolutionary algorithm," 2017. doi: 10.1109/ICEAA.2017.8065394.
- [7] C. C. Nadell, B. Huang, J. M. Malof, and W. J. Padilla, "Deep learning for accelerated all-dielectric metasurface design," *Optics Express*, vol. 27, no. 20, 2019, doi: 10.1364/oe.27.027523.
- [8] ANSYS Inc., <https://www.ansys.com>
- [9] Y. Sharma, H. H. Zhang and H. Xin, "Machine Learning Techniques for Optimizing Design of Double T-Shaped Monopole Antenna," in *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 7, pp. 5658-5663, July 2020, doi: 10.1109/TAP.2020.2966051.

- [10] L. Cui, Y. Zhang, R. Zhang, and Q. H. Liu, "A Modified Efficient KNN Method for Antenna Optimization and Design," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 10, 2020, doi: 10.1109/TAP.2020.3001743.
- [11] Z. Zheng, X. Chen, and K. Huang, "Application of Support Vector Machines to the Antenna Design," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 21, no. 1, 2011, doi: 10.1002/mmce.20491.
- [12] M. Awad and R. Khanna, "Support Vector Regression." In *Efficient Learning Machines*, pp. 67-80. Apress, Berkeley, CA, 2015.
- [13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-August-2016. doi: 10.1145/2939672.2939785.
- [14] XGboost Documentation by XGboost Developers, available online: <https://xgboost.readthedocs.io>
- [15] V. Roth, "The Generalized LASSO," *IEEE Transactions on Neural Networks*, vol. 15, no. 1, 2004, doi: 10.1109/TNN.2003.809398.

Development of an IoT-Based Low-Cost Multi-Sensor Buoy for Real-Time Monitoring of Dhaka Canal Water Condition

Ikbal Hasan

Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: 1722025@iub.edu.bd

Malobika Mukherjee

Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: 1822252@iub.edu.bd

Rumi Halder

Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: 1720884@iub.edu.bd

Farzana Yeasmin Rubina

Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: 1730796@iub.edu.bd

Md. Abdur Razzak

Department of Electrical and Electronic Engineering
Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: razzak@iub.edu.bd

Abstract—One of the most serious concerns for green globalization is water pollution. Being a riverine country, Bangladesh is focused to improve its water quality with its increasing population. Water quality monitoring in daily basis has always been a costly initiative in terms of time, human resources for large number of water bodies scattered throughout the country; the task become more daunting for ensuring sustainability after cleaning or improving water quality once. The consistency of drinking water must be monitored in real time in order to ensure its safety. Quality of water also needs to be addressed for irrigation and fish farming. In this paper, a multi-sensor buoy especially for canal water of Dhaka city has been designed for condition monitoring and predictive maintenance purpose and developed accordingly to facilitate the measurement of water quality in a regular interval; monitoring the water quality changes trends; reporting the data to the authority; and finally relating the trends to develop a model for the future. The design and implementation for this real-time water quality monitoring using the Internet of Things (IoT) is of a low-cost framework. The prototype has been deployed and validated with its operations towards the purpose. The system which consists of multiple sensors, is used to measure the water's physical and chemical parameters including the Temperature, pH, turbidity, and the TDS (total dissolved solids) of the water. The core controller will process and calculate values received from the sensors. Arduino is used as a central controller. Finally, the sensors data are sent to respective parties via short messages (SMS) and also viewed on internet using WI-FI system. The developed mechanism can also be implicated for similar water quality monitoring network including lakes and rivers.

Keywords—Internet of Things (IoT), pH sensor, Turbidity sensor, Temperature sensor, TDS sensor, Arduino, WI-FI module.

I. INTRODUCTION

In twenty-first century, there have been many developments but the trade-off with pollution, global warming, climate change and other issues. Clean drinking water is currently a huge issue due to the increased world's pollution. Water quality monitoring in real time is becoming more difficult as a result of global warming, limited water

supplies, and an increasing population, among other factors. Therefore, better methodologies for monitoring water quality parameters in real time are needed [1]. Potential of Hydrogen (pH) is one of the parameters of water quality which needs to be monitored regularly. It indicates whether the water is acidic or alkaline. The range of pH scale is 0 to 14. Pure water should have pH of 7 whereas a pH of less than 7 is acidic, more than 7 is alkaline. A pH value between 6.5 and 8.5 pH is acceptable for pure drinking. Turbidity is a measure dirtiness of water. Using turbidity sensor, we can measure the dirt present in the water. The chances of diarrhea disease increases with the drinking water having higher turbidity. The World Health Organization (WHO) establishes that the turbidity of drinking water should not be more than 5 NTU, and should ideally be below 1 NTU. The temperature sensor detects how much hot or cold the water is. Temperature will vary from time to time. Such as, it will be a higher temperature at day time whereas it will be a much lower temperature at night. A TDS meter is a small device which senses the Total Dissolved Solids in water. Dissolved ionized solids, such as salts and minerals, increase the conductivity of a solution. So, TDS meter measures the conductivity of the solution and estimates the TDS from that reading. Traditional water quality monitoring methods entail the manual collection of water samples from various locations.

Bangladesh Segment of the population, economic, and technological developments throughout the world have increased our capacity to consciously and inadvertently affect the environment in which we live in and that sustains us. Humans have emerged as the primary cause of environmental change. Our actions have an effect on the global environment, especially the climate. This, in turn, affects the volume and geographical and seasonal distributions of precipitation falling on watersheds, as well as the timing of runoff. We are changing the amount of our water sources on which we need to exist, both physically and economically, as a result of changes in landscapes, increased food and energy production,

and the influx of people into metropolitan areas. We rely on water not only for survival, but also for economic well-being. Water is used in the production of everything we make. There are no replacements, and while it is renewable, there is a limited supply. We have interrupted and over allocated river flow regimes, sometimes to the point of drying them out, as well as the lakes downstream. We have depleted groundwater aquifers, contaminated many, if not the majority, of our water bodies, including estuaries, coastal zones, and even seas, and destroyed ecosystems. We did this primarily to meet short-term economic aims, which may also not have included long-term ecological or even economic viability of the region or basin, as well as our own health. Water is becoming a more important policy problem on a global scale. The third World Water Development Report of the United Nations warns [1], the existing inequitable, unsustainable usage of water may have highly negative implications in an unprecedented way. Poor water management jeopardizes both economic progress and security. As a result, worries about a worldwide energy crisis have now been joined by worries about an impending global water catastrophe. The energy-water nexus, which includes the impacts of water usage on energy consumption as well as the consequences of energy production on water use, is garnering more attention. [2]. By 2050, the globe will have to feed and supply energy for an extra 2–2.5 billion people, in addition to meeting a billion people's existing unmet power demands. To fulfill the nutritional demands of this growing population, we must take into account the quantity of water used in the production of various items, particularly energy and food. Water managers have significant challenges in terms of energy and food security. Interactions and feedback loops connect energy production, water, food security, and climate change. Growing, transporting, processing, and trading food goods, for example, consume a lot of water and energy. The Comprehensive Assessment of Water Management in Agriculture provides a comprehensive examination [3]. This work demonstrates that in a business-as-usual scenario, water consumption in agriculture would almost double. The fast growth of the urban population is driving the construction of infrastructure and services like as road networks, water supply, sanitation, sewage, and drainage, as well as the city's expansion into floodplains and low-lying areas [2]. These urbanizations have a significant negative impact on water quality, particularly in humid tropical areas. [3].

However, in Dhaka, the relief-controlled structures of the region were successfully drained to the floodplain and low-lying area by streams and canals (local word 'KHALS') (JICA 1991) and eventually to the downstream via big rivers. The city's canals used to be the linking channels for the rivers that encircled the larger Dhaka area. [4]. Although there were many canals in the past, but the Institute of Water Modelling (IWM) has identified there exists 50 canals in the city at present; of which many has retained only the name due to the land fillings, waste disposal. Only 26 of them deemed recoverable by the authority [5].

Canal water needs to be monitored from time to time in order to keep the water clean since it's become very dirty over times as shown in Fig.1. There are many canals in our city which are under Dhaka City Corporation and all these canals are connected to rivers. We can be benefited from canal water in many scenarios. For instance, this water further can be used in irrigation, aqua culture, and also as drinking water. If not used in above cases, eventually this water goes back to river. Hence, in order to keep river water safe, it is of utmost importance to monitor water quality of canal water which further is connected to river. Monitoring water quality is extremely important and needs a system by which it can be monitored. For this purpose, a system has been made called a buoy with which four sensors will be connected which are pH sensor, temperature sensor, TDS sensor, and turbidity sensor. These sensors will sense the pH level, temperature, TDS value, and turbidity. A handsome number of data needs to be collected. We will take the collective measure. Then with the help of GSM module, via SMS, web, or apps, we will deliver the data collected to the authority/party. GSM module is being used because it is a remote data collection procedure.

This is done mainly so that river water does not get polluted. If the canal water is used for drinking water, it needs to be clean and TDS should be less than 50. If it is used for irrigation and aqua culture, and if it is not clean then it will have a huge impact on agriculture and the overall food cycle will be interrupted. Crops and fishes will die and living things will be affected badly. pH value should be 5.0 to 7.0 for irrigation and 6.5 to 9.0 for aqua culture. After monitoring, if the quality of water is bad then we can let Dhaka WASA or the Director of Environment to know about it and they will look into it. There can be many reasons of pollution of canal water such as garbage disposal, industrial waste, urban runoff, etc.

The purpose of this paper is dual. One is to present a complete assessment of recent work in the field of smart water quality monitoring in terms of application, communication technology utilized, sensor types used, and so on. The second goal is to propose a low-cost, less sophisticated smart water quality monitoring system that employs a controller with an integrated Wi-Fi module to monitor parameters such as pH, turbidity, and conductivity. The device also features an alarm feature that notifies the user when water quality metrics are out of range. The monitoring system analyzes data collected in real time and recommends appropriate corrective actions.



Fig. 1. A typical canal water condition of of Dhaka city

The necessity of user engagement in water quality management, as well as awareness of other aspects such as cleanliness, hygienic practices, storage, and disposal, are critical concerns in preserving water resource quality. Inadequate water quality promotes disease, kills people, and stifles socio-economic progress. Waterborne illnesses take the lives of around 5 million people globally [10-11, 13-16].

The organization of the paper is as follows, Section II presents the description of the proposed system, details of hardware and software requirements are provided in Section III. Section IV describes the hardware prototype. The obtained results are discussed in Section V and finally Section VI provides the conclusion.

II. PROPOSED SYSTEM

Figure 2 depicts the proposed system's block diagram, which monitors conductivity, turbidity, water level, and pH as critical parameters. The fundamental component of the IoT-enabled water quality monitoring system is a controller. The majority of IoT-based solutions employ a microcontroller with external Wi-Fi. Such design is not cost efficient, they are inefficient in terms of power, and result in complicated circuitry. The TI CC3200 in this work is a single chip microcontroller with an integrated Wi-Fi module and an ARM Cortex M4 core that can be linked to the closest Wi-Fi hot spot for internet access.

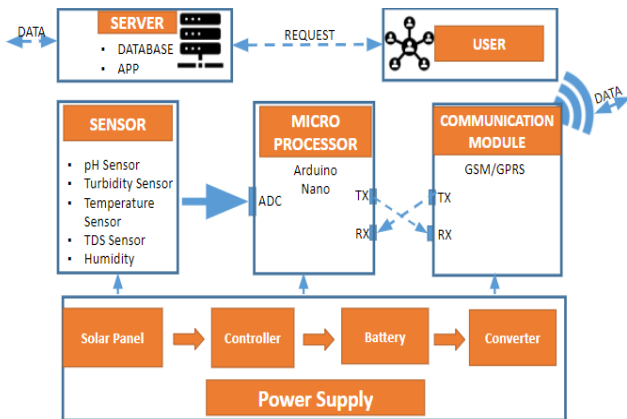


Fig. 2. Block diagram of the proposed system

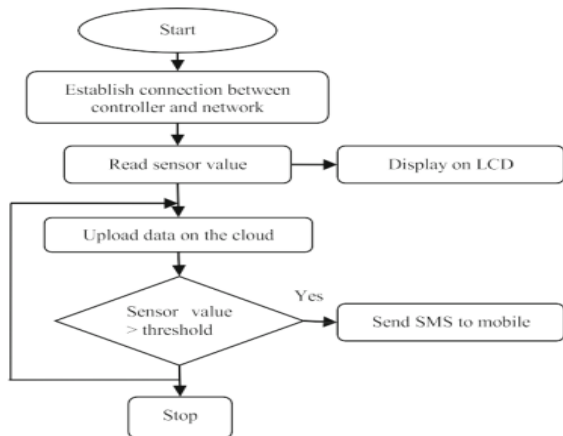


Fig. 3. Flow chart of the proposed system

As the suggested system is designed to monitor residential water quality, sensors are directly connected to the controller. Sensor characteristics such as conductance, clarity, water level, and pH are monitored by immersing the sensor in various water solutions. The measured parameters may be seen on an LCD display. The controller sends sensor data to the cloud. The threshold is determined in the system computing on WHO criteria. If the quantity exceeds the threshold, a message is delivered from the cloud to the user's mobile device. A mobile application was created that allows users to access the information acquired by each sensor in the cloud. This may be utilized by both water quality monitoring agencies and users.

III. SYSTEM DESIGN

A. Hardware components used in the project

- Arduino Nano
- pH sensor
- Turbidity sensor
- TDS Sensor
- Waterproof Digital temperature sensor: DS18B20
- Humidity Sensor Module: DHT22
- GSM module: Sim808

B. Software components used in the project

- Software: Arduino IDE and SMS marketing
- Web server: Ubidots and Cloud

TABLE I. SPECIFICATIONS OF THE HARDWARE COMPONENTS

Component Name	Specifications
DHT22	Measuring Range(0 to 14.00 pH), sensitivity (0.002 pH) Stability(0.02 pH/24 hrs), Temperature range(-5 to 95C degree Celsius) Pressure range(0 to 100 psi)
DS18B20	Usable temperature range: -55 to 125°C (-67°F to +257°F) Unique 64 bit ID burned into chip ±0.5° C Accuracy from -10° C to +85° C Usable with 3.0V to 5.5V power/data
GSM module	Quad-band 850 /900/1800/1900MHZ GPRS mobile station class B Supply voltage range 3.4V ~ 4.4V Supports 3.0V to 5.0V logic level
Arduino Nano	Operating voltage: 5 volts Input voltage: 6 to 20 volts DC per I/O pin: 40 m, DC for 3.3 V pin: 50 mA
pH sensor	Measuring Range(0 to 14.00 pH), sensitivity (0.002 pH). Stability(0.02 pH/24 hrs), Temperature range(-5 to 95C degree Celsius) Pressure range(0 to 100 psi)
Turbidity sensor	Operating Current: 40mA (MAX) Insulation Resistance: 100M (Min) Analog output: 0-4.5V Operating Temperature: 5°C~90 °C Storage Temperature: -10°C~90°C
TDS Sensor	Input Voltage: 3.3 ~ 5.5V Output Voltage: 0 ~ 2.3V Working Current: 3 ~ 6mA

IV. HARDWARE PROTOTYPE

A. Prototype testing

A water monitoring buoy system regularly measures a number of water factors, monitoring water quality. The system comprises of a few components, counting a data logger, buoy stage, solar control, temperature string, mooring equipment, & sensors. Rather than calling it system, we can call it a water quality monitoring buoy 'station'.

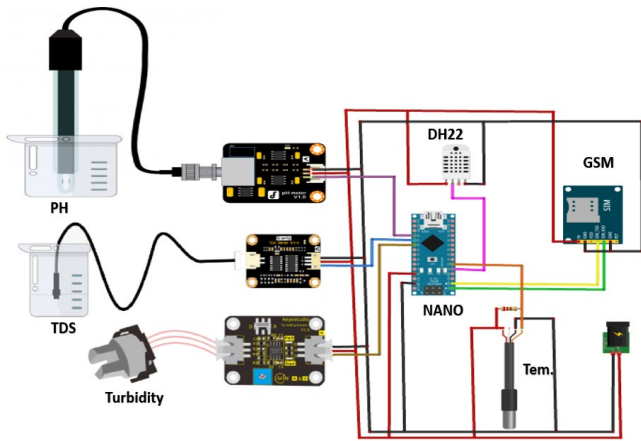


Fig. 4. Schematic of the proposed system

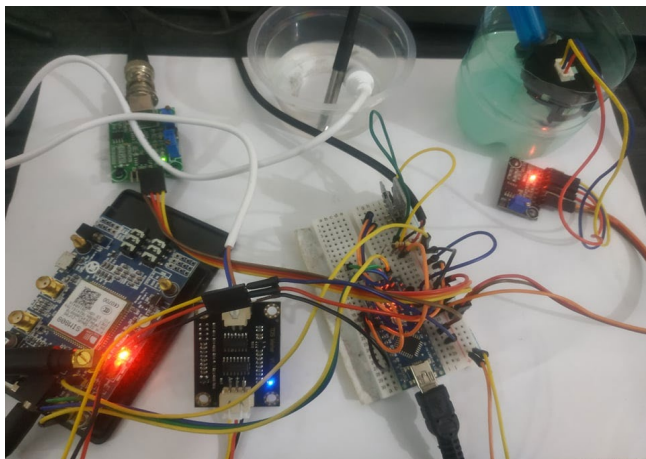


Fig. 5. Prototype testing in the lab

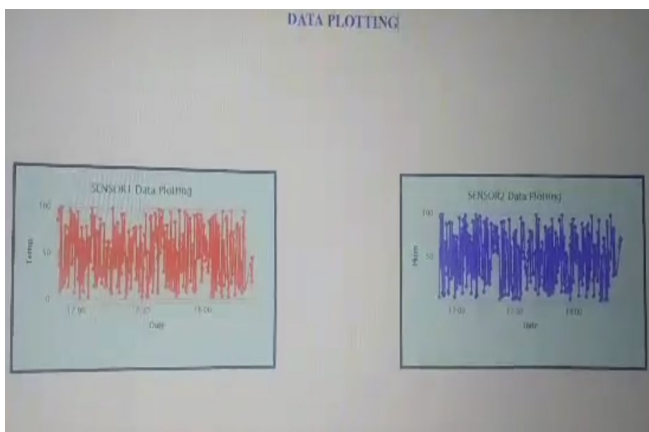


Fig. 6. Data hitting to the server during laboratory testing



Fig. 7. Prototype installation in the canal

Before installing in the canal, the prototype was built according to the circuit diagram shown in Fig.4 and tested in the laboratory after proper calibration as shown in Fig.5. Figure 6 depicts a graphical depiction of parameters observed and saved in the cloud over time. Experiments were then carried out by immersing the sensors in various water solutions collected on the campus grounds. the parameter values recorded for three distinct samples.

B. Installation

For installation, at first the required materials need to be lowered into the water. After placing the materials on the canal floor, the swimmers who will be installing the buoy come down to assemble some additional materials such as to install concrete block with chain and assembled with the rope of the buoy. In some cases, some markers can be placed and tied with the rope attached to the concrete. This is done to avoid any missing trail of buoy in case there is a change in the location of the anchor in canal floor or due to current. After all the materials are ready, the buoy is attached which has been tied with rope on water surface. Finally, the buoy is installed in local canal in the Bashundhara area of Dhaka city as shown in Fig.7.

Sensors are located at the bottom of the block buoy. Water quality metrics may be monitored using a variety of sensors. These sensors are submerged in the water to be examined, which can be either stored or flowing water. Sensors translate physical parameters into quantifiable electrical quantities, which are sent onto controllers through an optional wireless communication device. The controller's primary role is to read data from sensors, potentially process it, and communicate it to the application through appropriate communication technologies. The communication technology and parameters to be monitored are determined by the application's requirements. Data management tools, data analysis, and an alarm system based on monitored parameters are all part of the application.

C. Data Collection

There are three main ways in which data can be collected. Data can be collected via SMS, Apps, or Web. In this

research, both SMS and Web is used. The data sent by the controller is saved in the "Ubidots" cloud. "Ubidots" offers a framework for developers to collect data and transform it into valuable information. The capabilities it includes are real-time dashboard for analyzing data or controlling devices, as well as the ability to share data via public connections. Data saved in the cloud may be utilized for in-depth analysis. When the monitored parameter surpasses the threshold limit, the cloud is configured to send an SMS alert.

The system is linked to the IoT platform cloud by performing the following steps:

1. Use a mobile phone or a personal computer to connect to the access point by entering the SSID and password.
2. Wi-Fi is then used to link the controller to the access point.
3. Log in to the cloud platform and generate a token.
4. In the program, enter the token id.
5. Controller data is put into the cloud.
6. On the cloud platform, data may be examined.

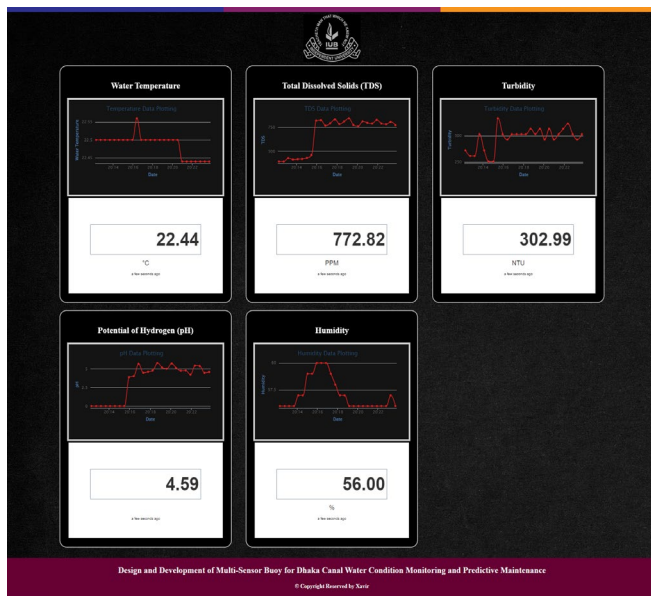


Fig. 8. Data collected through website

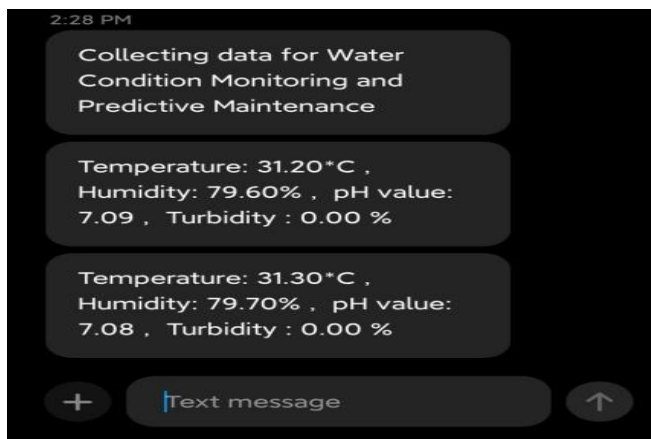


Fig. 9. Data sent via SMS

The buoy measures five parameters: conductivity, pH, turbidity, temperature, and water level. At first, all of the sensors' data have been calibrated and then they have been compared between raw data that we have got through message and actual data. After comparing all data, the accurate result or data has been found out. The configuration is linked to the Ubidots platform. The sensing data from developed prototype has been received via web as shown in Fig.8 and as text message via GSM module Sim808a as shown in Fig.9. The measured values are the compared to WHO drinking water quality requirements.

V. RESULT ANALYSIS

We have collected TDS, pH, turbidity and temperature data for 10 days and then average data has been plotted to see the difference as shown in Figs. 10, 11, 12 and 13, respectively. It has been observed that water condition changes slowly. As we can see, our very first day's average TDS data is 1500 ppm and our last days data is 1602ppm which means water is getting dirty. Within 10 days, it can be seen that the canal water shows big difference of pH which is in very bad condition. Values of turbidity is found within the standard values. However, temperature changes between 22.5°C to 24.5°C as expected since it may vary throughout the day and night.

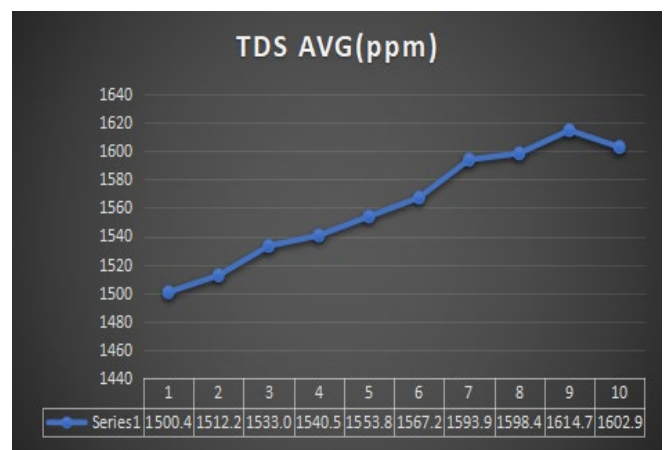


Fig. 10. Average TDS values in ppm

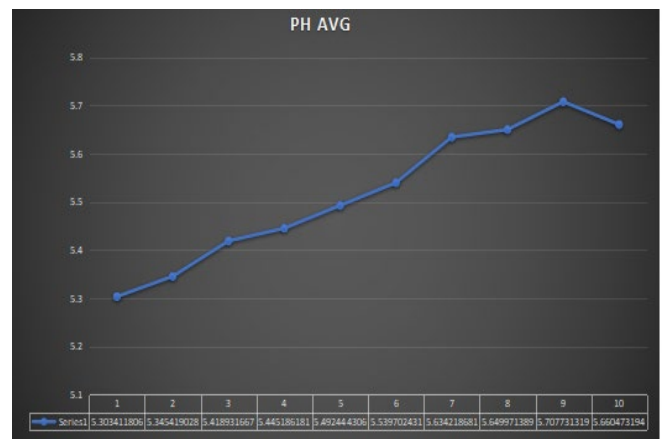


Fig. 11. Average values of pH over time

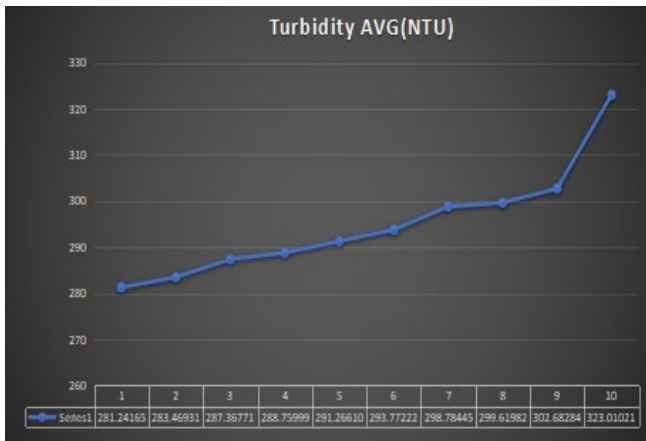


Fig. 12. Average values of turbidity in NTU

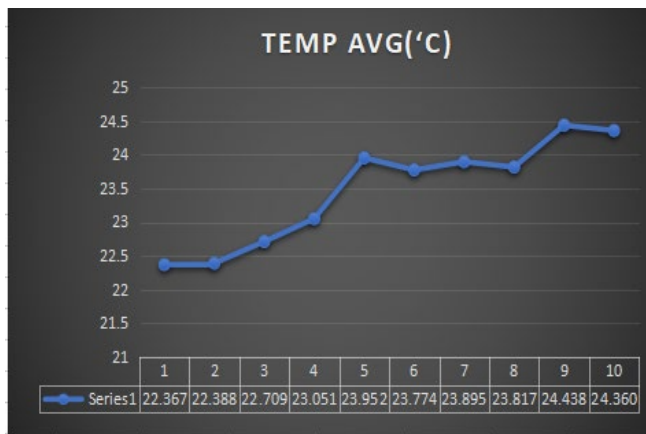


Fig. 13. Average values of temperature in degree C

TABLE II. COMPARISON BETWEEN STANDARD VALUE AND COLLECTED DATA VALUE

Sensor	Measured value	Standard value
Temperature	23.47 °C	25 °C
pH	5.51	7.85
TDS	294.94 ppm	280 ppm
Turbidity	1561.75 NTU	1000 NTU

The measured data are compared with that of the standard values as listed in Table II to understand the condition of water quality of Dhaka canal. From the above data, it can be observed that the values of TDS, Turbidity, and pH all tend to increase very slowly which correlates the water pollution situation in Dhaka city, which is worsening day by day.

VI. CONCLUSION

In this paper, a low-cost, IoT-based, real-time water quality monitoring system of Dhaka city canal water has been presented. The multi-sensor buoy, which costs approximately USD 200, measured and sent the water’s electrochemical data

to the appropriate parties via sms and web successfully. Its materialness was credited to its long length activity, adaptability, and reproducibility. The system used economically accessible electrochemical sensors to screen water quality boundaries precisely and show the outcome in the web utilizing GSM technology. This multi-sensor buoy can be used to monitor the water quality covering a huge region like lakes and other different waterways which need consistent observation because of its significance both to humanity and nature.

REFERENCES

- [1] United Nations World Water Assessment Program (UN WWAP), 2009.
- [2] Hoff, World Economic Forum Water Initiative (WEFWI), 2011, UN WWAP, 2011, 2012, 2014.
- [3] International Water Management Institute (IWMI), 2007.
- [4] Islam MS *et. al.*, “Changes in Wetlands in Dhaka City: Trends and Physico environmental Consequences”, *J. Life Earth Sci.* 5, pp.37-42 (2010).
- [5] Mahmud MS *et. al.*, “Remote Sensing & GIS Based Spatio-Temporal Change Analysis of Wetland in Dhaka City, Bangladesh”, *Journal of Water Resource and Protection*, 3, pp.781-787 (2011).
- [6] Chowdhury JU *et. al.* “Impact of land use change upon storm water drainage and wetlands in the eastern part of Dhaka city”, *Integrated Water Flow Model (IWFm)*, Bangladesh University of Engineering & Technology (BUET). Dhaka, Bangladesh (2011).
- [7] Mooring Buoy Installation, 2015.
- [8] N. Vijayakumar and R. Ramya, “The real time monitoring of water quality in IoT environment,” in 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015, pp.1-5.
- [9] Jianhua D, Guoyin W, Huyong Y, Ji X, Xuerui Z, “A survey of smart water quality monitoring system”, *Environ Sci Pollut Res* 22(7), pp.4893–4906 (2015).
- [10] Zahoor Ahmad, Rubab Khalid, Abubakr Muhammad, “Spatially Distributed Water Quality Monitoring using Floating Sensors”, 44th Annual Conference of the IEEE Industrial Electronics Society, 2018 (IECON 2018).
- [11] Yuling PEI, Qian WU, Xiaoyi YANG, “Design on integrated monitoring system for main waterway station”, *Proceedings of the 10th World Congress on Intelligent Control and Automation*, 2012.
- [12] Nenad Zoric, Sohail Sarang, Stefano Tennina, “Wireless Communication System for River Monitoring: An Energy-based Study”, 2020 International Conference and Exposition on Electrical And Power Engineering (EPE 2020).
- [13] Radityo Putro Wibisono, Novian Anggis Suwastika, Sidik Prabowo, Tri Djoko Santoso, “Automation Canal Intake Control System Using Fuzzy Logic and Internet of Things (IoT)”, 2018 6th International Conference on Information and Communication Technology .
- [14] Ubidots (2017) IoT platform | Internet of Things. <https://ubidots.com/>. Accessed 22 June 2021.
- [15] Hsia S. C., Hsu S. W. and Chang Y. J., “Remote monitoring and smart sensing for water meter system and leakage detection”, *IET Wireless Sensor Systems* 2, pp.402-408 (2012).
- [16] Ganjikunta RK, Devaki K, Gadamsetty V Mohan Ganesh, Avila J, Thenmozhi K, Rengarajan Amirtharaja, Padmapriya Praveenkumar, “Waste Contamination in Water – A Real-time Water Quality Monitoring System using IoT”, 2021 International Conference on Computer Communication and Informatics (ICCCI).

A High Gain Cascaded DC-DC Boost Converter for Electric Vehicle Motor Controller and Other Renewable Energy Applications

Md. Rezanul Haque

Department of Electrical and
Electronic Engineering

Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: hrezanulsets@iub.edu.bd

K. M. A. Salam

Department of Electrical and
Computer Engineering

North South University
Plot-15, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: kazi.salam@nsu.edu

Md. Abdur Razzak

Department of Electrical and
Electronic Engineering

Independent University, Bangladesh
Plot-16, Block-B, Bashundhara R/A,
Dhaka-1229, Bangladesh
E-mail: razzak@iub.edu.bd

Abstract—An impetuous popularity of electric vehicles (EVs) are of great concern in the field of automobile research. Since electric vehicles run through electric motor so it is important to have reliable DC-DC boost converter which can give stable required voltage for the motor controller. In this paper a modified cascaded DC-DC boost converter has been presented. The proposed boost converter can achieve high gain at intermediate duty cycle thereby lowering down the conduction loss, which will be ideal for EVs motor controller and other renewables energy applications. The analysis of the proposed converter is performed both in ideal and non-ideal conditions. The analysis of MOSFET's junction temperature has also been performed in order to confirm its viability for EV's motor controller and other renewable energy applications.

Keywords—DC-DC boost converter, Boost Converter, Electric vehicle, Pulse Width Modulation, Loss profile, Heat sink.

I. INTRODUCTION

The demand of Electric Vehicles (EVs) is increasing rapidly due to its efficiency and clean environmental aspects including the positive impact on climate change. In EVs the electric motor is one of the important system which gets power from low voltage battery system and in most cases the battery voltage is amplified to the required motor voltage to have smooth operations [1-2]. A high gain DC-DC boost converter is needed to full fill this requirement.

A DC-DC boost converter can be isolated and non-isolated type. In isolated boost converter high frequency transformer has been used which add extra cost and the converter will be bulky due to increased turns ratio at the secondary side. On the other hand, in the most non isolated boost converter the high gain has been achieved at high duty cycle which increases the conduction losses and decays the life cycle of the active switches those have been used in the converter [3-9].

High duty cycle conveys the active switch has to be turn on at most of the time and due to the internal resistance of the active switch an extensive amount of power will be lost which causes the thermal issues of the switch [10]. In most cases the high gain dc-dc boost converter might consist of multiple switched inductor and capacitor which get charged when the

switch conducts and discharge when the switch does not conduct. So, the high duty cycle means these passive elements get a smaller time to get discharged. As a result, these passive elements can be saturated which will be exaggerated for the converter. Therefore, high gain at intermediate duty cycle is a great concern in the field of EVs and other renewable applications.

For the switch, thermal cooling heat sink is a great option due to its robustness and cost effectiveness. The heat sink increases the area that is connected with the environment and helps the heat to flow easily from case to ambient [11-12]. Consequently, heat sink helps to maintain the temperature that is being generated in the junction of the MOSFET.

This paper presents a high gain dc-dc boost converter at intermediate duty cycle for EV's motor controller and other renewable applications. A detailed power loss calculations and active switch junction temperature analysis with three different cases are evaluated.

II. PROPOSED SYSTEM

The proposed cascaded dc-dc boost converter system has been depicted in Fig.1. The converter consists of two active switches. The switches are being controlled in complementary fashion which made the control system simple. In the first stage, the two inductors L4, L5, three diodes D9, D10 and D11 and one capacitor help to amplify the input voltage. In the second stage, two inductors L1, L2, three diodes D1, D2, D3 amplify the voltage. In the final stage, voltage again amplified via two capacitors C1, C2, and two D4, D5. Lastly, the L3, C3 filter provide the final output voltage.

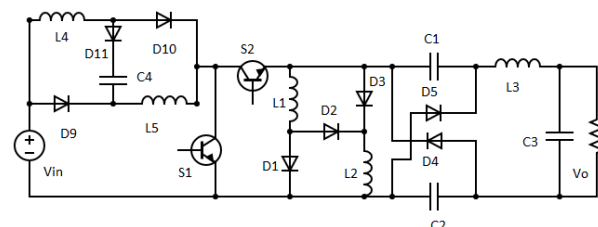


Fig.1. Proposed System

III. STEADY-STATE ANALYSIS

A. Ideal Condition

For this analysis all the components are assumed to be ideal and lossless, and the inductors and capacitors are chosen in such a way that can withstand ripple. The switching frequency is fixed, and the system is being analyzed in open loop CCM conditions. The Proposed converter has two different modes.

I. Mode 1

The converter will be analyzed at $0 < t < DT$ period. This time the switch one (S1) is closed and switch two (S2) is open. The equivalent circuit is shown in Fig. 2 (a). this time the L4 and L5 are being charged since the S1 in closed. The L1 and L2 are being discharged since S2 in open consequently the C1 and C2 are being charged via D5 and D4 besides L3 and C3 are responsible to have smooth continuous output current and voltage. The important equations for these modes are

$$v_{L1} = v_{L2} = v_{in} \tag{1}$$

$$v_{L3} = v_{in} + 2v_{C1} - v_0 \tag{2}$$

$$v_{L4} = v_{L5} = v_{in} - v_0 \tag{3}$$

$$v_{L4} = \frac{v_{in}-v_{C4}-v_0}{2} \tag{4}$$

II. Mode 2

The converter will be analyzed at $DT < t < T$ period. This time the switch two (S2) is closed and switch one (S1) is open. The equivalent circuit is shown in Fig. 2 (b). This time the inductors L4 and L5 are being discharged along with input voltage. The L1 and L2 are being charged through D1 and D3. Besides the C1 and C2 are being discharged to the load. L3 and C3 are responsible for continuous operations. The important equations for these modes are

$$v_{L1} = v_{L2} = -\frac{1}{2}v_{C1} \tag{5}$$

$$v_{L3} = v_{C1} - v_0 \tag{6}$$

$$v_{L4} = v_{L5} = v_{in} \tag{7}$$

$$v_{L4} = \frac{v_{in}-v_{C4}}{2} \tag{8}$$

After manipulating all the equations, the gain for the proposed converter appeared as

$$G = \frac{4D+3D'}{D'} \tag{9}$$

Figure 3 implicate that the theoretical vs simulated gain result when all the components are assumed to be ideal and lossless. It is clear that the proposed converter can give 5.5 to 9 times gain at intermediate duty cycle however 39 times gain can be achieved through this converter.

Figure 4 shows the theoretical gain comparison between the proposed converter and the references when all the components are assumed to be ideal and lossless. The results show that the proposed converter can give high gain at low, intermediate and high duty cycle compared with the references. Reference converters can give high gain at high duty cycle instead of low and intermediate duty cycle at ideal

conditions. Which clearly says that the conduction loss will be higher for the references whereas for the proposed converter conduction loss would be lower since high gain is being achieved at low and intermediate duty cycle.

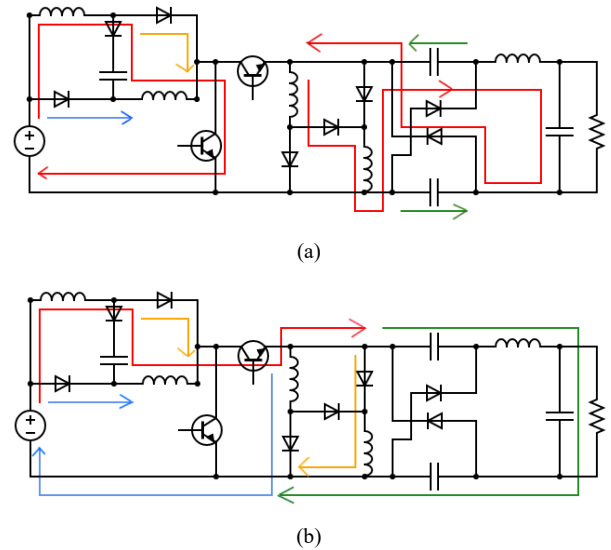


Fig.2. Working modes of the converter (a) Mode 1 and (b) Mode 2

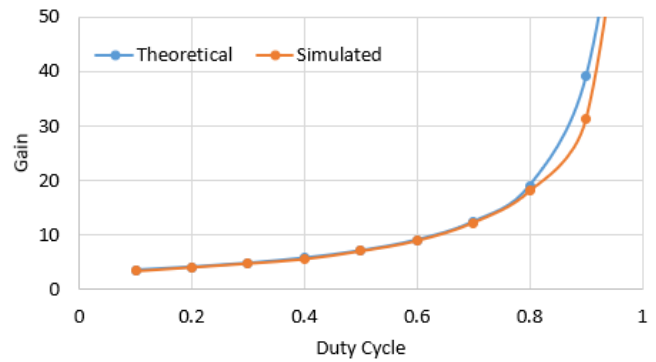


Fig.3. Ideal Gain of the proposed converter Theoretical vs Simulated.

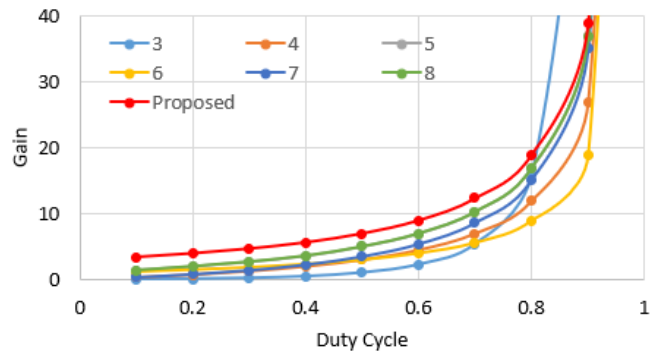


Fig.4. Theoretical Ideal Gain of the proposed converter vs references.

B. NON-IDEAL CONDITIONS

For this analysis all the components are assumed to be non-ideal and are related to real world with losses. The effect of the small ripple approximation is not excluded. For switch S1 and S2 MOSFET SUP60020E has been chosen. The switching frequency is fixed, and the system is being

analyzed in open loop continuous conduction mode (CCM) conditions. The Proposed converter has two different modes in non-ideal conditions.

I. Mode 1

The important equations for this mode are

$$v_{L4} = v_{in} - I_{RL4}R_{L4} - v_{D10} - I_{RD10}R_{D10} - v_o \quad (10)$$

$$v_{L5} = v_{in} - v_{D9} - I_{RD9}R_{D9} - I_{RL5}R_{L5} - v_o \quad (11)$$

$$v_{L1} = v_{in} - v_{s2} - I_{RS2}R_{S2} - I_{RL1}R_{L1} - v_{D1} - I_{RD1}R_{D1} \quad (12)$$

$$v_{L2} = v_{in} - v_{s2} - I_{RS2}R_{S2} - v_{D3} - I_{RD3}R_{D3} - I_{RL2}R_{L2} \quad (13)$$

$$v_{L3} = v_{in} - v_{s2} - I_{RS2}R_{S2} + 2v_{c1} - I_{RC1}R_{C1} - I_{RL3}R_{L3} - v_o - I_{RC2}R_{C2} \quad (14)$$

II. Mode 2

The important equations for this mode are

$$v_{L4} = v_{in} - I_{RL4}R_{L4} - v_{D10} - I_{RD10}R_{D10} - I_{S1}R_{S1} - v_{s1} \quad (15)$$

$$v_{L5} = v_{in} - v_{D9} - I_{RD9}R_{D9} - I_{RL5}R_{L5} - I_{S1}R_{S1} - v_{s1} \quad (16)$$

$$v_{L3} = -2v_{L1} - I_{RL1}R_{L1} - v_{D2} - I_{RD2}R_{D2} - I_{RL2}R_{L2} - v_{D5} - I_{RD5}R_{D5} - I_{RL3}R_{L3} - v_o - v_{D4} - I_{RD4}R_{D4} \quad (17)$$

$$v_{L1} = \frac{1}{2}(-I_{RL1}R_{L1} - v_{D2} - I_{RD2}R_{D2} - I_{RL2}R_{L2} - v_{D5} - I_{RD5}R_{D5} - I_{RC1}R_{C1} - v_{c1}) \quad (18)$$

$$v_{L1} = \frac{1}{2}(-I_{RL1}R_{L1} - v_{D2} - I_{RD2}R_{D2} - I_{RL2}R_{L2} - v_{c2} - I_{RC2}R_{C2} - v_{D4} - I_{RD4}R_{D4}) \quad (19)$$

After manipulating all the equations, the non-ideal gain is found as

$$G = \frac{1}{D} \times A \times (D + B + C) \quad (20)$$

where, $A = 1 - \frac{I_{RL4}R_{L4}}{v_{in}} - \frac{v_{D10}}{v_{in}} - \frac{I_{RD10}R_{D10}}{v_{in}} - \frac{I_{S1}R_{S1}}{v_{in}} D' - \frac{v_{s1}}{v_{in}} D'$ (21)

$$B = \left[\frac{2D}{D'} - \frac{2Dv_{s2}}{D'v_{in}} - \frac{2(I_{RS2}R_{S2})D}{D'v_{in}} - \frac{2(I_{RL1}R_{L1})D}{D'v_{in}} - \frac{2Dv_{D1}}{D'v_{in}} - \frac{2(I_{RD1}R_{D1})D}{D'v_{in}} - \frac{(I_{RL1}R_{L1})}{v_{in}} - \frac{v_{D2}}{v_{in}} - \frac{(I_{RD2}R_{D2})}{v_{in}} - \frac{(I_{RL2}R_{L2})}{v_{in}} - \frac{v_{D5}}{v_{in}} - \frac{(I_{RD5}R_{D5})}{v_{in}} - \frac{(I_{RC1}R_{C1})}{v_{in}} \right] (2D + D') \quad (22)$$

$$C = \left[-\frac{Dv_{s2}}{v_{in}} - \frac{(I_{RS2}R_{S2})D}{v_{in}} - \frac{(I_{RC1}R_{C1})D}{v_{in}} - \frac{(I_{RL3}R_{L3})D}{v_{in}} - \frac{(I_{RC2}R_{C2})D}{v_{in}} + \frac{(I_{RC1}R_{C1})D'}{v_{in}} - \frac{(I_{RD5}R_{D5})D'}{v_{in}} - \frac{(I_{RL3}R_{L3})D'}{v_{in}} - \frac{v_{D4}D'}{v_{in}} - \frac{(I_{RD4}R_{D4})D'}{v_{in}} \right] \quad (23)$$

Figure 5 implicate the non-ideal theoretical vs simulated gain result. It is clear that in non-ideal case the proposed converter can give 4 to 7 times gain in intermediate duty cycle however 13.3 times gain can be achieved through this converter. Besides due to non-ideal characteristics the gain become zero at max duty cycle.

Table I compare the gain and number of switches between the proposed converter and the references.

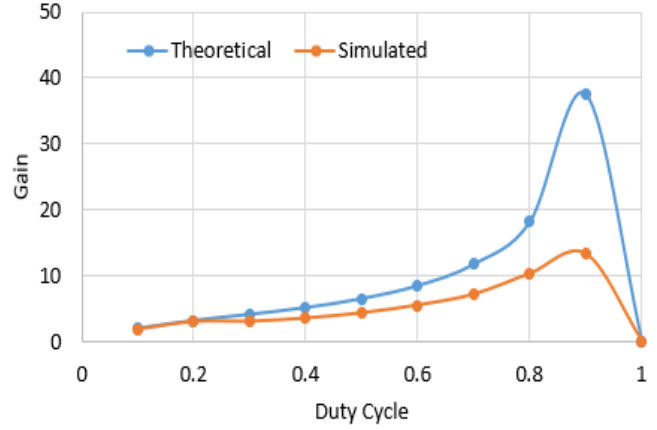


Fig.5. Non-Ideal Gain of the proposed converter Theoretical vs Simulated

TABLE I: IDEAL GAIN AND NUMBER OF SWITCH ANALYSIS WITH REFERENCES

Reference	Gain	Switch
3	$\frac{D^2}{(1-D)^2}$	1
4	$\frac{3D}{1-D}$	1
5	$\frac{1+3D}{1-D}$	2
6	$\frac{1+D}{1-D}$	2
7	$\frac{(3+D)D}{1-D}$	1
8	$\frac{1+3D}{1-D}$	4
Proposed converter	$G = \frac{4D + 3D'}{D'}$	2

IV. POWERLOSS CALCULATION OF MOSFET

To determine of a converter's reliability, the power loss of a switch must be in concern [10]. The power loss of MOSFET can be determined by the conduction loss and switching loss and shown in equation 22.

$$P_s = P_c + P_{sw} \quad (24)$$

The conduction loss can be calculated as

$$P_c = I_D^2 \times R_{DS(ON)} \times D \quad (25)$$

The switching loss can be calculated as

$$P_{sw} = P_{sw(ON)} + P_{sw(OFF)} \quad (26)$$

$$P_{sw(ON)} = I_D \times V_D \times t_{ON} \times \frac{1}{2} \times f_{sw} \quad (27)$$

$$P_{sw(OFF)} = I_D \times V_D \times t_{OFF} \times \frac{1}{2} \times f_{sw} \quad (28)$$

After manipulating equations 8 to 25 the power losses of MOSFETs have been analyzed and depicted in Fig.6. This graph also relates to efficiency (η) of the converter by assuming non-ideal characteristics. At low to intermediate duty cycle the power losses (P_{s1} and P_{s2}) in both S1 and S2 switches are around 3W to 9.5W. And with the increase in duty cycle the power loss also increases consequently decay the converter efficiency.

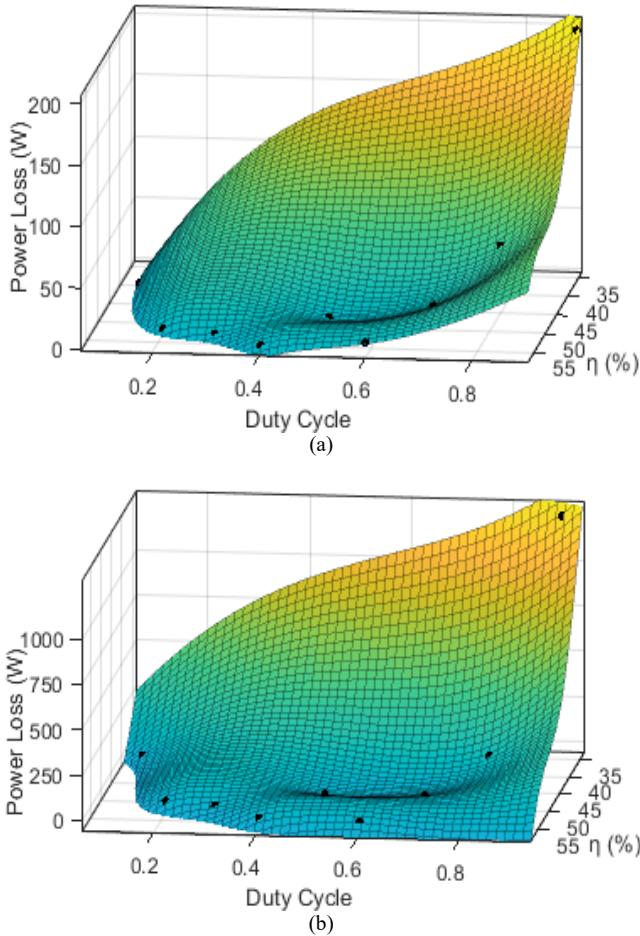


Fig.6. Non-ideal MOSFET power loss results of proposed converter (a) S1 (b) S2

V. THERMAL ANALYSIS OF MOSFET

The MOSFET thermal analysis has been performed based on the conduction method. The MOSFET junction structure and connection between the thermal conductor and heat sink has been expressed in equation 26.

$$T_j = P \times (R_{\theta ca} + R_{\theta cs} + R_{\theta jc}) + T_a \quad (29)$$

where

T_j = MOSFET junction temperature

P = MOSFET power loss

$R_{\theta ca}$ = Case to Air thermal conduction resistance
 $R_{\theta cs}$ = Case to Heat sink thermal conduction resistance
 $R_{\theta jc}$ = Junction to Case thermal conduction resistance
 T_a = Ambient temperature

Three different modes have been performed to determine the MOSFET junction temperature. Power switch without heatsink and natural cooling (T1). Power switch with heatsink and natural cooling (T2). Power switch with heatsink and forced cooling (T3). For heatsink EV-T220-51E has been assigned to perform the analysis. Besides for forced cooling 500 (ft/min) Air flow has been chosen at ambient temperature and assumed to be constant.

After manipulating equations 22 to 26 the non-ideal MOSFET thermal analysis has been performed and depicted in Fig.7(a) for switch 1 and (b) for switch 2. These graphs implicate that the with the increase in duty cycle the temperature also increase in junction of the MOSFETs. At natural cooling with no heatsink attached with MOSFET the junction temperature is 459.8°C and 397.348°C for S1 and S2 consistently. Which is exaggerated for MOSFET. With heatsink and natural cooling the MOSFET junction temperature drops to 121.25°C and 109.78°C at 0.6 duty cycle for S1 and S2 consistently. After adding fan with 500 (ft/min) air flow the temperature drops to 85.00°C and 76.38°C at 0.6 duty cycle for S1 and S2 consistently.

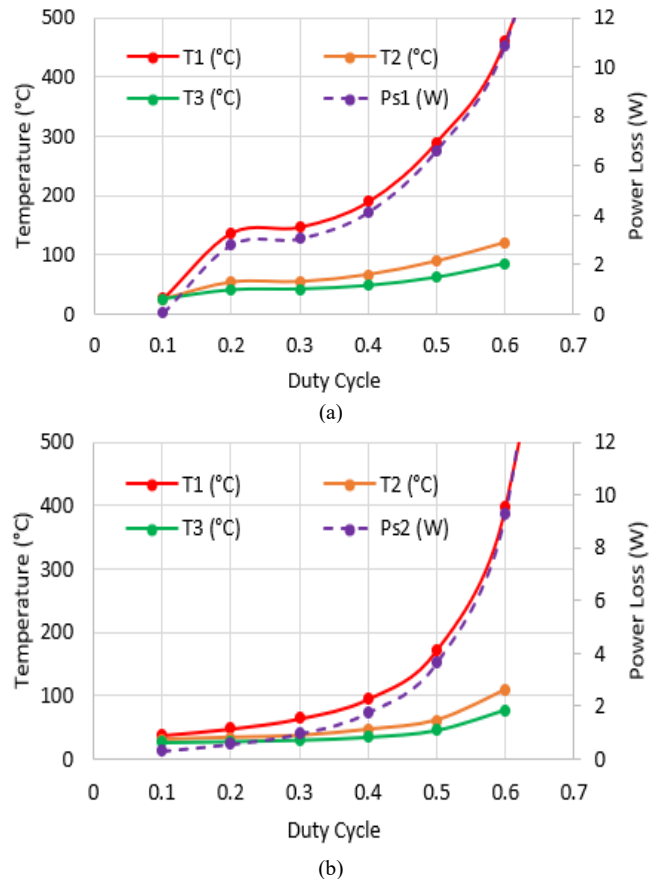


Fig.7. Non-ideal MOSFET Thermal analysis with power loss of proposed converter (a) S1 (b) S2

VI. CONCLUSION

This paper proposes a modified cascaded dc-dc boost converter for electric vehicles motor controller and other renewable applications. The power loss for MOSFET with thermal analysis are also presented. The proposed converter can achieve 4.5 to 9 times gain at 0.4 to 0.6 duty cycle. Due to high gain at intermediate duty cycle the conduction loss is minimum and with heatsink attached to MOSFET the junction temperature is at acceptable range and the temperature even more decayed should a cooling fan is added along with a heatsink.

REFERENCES

- [1] M. R. Haque, S. Das, M. R. Uddin, M. S. Islam Leon and M. A. Razzak, "Performance Evaluation of 1kW Asynchronous and Synchronous Buck Converter-based Solar-powered Battery Charging System for Electric Vehicles," 2020 IEEE Region 10 Symposium (TENSYP), 2020, pp. 770-773, doi: 10.1109/TENSYP50017.2020.9230833.
- [2] M. R. Haque and S. Khan, "The Modified Proportional Integral Controller for the BLDC Motor and Electric Vehicle," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1-5, doi: 10.1109/IEMTRONICS52119.2021.9422548.
- [3] N. ZHANG, G. Zhang, K. W. See and B. Zhang, "A Single-Switch Quadratic Buck-Boost Converter with Continuous Input Port Current and Continuous Output Port Current," in IEEE Transactions on Power Electronics, vol. PP, no. 99, pp. 1-1.
- [4] M. R. Banaei and S. G. Sani, "Analysis and Implementation of a New SEPIC-Based Single-Switch Buck-Boost DC-DC Converter With Continuous Input Current," in IEEE Transactions on Power Electronics, vol. 33, no. 12, pp. 10317-10325, Dec. 2018, doi: 10.1109/TPEL.2018.2799876.
- [5] V. F. Pires, A. Cordeiro, D. Foito, and J. F. Silva, "High Step-Up DC-DC Converter for Fuel Cell Vehicles Based on Merged Quadratic Boost-Cuk," IEEE Transactions on Vehicular Technology, vol. 68, no. 8, pp. 7521-7530, 2019.
- [6] S. Sadaf, S. B. Mahajan, M. Meraj, A. Iqbal and N. Alemadi, "A Novel Modified Switched Inductor Boost Converter with Reduced Switch Voltage Stress," in IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2020.2970648.
- [7] S. Arfin, A. Al Mamun, T. Chowdhury and G. Sarowar, "Zeta based Hybrid DC-DC Converter using Switched Inductor and Switched Capacitor Combined Structure for High Gain Applications," 2019 IEEE International Conference on Power, Electrical, and Electronics and Industrial Applications (PEEIACON), 2019, pp. 1-4, doi: 10.1109/PEEIACON48840.2019.9071940.
- [8] H. Mashinchi Maheri, E. Babaei, M. Sabahi and S. H. Hosseini, "High Step-Up DC-DC Converter With Minimum Output Voltage Ripple," in IEEE Transactions on Industrial Electronics, vol. 64, no. 5, pp. 3568-3575, May 2017, doi: 10.1109/TIE.2017.2652395.
- [9] S. Khan et al., "A Positive Output Step Up Boost Converter for Renewable Energy Applications," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021, pp. 1-6, doi: 10.1109/GUCON50781.2021.9573603.
- [10] M. R. Haque, S. Z. Eka, S. Ferdous and M. A. Razzak, "Analysis of Loss Profile and Thermal Distribution of Heat Sink of IGBT-Based Asynchronous and Synchronous Buck Converters for EV Charging System," 2021 5th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 2021, pp. 1-6, doi: 10.1109/IEMENTech53263.2021.9614827.
- [11] D. Karimi, H. Behi, J. Jaguemont, M. El Baghdadi, J. Van Mierlo and O. Hegazy, "Thermal Concept Design of MOSFET Power Modules in Inverter Subsystems for Electric Vehicles," 2019 9th International Conference on Power and Energy Systems (ICPES), 2019, pp. 1-6, doi: 10.1109/ICPES47639.2019.9105437.
- [12] C. Qian et al., "Thermal Management on IGBT Power Electronic Devices and Modules," in IEEE Access, vol. 6, pp. 12868-12884, 2018, doi: 10.1109/ACCESS.2018.2793300.

Performance Evaluation of Secured Blockchain-Based Patient Health Records Sharing Framework

Meryem Abouali, Kartikeya Sharma, Oluwaseyi Ajayi, Tarek Saadawi

Department of Electrical Engineering

City University of New York, City College, New York, USA, 10031

maboual000@citymail.cuny.edu, Kartikeyasharma04@gmail.com, Oajayi000@citymail.cuny.edu, Saadawi@ccny.cuny.edu

Abstract—With the healthcare system’s ongoing digital transformation and the need for patient data sharing to become an essential step to understanding the patient’s health history, cyber security must stay at the forefront and be made a top priority. As a result, most existing data-sharing systems depend on trusted third parties. As a result, these systems lack interoperability, data fragmentation, integrity, security, and privacy. In our previous work, we designed a framework based on Blockchain to secure patient health records exchange (SPHRS) that is fully controlled by the patient in terms of revoking or granting access and creating access policies for care providers. The framework achieves security by using smart contracts for user identity authentication and verification. The distributed IPFS storage is applied to store the encrypted patient health records and ensure immutability. In addition, NuCypher software takes advantage of a proxy re-encryption protocol to store the encryption and decryption keys securely. In this study, we assess the framework’s performance by testing metrics such as blockchain transactions’ gas consumption, throughput, Average response time, and average. Bytes. Furthermore, the security of the framework is discussed. SPHRS demonstrates how we can establish a novel approach to efficiently secure patient health record sharing. However, it shows a promising result that can potentially transform the digital patient healthcare system.

Index Terms—Ethereum blockchain, throughput, Avg.Bytes, IPFS, security, transaction, proxy re-encryption, privacy, health records.

I. INTRODUCTION

The healthcare industry generates large volumes of patient health data. This information must be accessible by the care provider promptly to provide immediate care in an emergency and must be kept private and stored securely. Moreover, the lack of trust and interoperability makes it difficult to exchange records between different health systems and among providers; as a result, doctors frequently have incomplete patient records. Meanwhile, the massive amount of PHR collected daily forces the healthcare system to adopt a centralized storage mechanism that lacks transparency or relies on a third party to store and manage these health records, which can be costly. Moreover, most healthcare systems use centralized storage to hold their patient’s data or traditional paper records without backup in case of a threat that can cause data loss or damage. Even if the healthcare

systems moved to cloud storage, they would face security threats [1].

Lately, the healthcare system has been struck by different cyber-attacks such as ransomware and distributed denial of service attacks with these traditional database practices. In addition, while the health sector is under unexpected pressure from the Covid-19 pandemic, patient health records are under cyber-attacks. Cybersecurity breaches hit a record height in 2021, exposing the patient’s health records. In 2021, 45 million individuals were impacted by healthcare attacks which is about a 26.5 percent increase from the previous year and three times higher than three years ago, according to a report that inspects the breach data reported to the US department of health and services (HHS) by healthcare organizations [2]. This suggests that more and more records are breached each year. Moreover, the healthcare industry is a prime target for attackers to monetize PHR and sell them to whoever can pay more. As a result, cybercriminals are expanding their activities to exploit security vulnerabilities across the healthcare supply chain.

As we continue into 2022, healthcare organizations must guard their cybersecurity stance and third-party vendors with access to data and networks. To avoid the consequences of increasingly sophisticated attacks, our focus should be on ensuring a solid framework to enhance PHR sharing, improve interoperability, strengthen health data sharing security, and ensure that patients are put first [3]. According to HIPAA Journal [4], as evidenced in the below Figures, the healthcare industry’s data breaches have increased nearly every year since the Department of Health and Human Services Office for Civil Rights first started publishing summaries of healthcare data Breaches. Figure 1.

According to figure 1, between 2009 and 2021, about 4.5 thousand healthcare data breaches of 500 or more records have been reported to the HHS’ Office for Civil Rights. Those breaches lead to the loss, damage, and exposure of millions of healthcare records.

As illustrated in figure 2, the exposure of the personal identity information of patients in healthcare sectors is on the rise. From the figure 2, there is an elevation in the number of records that are exposed each year, with a huge increase in 2015 due to a massive healthcare

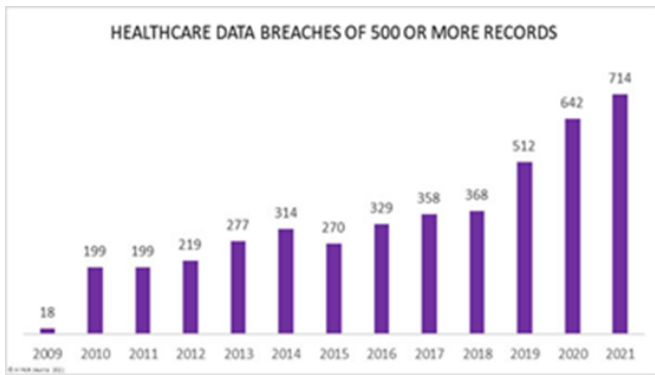


Fig. 1: Healthcare Data Breaches of 500 or more records

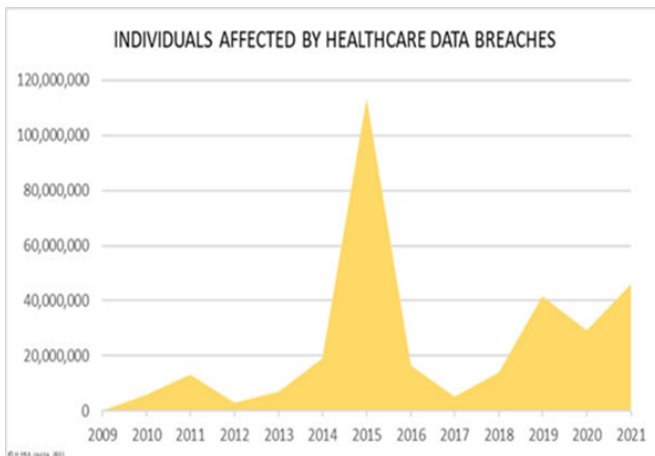


Fig. 2: Individuals Affected by the healthcare data Breaches

record breach where more than one hundred thousand records were exposed, stolen, or unauthorized disclosed and affected these health plans: Anthem Inc, Premera Blue Cross, and Excellus [5]. Healthcare system security is missing something critical if data breaches are on the rise almost every year, costing healthcare providers millions of dollars and impacting millions of patients. The use of blockchain technology can enhance HIPAA compliance and protect patient records. More importantly, as it relates to HIPAA, the security and confidentiality features of Blockchain improve cybersecurity and compliance with privacy laws. In other words, blockchain cybersecurity integrated into distributed storage solutions provide an innovative approach to managing HIPAA-compliant health records systems[6].

The remainder of our paper is arranged as follows: In section 2, background. In Section 3, motivation. In section 4, related work. Section 5 proposed the model and its implementation. In section 6, discussion of the proposed framework performance aspects. Finally, a conclusion is in section 7.

II. BACKGROUND

This section briefly describes some technologies related to the proposed framework.

A. Blockchain

Blockchain provides an immutable ledger of transactions for the data exchange system [7]. An unauthorized user cannot change or modify the recorded transactions on the Blockchain. This ensures system reliability and integrity. Furthermore, adding access control using Blockchain can provide system transparency and solve the problem of data breaches. Any illegal activity of an unauthorized user is detected and can prevent malicious transactions. As well as the smart contracts [8] to achieve authentication and user verification.

B. Interplanetary File System (IPFS)

It is a peer-to-peer distributed file system. It provides a content-addressed block storage model to uniquely identify files in a global file system[9]. IPFS gives a high throughput. IPFS has no access control, and files over IPFS can be directly accessed by their respective hash value, making it easier to integrate with different blockchains as an off-chain storage solution where Blockchain acts as an access control system.

C. NuCypher

NuCypher[10] software has two main components. It is a decentralized blockchain-based key management system (KMS) encryption and access control service that can be implemented on a private or Ethereum network. In our framework, it manages the encryption and decryption keys and enables data sharing between users using proxy-re-encryption over the network. NuCypher uses decentralized network to delegate encryption/decryption rights. Hence, decentralized KMS helps conditionally grant and revoke access to sensitive data to as many recipients as the patient like by just creating policies that can be either time-based or condition-based.

D. Ethereum Smart Contracts

The smart contracts are automatically executed when the rules in the contract are met. A smart contract automatically enforces predetermined rules between parties without the need for intermediaries, with a credible public ledger. It can store metadata and program some events according to some conditions. They are transaction protocols that intend to enforce the terms of the contract and can be used by the developers in creative ways in programming functionality [11].

E. Proxy re-encryption (PRE)

Proxy re-encryption is a public key cryptographic primitive where a data owner can grant the right to decrypt his data to other data requesters [12]. After a semi-trusted proxy re-encrypts the ciphertext, which is under an owner's public key, a data requestor can decrypt the new ciphertext through his secret key. Identity-based proxy re-encryption (IBPRE) is a PRE scheme in which data owners and requestors take their identities as public keys and no longer need public infrastructure (PKI).

III. MOTIVATION

Patient health records (PHR) are the foundation of any patient health record exchange system. The data must be shared among various healthcare entities so that health professionals have the information they need to make informed decisions that can impact their lives. However, accessing and sharing patient health records is facing a big challenge by compromising the patient's confidentiality and privacy without breaking confidentiality or compromising patient privacy is an ongoing challenge in today's increasingly digitized world [14].

Patient health record sharing plays a huge part in the healthcare system by enhancing interoperability. To address the issues of the existing PHR sharing and ensure data privacy and security during data sharing, the need for a secure, reliable system that enables care providers to diagnose and treat patients becomes an even greater priority.

Motivated by the need for a more reliable system that will guarantee the security and the privacy of the PHR sharing and solves the problem of the lack of interoperability, the lack of clear definition of the data ownership, and the protection of sensitive records while making the information accessible to authorized providers, we designed a framework that can ensure all the above (SPHRS).

The built system can facilitate secure, trustable management and aggregation of PHR data. In addition, the system will ensure patient privacy protection and guarantee security where patients are allowed to manage their health records efficiently among multiple providers. These features can only be achieved using blockchain technology [15], an ideal way to lay the foundation for decentralization and supporting healthcare data integration across various use cases. It also has the potential to solve existing issues in health records sharing.

It is impossible to store a large amount of data within the Blockchain. Ethereum blockchain has a cost associated with performing a transaction, known as a "gas fee," if one wants to store data, the gas price will be high and impractical. The alternative will store the data on secured off-chain decentralized IPFS storage and store the hash on the Blockchain. The main advantage is that IPFS allows patients to keep their large medical files securely, such as X-ray scans, MRI scans, and more, to solve the issue of the latency and block size limit in the Blockchain.

In our previous paper, we built a fully decentralized model in all its essence and discussed our framework's contribution in detail. In contrast, in this paper, we will focus more on testing and evaluating the performance measure of our framework. The proposed architecture makes the following contributions:

- This study presents a solution for patient health records exchange using the Ethereum blockchain for verification and authentication of the users and to prevent and deny an unauthorized provider access to the patient's sensitive information.

- The Patients own and control the access to their data which eliminates all the barriers for the patients to request copies of their health records or send them to another doctor.
- The health records are stored in a distributed tamper-proof Interplanetary File System IPFS, which is fully distributed off-chain storage and makes it possible to efficiently distribute high volumes of data. Furthermore, it provides fast PHRs retrieval capability.
- Data is encrypted and stored in IPFS and can only be decrypted with the provider's private key. Therefore, if a malicious party compromises the network, there is no way to read patient data.
- This study uses the Ethereum blockchain smart contracts Smart for identity verification and authentication of the doctor and the healthcare entities, making it difficult for an unauthorized doctor to access the data.
- This study performs the vulnerability and security analysis on the unauthorized doctor who wants to access the records and the ability of the patient to revoke the doctor's access to these PHRs at any time.

IV. RELATED WORK

Blockchain is taking the world by storm. It is implemented almost in every existing field. The blockchain potential in the healthcare sector is undeniable to increase healthcare data security, privacy, and interoperability and is seen as the new model for health data sharing. However, most existing schemes use the centralized approach for storing data. This section reviews the current systems related to blockchain-based on PHRs sharing and access management control. In [16], the authors introduced how Blockchain would facilitate the health sector. The specified data sharing is why Blockchain should be used in healthcare to deal with authorized users' access, data availability, and obtaining of health records. It also considers the data storage, either on-chain or off-chain storage. However, the challenges and barriers to using Blockchain were the huge volumes of health records, patient access control, security, and privacy.

The authors [17] designed a Medshare system based on a blockchain approach for healthcare information exchange using cloud services. The proposed method provides security of health records and access control. However, they had not considered the problem of key management and scalability. The evaluation of the performance of the proposed system was done on network latency measurements using theoretical analysis. In [18], the authors proposed a secure ABE system with multiple authorities for Blockchain in Electronic Health Records. They used an attribute-based signature scheme for numerous electronic health record management users with Blockchain. In this attribute-based, their objective is to achieve security and privacy of the system and achieve the immutability of the information ledger. However, the system lacks interoperability and privacy.

In [19], the authors proposed an implementation of MedChain based on a Hyperledger blockchain system to

manage health data. The system has three components: an access control scheme, cloud storage, and a user interface. Patient health data is encrypted and stored on the cloud while the hash of the data is stored in the Blockchain. However, the system does not consider how different healthcare entities can connect and share patient health data. In [20], the author presents a blockchain-based health record exchange system that guarantees patient privacy and integrity of his health data. The model provides a patient’s record immutability. However, the model has not considered another alternative to store the huge amount of data that the Blockchain cannot support.

In [21][22], the authors present a framework based on Blockchain to address the data scattered among different healthcare systems. The medical data is stored in a decentralized structure and uses the blockchain immutability property to provide security. In addition, the authors make use of smart contract transactions to transfer medical records. However, the system stores medical data cloud-based, which can be easily compromised, and unauthorized users can access sensitive medical data.

In [23], the authors present Medrec system, an Ethereum-based record management system for EHRs. The method leverages Blockchain to administer user authentication, the confidentiality of health records, accountability, and data exchange. MedRec provides patients with a complete history of their health records through an inalterable log easily accessed across different providers. However, MedRec only integrates with the provider’s current storage solution. Therefore, the system is not patient-centric.

All the above approaches have no restriction on privacy security and hence pose a threat to the integrity of the health records. Furthermore, relying on centralized storage to store health records which means a third-party dependency, is a major disadvantage. These proposed approaches have not considered an alternative solution to store huge data that is not supported by the Blockchain. In this study, we have resolved most of the healthcare sector’s issues by integrating technologies such as Blockchain, IPFS storage, and other emerging technologies like NuCypher to provide higher quality healthcare for the patient.

V. PROPOSED MODEL AND ITS IMPLEMENTATION

This section offers a solution that can help improve patient care, care coordination, and the nation’s overall health. SPHRS [24] uses the Ethereum blockchain that features smart contracts to safeguard security and privacy by verifying and authenticating all the involved parties who can access the PHRs. In addition, the IPFS is used to resolve the storage cost and scalability issues. Our solution removes the need for a reliable centralized authority. It gives security and privacy to the patient health records exchange. The robustness of the infrastructure with high integrity and resiliency is achieved by using NuCypher software to store safely and securely the encryption re-encryption keys. Figure 3.

The system model has the following entities:

- The data owner, who is the patient, can grant or revoke access to an unauthorized provider.
- The encryptor is the Healthcare entity and is the PHRs source that encrypts and stores records on IPFS.
- The provider retrieves records from IPFS and grants access to the patient.
- Ursulas are nodes on the NuCypher network representing the proxy re-encryption proxy. They are responsible for re-encrypting data and enforcing the access policy created by the patient.

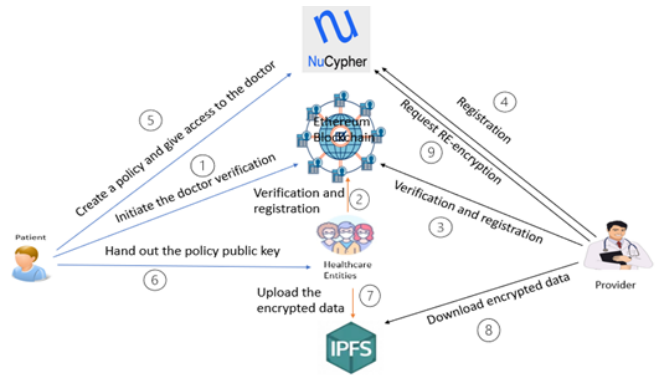


Fig. 3: : Secured Patient Health Records Sharing based on Blockchain

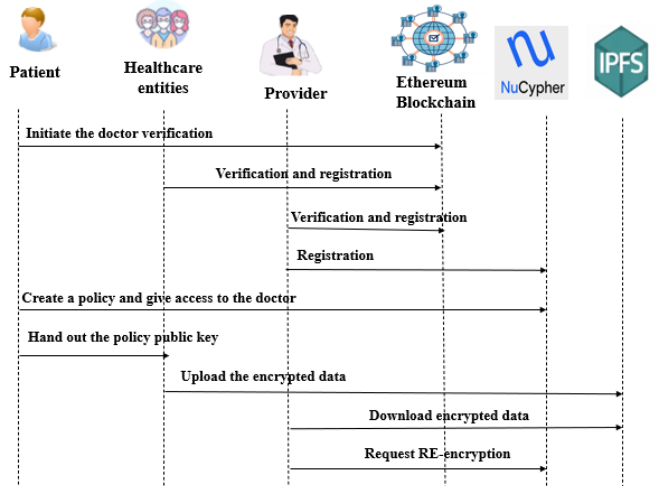


Fig. 4: Sequence Diagram for Sharing Patient Health Records

Our proposed framework consists of four parts:

A. Users Authentication and Registration

There is currently a provider and healthcare entities verification smart contract initiated by the patient and deployed on the Ethereum test network. There are two parts to this smart contract:

- Information receiving: The provider and the healthcare entities can submit their basic information, which will

be used to verify their credentials and identity : A function in the smart contract that tells us whether a provider and the healthcare entities with a specific id have been verified yet. After that, the provider must be authenticated to be able to create an account on the Blockchain.

B. Patient grants the provider access by creating a policy

For any authorized provider to permit access to the patient’s encrypted health records, the patient must create a policy and upload it to the NuCypher Network. The patient combines his private key and the provider’s public key to get the re-encryption key and send it to the proxy servers. Next, the patient collects information about the nodes or Ursulas who provide the proxy re-encryption service linked with the policy. Each Ursula gives their encrypted key, and the patient generates n re-encryption key shares (kFrag). The Ursulas remain ready to re-encrypt data. The list of Ursulas and other information are stored in a Treasure Map. Every time the patient creates a policy, this policy has an associated policy public key, which the healthcare entity can use to encrypt data on the patient’s behalf.

C. Encryption and upload of PHR to IPFS by the Healthcare Entities

The healthcare entities use the public policy key generated when the policy was first created from the patient and encrypt PHRs. Next, the Healthcare entities use a randomly generated symmetric key to encrypt the PHRs and then encrypt the symmetric key with the patient’s policy public key. Finally, the Healthcare entities upload the encrypted symmetric key and the encrypted PHR and store them at IPFS.

D. Retrieval of Patient’s Health Records by the provider

For the provider to access the patient health records, he must use the treasure map to allocate the list of Ursulas who stand ready for the re-encryption operation. Then, the provider obtains the encrypted data from the IPFS and sends a re-encryption request to the relevant Ursulas on the network. If the policy is satisfied and the provider is an authorized user, Ursula decrypts the provided re-encryption key share and re-encrypts the provider’s public key. Afterward, the provider can decrypt the capsule with the private key to obtain the symmetric key used in data encryption. Then can access the data after decrypting the records.

VI. PERFORMANCE EVALUATION

This section aims to provide the performance evaluation by conducting several tests to validate the performance of the proposed healthcare system and reduce the risks associated with this novel approach. This study assesses the transaction’s performance in terms of gas consumption, transaction throughput, average time, and sent and received byte. This section explains the evaluation of different security perspectives and user data requirements.

A. Transaction Gas fees

We evaluate the performance of the proposed framework by determining the cost of the transaction in terms of gas consumption. Every transaction on The Ethereum blockchain contains data. Therefore, we calculate the gas fee associated with different transactions of the proposed framework. In the Ethereum, ” ETH” is used to calculate the transaction fees [25], which is an Ethereum coin, and its units are Wei and gwei. The transaction fee is the product of the gas consumed and the gas price. Table 1 relates the gas consumption values to the various functions used in the smart contract. After deploying the blockchain smart contract to the Ethereum, a public Rinkeby testing network (aka testnet) was used for this, which is a network that operates as the main Ethereum network, but where Ether has no value and is free to acquire-making them ideal for testing your contracts at no cost. The transaction and the execution gas were recorded. Hence, the transaction cost is the amount of gas required to send data on Ethereum, and the execution cost is the computational fee necessary to execute the function.

TABLE I: Gas Consumption of the Ethereum Transaction

Smart Contract	Function	Transaction Fee (Ether)		
		Trial 1	Trial 2	Trial 3
Doctor Verification	setDoctor	0.000046317	0.000046329	0.000046329
	changePatient	0.000028777	0.000028777	0.000028777
Doctor Authentication	setPrimaryInfo	0.000089364	0.000040201	0.000040213
	set Doctor Info	0.000137441	0.000062734	0.000062710
Healthcare Entity Authentication	setPrimaryInfo	0.000089364	0.000040249	0.000040237
	sethealthcareEntityInfo	0.000137335	0.000062784	0.000062808

We implemented and deployed three smart contracts. Table 1 exhibit the transaction costs of the used functions. The setDoctorInfo and sethealthcareEntityInfo, their general information inside the smart contracts, make up the major share of the transaction costs. On the other hand, patient function changes have a much lower cost. As for the remaining functions, starting with setPrimaryInfo, the costs are slightly increased as they perform expensive checks using the primary information of the doctor and the healthcare entities.

B. Performance Testing

SPHRS scalability is very important for the application. To spin up a network for scalability testing, we create a network of a varied number of users(Threads). SPHRS was tested with varying numbers of users in terms of the NuCypher endpoint for the policy encryption key for the label. The results show that NuCypher, even a decentralized

PREservice, performs well for endpoint policy encryption keys. To evaluate the performance, we are using Apache JMeter, a desktop performance testing tool to analyze and test applications [26]. Using JMeter, we mimic the number of users from 10 to 5000 users using the system and performing its several features. In JMeter, the throughput signifies the number of requests made in a given period, and its unit is KB/Sec. The average represents the average response time of all the total samples. The values for received and sent represent how many kilobytes of data were received and sent, respectively. While the Avg. Byte means the average data size passed through each API call. We repeatedly called the same API with the same data size for each experiment.

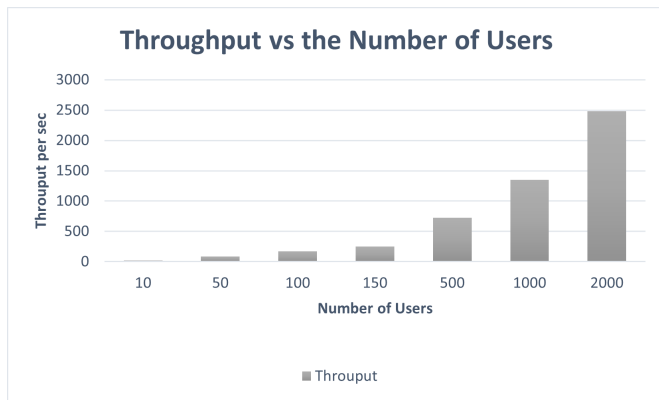


Fig. 5: Throughput vs the Number of Users

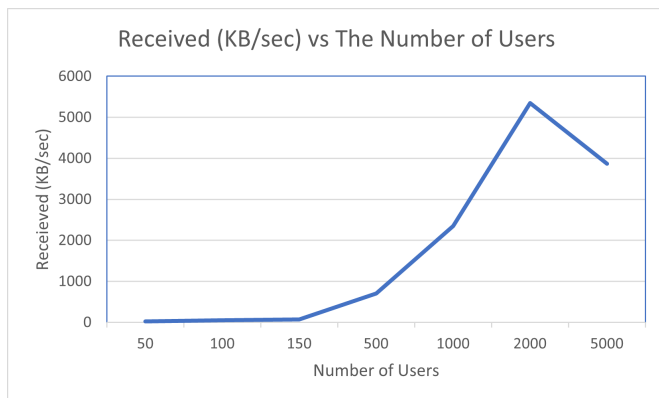


Fig. 6: Received KB per second vs. the Number of Users

As findings for the above figures 5 to 9, we can conclude that as the number of users increases, the throughput, received and sent data, and the Average response time all increase. While, in threads 10, 50, 100, and 150, the Avg. Bytes are 302. The average started to increase later because some API calls started to return errors, and therefore, the response included more bytes. The errors partly came because we were experimenting on my local machine, and it is very machine intensive to have 500+ threads open at once. With the above experimentation and comparison, we conclude that SPHRS is viable. A decentralized alternative

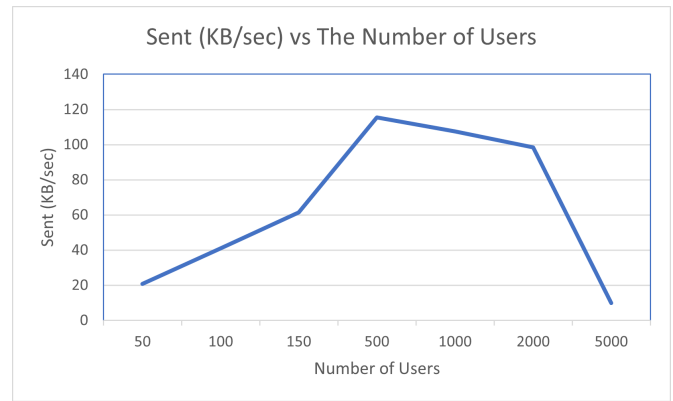


Fig. 7: Send KB per second vs. The Number of Users

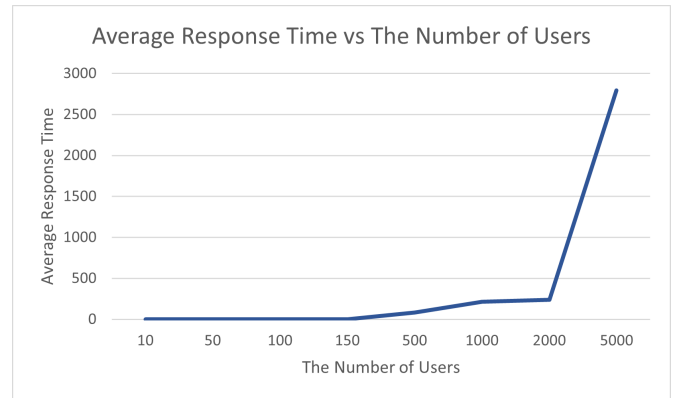


Fig. 8: Average Response Time of all the Users

to the currently present centralized option. Some outlined benefits of SPHRS will be:

- SPHRS can perform better in response to throughput
- Decentralization also makes sure no single point of failure.
- SPHRS provides data governance to the patient, thus increasing transparency, decreasing delay, and improving services.

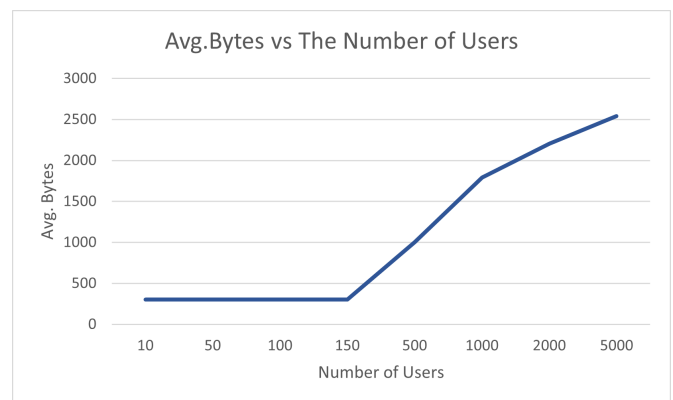


Fig. 9: Avg. Bytes vs. The Number of Users

VII. CONCLUSION

This paper tests the performance metrics of our proposed secured PHRs framework. Our Blockchain framework based on secure PHR sharing includes all the expected health records sharing system criteria. We identify the security vulnerability in the existing approaches, and we tackled these issues in our proposed framework by proposing efficient solutions through an actual prototype implementation. Furthermore, our framework surpasses the current systems because it enhances the privacy and the security of the patient health records exchange. Finally, we evaluated the performance metrics of our proposed framework. For future work, we will investigate how the framework is protected against inside and outside threats by testing the vulnerability of the framework against various types of cyber-attacks [27]. Moreover, we will extend this implementation work in terms of scalability by increasing the number of patients, providers, and the size of the health records. Finally, we will enhance the performance of the existing framework and test the performance metrics in terms of transaction dissemination latency, security, and scalability. The measurement will be evaluated using the following platforms: the GRE tunnel setup between Kyutech in Japan and CCNY in the US and Extending the CCNY- Japan tunnel experiment to the COSMOS experiment.

ACKNOWLEDGMENT

This work is supported by the following: NSF Grant, INRC Testbed: COSMOS Interconnecting Continents, and NSF Grant; Japan USA Networking Opportunities JUNO3.

REFERENCES

[1] S. T. Argaw, N. E. Bempong, B. Eshaya-Chauvin, and A. Flahault, "The state of research on cyberattacks against hospitals and available best practice recommendations: A scoping review," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, p. 10, Dec. 2019.

[2] Jessica Davis "The 10 Biggest Healthcare Data Breaches of 2021 impact over 22.6 M patients" byHealth IT Security. <https://www.scmagazine.com/feature/ransomware/10-biggest-healthcare-data-breaches-of-2021-impact-over-22-6m-patients>

[3] W. W. Koczkodaj, M. Mazurek, D. Strzałka, A. Wolny-Dominiak, and M. Woodbury-Smith, "Electronic health record breaches as social indicators," *Social Indicators Res.*, vol. 141, no. 2, pp. 861–871, Jan. 2019.

[4] Healthcare Data Breach Statistics for 2021.URL <https://www.hipaajournal.com/healthcare-data-breach-statistics/>

[5] Jessica Davis "The 10 Biggest Healthcare Data Breaches of 2021 impact over 22.6 M patients" health IT Security. <https://www.scmagazine.com/feature/ransomware/10-biggest-healthcare-data-breaches-of-2021-impact-over-22-6m-patients>.

[6] Alamri B, Javed IT, Margarita T. A GDPR-compliant framework for IoT-based personal health records using Blockchain. In2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS) 2021 Apr 19 (pp. 1-5). IEEE.

[7] M. Hölbl, M. Kompara, A. Kamišalić, and L. N. Zlatolas, "A systematic review of the use of blockchain in healthcare," *Symmetry*, vol. 10, no. 10, p. 470, 2018.

[9] Interplanetary File System (IPFS). Accessed: Feb. 4, 2019. [Online]. Available: <https://ipfs.io/>

[8] S. Wang, Y. Zhang, and Y. Zhang, "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems," *IEEE Access*, vol. 6, pp. 38437–38450, Jun. 2018

[10] Egorov, Michael, Wilkison, MacLane, Nunez, and David, "Nucypher kms: Decentralized key management system," URL <https://arxiv.org/abs/1707.06140>

[11] Wood, G., et al. (2014) Ethereum: A Secure Decentralized Generalised Transaction Ledger. Ethereum Project Yellow Paper, 151, 1-32

[12] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," *ACM Trans. Inf. Syst. Secure.*, vol. 9, no. 1, pp. 1-30, 2006

[13] M. Green and G. Ateniese, "Identity-based proxy re-encryption," *Conference Proceedings*, pp. 288-306.

[14] D. Spatar, O. Kok, N. Basoglu, and T. Daim, "Adoption factors of electronic health record systems," *Technol. Soc.*, vol. 58, Aug. 2019, Art. no. 101144.

[15] A. A. Siyal, A. Z. Junejo, M. Zawish, K. Ahmed, A. Khalil, and G. Soursou, "Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives," *Cryptography*, vol. 3, no. 1, p. 3, Jan. 2019

[16] Kumar R, Marchang N,Tripathi R Distributed off-chain storage of patient diagnostic reports in healthcare system using IPFS and Blockchain; International Conference on COMMunication Systems NETworkS (COMSNETS); 2020; Bengaluru, India; 1-5, <https://doi.org/10.1109/COMSNETS48256.2020.9027313>.

[17] Rouhani, Sara, Butterworth, Luke, Simmons, Adam, Humphery, Darryl G Ralph Deters: MediChainTM: A Secure Decentralized Medical Data Asset Management System. CoRR abs/1901.10645 (2019).

[18] Oliveira,M.T,Reis,L.H, Carrano,R. C, Seixas,F. L, Saade,D. C., Albuquerque, C. V. Maos, D. M.(2019, May). Towards a Blockchain-Based Secure Electronic Medical Record for Healthcare Applications.In ICC 2019-2019 IEEE International Conference on .pp.1-6. IEEE

[19] Gupta R, Shukla A, Tanwar S. AaYusH: a smart contract-based telesurgery system for Healthcare 4.0. IEEE International Conference on Communications Workshops (ICC Workshops); 2020; IEEE; 1-6 15.

[20] T. Kumar, V. Ramani, I. Ahmad, A. Braeken, E. Harjula, and M. Ylianttila, "Blockchain utilization in healthcare: Key requirements and challenges," 2018 IEEE 20th International Conference on e-Health Networking Applications and Services (Healthcom), pp. 1-7, Sep. 2018.

[21] Thomas K. Dasaklis, Fran Casino," Blockchain Meets Smart Health: Towards Next Generation Healthcare Services," 978-1-5386-8161-9/18©2018 European Union

[22] Abugabah A,Nizam N,AlzubiAA. Decentralized telemedicine framework for a smart healthcare ecosystem. IEEE Access. 2020;8:166575-166588.

[23] Kumar R, Tripathi R. A secure and distributed framework for sharing COVID-19 patient reports using consortium Blockchain and IPFS. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC); 2020; IEEE.

[24] Meryem Abouali, Kartikeya Sharma, Oluwaseyi Ajayi, Tarek Saadawi, "Blockchain Framework for Secured On-Demand Patient Health Records Sharing" 2021 IEEE 12th Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON 2021) December 1-4, 2021, New York, USA.

[25] Pierro GA, Rocha H, Tonelli R, Ducasse S. Are the gas prices oracle reliable? A case study using the eth gas station. In2020 IEEE International Workshop on Blockchain Oriented Software Engineering (IWBOSE) 2020 Feb 18 (pp. 1-8). IEEE.

[26] M. Niranjanamurthy, K. Kumar S, A. Saha, and D. D. Chahar, "Comparative study on performance testing with JMeter," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 2, pp. 70–76, 2016.

[27] O.Ajayi,M.Abouali and T.Saadawi,"Secured Inter-Healthcare Patient Health Records Exchange Architecture,"2020 IEEE International Conference on Blockchain (Blockchain),2020, pp.456-461.

Input Fuzzing for Network-based Attack Vector on Smartphones

Micah Noyes
Computer Science

University of North Carolina Wilmington
Wilmington, North Carolina
mrn8503@uncw.edu

Hosam Alamleh
Computer Science

University of North Carolina Wilmington
Wilmington, North Carolina
hosam.amleh@gmail.com

Abstract—In the last decade, smartphones have evolved tremendously. They became powerful computers that fit in our pockets. Today, with the huge adoption of smartphones. Smartphone security has become essential. Nowadays, a smartphone has become a complex piece of equipment that a single device incorporates users, Operating systems, manufacturers, carriers, and app developers. With such complexity, smartphones can be subject to many attack vectors. This paper explores one of them which is network-based attacks on cell phones. This paper discusses the several types of attacks then, it proposes an input fuzzing system for network-based attacks on phone. The proposed system is novel, as today for the previously mentioned attack vector, fuzzing is done using an emulator. This paper introduce a new fuzzing system for smartphones that addresses network-based attacks. This is done by setting up a network environment using SDRs. The proposed setup allows fuzz testing for Smartphones and baseband OSeS for network-based attacks coming through SMS, WAP, and calls.

Index Terms—Mobile Network, Attacks, Network-based attacks

I. INTRODUCTION

A smartphone is essentially a small computer that can fit in one's hand. The beauty of using a smartphone lies in its ability to perform different functions and allow its users to access networks such as the Internet while providing convenience. In the past two decades, there has been rapid growth in the number of mobile devices utilized, especially with the global smartphone boom. Smartphones took over the wider consumer market, and as of 2021 according to the Pew Research Center, 85% of US adults use smartphones [1]. Smartphone revolution is global. The number of smartphones has reached 6.2 Billion globally in 2021. Smartphones are used for many applications in our daily life. This includes banking, games, emails, and e-commerce. With the wide range of applications and use cases, smartphones holds very sensitive data, this includes pictures, call logs messages emails. Also, smartphone has several ways to sense its surrounding such as cameras and microphones. Therefore, smartphones became a valuable target for attackers. This created new security challenges. Attacks on smartphones have been on the rise in the last few years. According to the 2021 Mobile Security Report, 40% of all mobile devices

are vulnerable to attacks [2]. Moreover, The same report also stated that nearly every company witnessed at least one smartphone malware attack last year. 93% of the said attacks stemmed from the mobile network.

Many of attacks on phones comes to the mobile network. A mobile network allows users to access the network while on the move. It consists of base stations where each cover a limited area or "cell." When joined together these cells provide radio coverage over a wide geographic area. This enables a large number mobile devices to communicate with each other and to the internet. To connect to a mobile network, a smartphone uses a wireless protocol (e.g. GSM, 3G, LTE). Smartphones include a "side" OS known as the baseband OS. This OS implements the wireless protocol. The baseband runs on a dedicated processor that is different from the one where the smartphone OS runs. which is known as the baseband processor. A baseband processor is a chip in a smartphone, that helps convert digital data into radio frequency signals which can then be transmitted over a Radio Access Network. This processor is different from the processor than runs the smartphone OS and applications. The baseband chip incorporates encryption keys used to authenticate to the mobile network and are stored on a tamper-resistant smart card known as the SIM card along with other user identification information that allows the phone to connect to the network.

Network-based attacks on smartphones attacks either coming from network or an entity intercepting the communication (man-in-the-middle attack). Such attacks include denial of service, malware injection, and others. With many attack vectors, researchers attempt to find vulnerabilities before they are discovered and exploited by hackers. One technique to discover vulnerabilities is input fuzzing. Input fuzzing is a process in which malformed inputs are sent to the targets with the objective to trigger bad behaviors, such as crashes, infinite loops, and/or memory leaks. Input fuzzing has been a tool used by programmers and penetration testers to detect problems in the systems. In this paper, we implement a system of input fuzzing targeting Smartphones. In this system it is possible to input fuzz the network-based attack vector on

smartphones. This would include, Short messages service (SMS), wireless application protocol (WAP) Pushes, pagings, and phone calls. We show details about this implementation using Two different Software defined Radio (SDR) chipsets.

II. BACKGROUND

Mobile networks are maintained by the carrier that provides SIM cards. The smartphone maintains a baseband chip that uses authentication information on the Simcard to access the network. There are many ways to attack this setup. For one, attackers can target the Simcard, in a demonstration, it was shown that a text message from a spoofed carrier can obtain the 56-bit DES encryption key of a SIM [3]. Alternatively, an attacker could target the baseband OS [4] [5]. One attack used a carrier Over-The-Air (OTA) update mechanism to bypass the lock screen [6], or flash a Samsung Galaxy S6 baseband remotely [7]. On the other hand, Man-in-the-middle attacks targets communications between the smartphone and the cellphone towers. GSM has a short encryption key which makes it very easy to attack [8]. Another issue with GSM is its lack of authentication of the base station, Thus an attacker with a spoofed base station can capture the SIM and the phone information and intercept outgoing calls and texts [9].

Alternatively, attackers target the OS on smartphones. This would be trying to exploit weakness or poor design on Oses of smartphones. There are has been several exploits reported on attacks targeting system using SMS [10] [11]. For example, one vulnerability allows a setup message to create buffer overflow [12] Other attacks are done through WAP [13] [14]. For example on Samsung Galaxy S4 through S7 a a malformed WAP PUSH SMS causes the Android runtime to continually crash, rendering the device unusable until a factory reset is performed [15]. Also, attacks using calls [16] [17], for example, An issue was discovered on LG mobile devices with Android OS 11 software. Attackers can bypass the lockscreen protection mechanism after an incoming call has been terminated [18]. Exploits also targets iMessage the iPhone messaging App. A zero-click exploit against iMessage targets Apple's image rendering library, that would allow arbitrary code execution.

One technique to discover such attacks is input fuzzing. Fuzzing does not have to detect an exploit. But it can detect weakness that can be investigate further, which could lead to detect exploits. Fuzzing also discover how a phone would react to different input. For example, Apple operating system iOS crashed when received a special character [19]. There has been several implementation to input fuzz smartphones. Inputting fuzzing targets different aspect of the smartphone. Including the Operating system [20], Applications [21], and drivers [22]. Also, there are systems to input fuzz the baseboard OS using an emulator [23]. This paper introduce a new fuzzing system for smartphones that addresses network-based attacks. This is done by setting up a network environment using SDRs. The proposed setup allows fuzz testing for

Smartphones and baseband Oses for network-based attacks coming through SMS, WAP, and calls.

III. SYSTEM MODEL

The proposed Fuzzer for Network-based attack vector on smartphone is as shown in Fig.1 and consists of the following elements:

- 1) **Fuzzer dictionary:** This dictionary consist of the list of the inputs to be tried by the fuzzer. It is very useful to provide fuzz target with a set of characters, values, and different text format that are expected to find in the input. Adding a dictionary highly improves the efficiency of finding new weakness. In general, a fuzzer is as good as the dictionary. A well-rounded dictionary that has a list of letters, characters, and inputs in different unicodes would be a great testing tools to find weakness. Alternatively, a dictionary could include a set of network configuration messages to study their impact on the targeted phone.

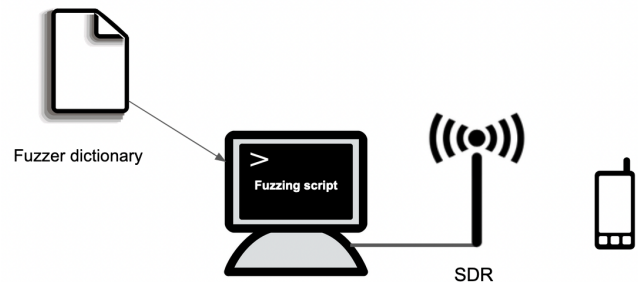


Fig. 1. System overview

- 2) **Fuzzing script:** The fuzzing scripts reads the input from the fuzzer dictionary and sends that to be transmitted by the SDR. The script usually would include a loop that goes through a elements of the fuzzer dictionary. It is important that the loop has enough sleep time to incorporate the the transmission ques and the buffer size of the SDR.
- 3) **SDR:** An SDR uses software for the modulation and demodulation of radio signals and can be used to imitate the performance of a radio communication protocol (e.g., GSM, 3G, UMTS, 5G). In general, different protocols entails different types of communication format, which leads to different nature for the attack vector. Therefore, it is important to input fuzz the smartphone with the appreciate protocols and communication format with the appropriate fuzzing dictionary.
- 4) **The target:** The target is usually a Smartphone. Usually, in the proposed fuzzing system, both Baseband

OS, and Smartphone OS can be targeted with the input fuzzing.

IV. IMPLEMENTATION

The implementation of the proposed system follows the element discussed in the in the previous section and are implemented as follows:

- 1) **Fuzzer dictionary:** a text file. Each input is written on a newline. The fuzzing system is running on DragonOS, which is a linux disturbiton that is tailored for wireless applications.
- 2) **Fuzzing script:** The fuzzing scripts reads the input from the fuzzer dictionary and sends that to be transmitted by the SDR. The script is written using python.
- 3) **SDR:** Two different models of SDR ware used. The USRP b205mini [24], which was used with an open-source communication package referred to as osmocomb [25]. Osmocom includes software and tools that implements a variety of mobile communication standards, such as GSM, DECT, TETRA and others. The other type of SDR used was BladeRF 2.0 micro xA5 [26] with software package known as YateBTS [27].
- 4) **The target:** For the purpose of the implementation a Samsung galaxy S6 was used.

The implementation is shown in Fig 2.

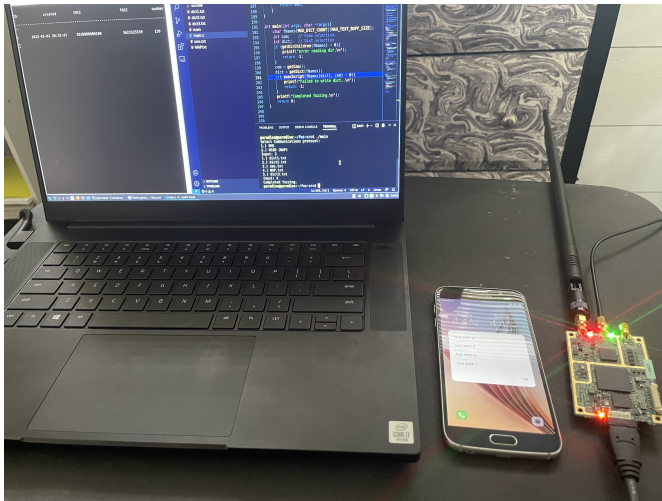


Fig. 2. Proposed fuzzing system implementation

V. CONCLUSION

In the last decade, smartphones have evolved tremendously and today became essential in our daily lives. With the Smartphones carrying many valuable information about their users, they have been a target for attackers. Smartphones can be targeting through several attack vector. This paper discussed network-based attack vector on smartphones. Such attack vector requires smartphone developers to test them

and fortify them against such attacks. One technique to test smartphones against such attacks is input fuzzing. This paper introduced a new fuzzing system for smartphones that addresses network-based attacks. This is done by setting up a network environment using SDRs. The proposed setup allows fuzz testing for Smartphones and baseband OSe for network-based attacks coming through SMS, WAP, and calls. The proposed system was implemented with two different models of SDRs. The proposed system is a powerful tool to automatically test smartphones against a vast number of inputs sent by the network.

REFERENCES

- [1] P. Research, "Mobile fact sheet," 2021. [Online]. Available: www.pewresearch.org/internet/fact-sheet/mobile/
- [2] C. S. technologies, "Mobile security report 2021," 2021.
- [3] K. Nohl, "Rooting sim cards," 2013.
- [4] "Baseband attacks: Remote exploitation of memory corruptions in cellular protocol stacks," in *6th USENIX Workshop on Offensive Technologies (WOOT 12)*. Bellevue, WA: USENIX Association, Aug. 2012. [Online]. Available: <https://www.usenix.org/conference/woot12/workshop-program/presentation/Weinmann>
- [5] C. Mulliner, N. Golde, and J.-P. Seifert, "SMS of death: From analyzing to attacking mobile phones on a large scale," in *20th USENIX Security Symposium (USENIX Security 11)*. San Francisco, CA: USENIX Association, Aug. 2011. [Online]. Available: <https://www.usenix.org/conference/usenix-security-11/sms-death-analyzing-attacking-mobile-phones-large-scale>
- [6] M. Solnik and M. Blanchou, "cellular exploitation on a global scale: The rise and fall of the control protocol," 2014.
- [7] X. C. Marco Grassi, "Over the air baseband exploit: Gaining remote code execution on 5g smartphones," 2021.
- [8] P. K. Gundaram, S. Allu, N. Yerukala, and T. Appala Naidu, *Rainbow Tables for Cryptanalysis of A5/1 Stream Cipher*, 02 2021, pp. 251–261.
- [9] F. Broek, R. Verdult, and J. Ruiter, "Defeating imsi catchers," 10 2015, pp. 340–351.
- [10] NIST, "Cve-2022-25821," 2022. [Online]. Available: nvd.nist.gov/vuln/detail/CVE-2022-25821
- [11] Mitre, "Cve-2021-25426," 2021. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-25426
- [12] —, "Cve-2020-25279," 2020. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-25279
- [13] —, "Cve-2018-10751," 2018. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-10751
- [14] —, "Cve-2016-7991," 2016. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2016-7991
- [15] —, "Cve-2016-7989," 2016. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2016-7989
- [16] —, "Cve-2021-41181," 2021. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-41181
- [17] —, "Cve-2021-39707," 2021. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-39707
- [18] —, "Cve-2021-30161," 2021. [Online]. Available: cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-30161
- [19] N. security, "Apple fixes that "1 character to crash your mac and iphone" bug," 2018. [Online]. Available: nakedsecurity.sophos.com/2018/02/20/apple-fixes-that-1-character-to-crash-your-mac-and-iphone-bug
- [20] D. Cotroneo, A. K. Iannillo, and R. Natella, "Evolutionary fuzzing of android os vendor system services," *Empirical Software Engineering*, vol. 24, 12 2019.
- [21] H. Ye, S. Cheng, L. Zhang, and F. Jiang, "Droidfuzzer: Fuzzing the android apps with intent-filter tag," 12 2013.
- [22] J. Corina, A. Machiry, C. Salls, Y. Shoshitaishvili, S. Hao, C. Kruegel, and G. Vigna, "Difuze: Interface aware fuzzing for kernel drivers," 10 2017, pp. 2123–2138.

- [23] D. Maier, L. Seidel, and S. Park, "Basesafe: Baseband sanitized fuzzing through emulation," 05 2020.
- [24] Ettus, "Usrcp-b205mini-i," 2022. [Online]. Available: www.ettus.com/all-products/usrp-b205mini-i/
- [25] Osmocom, "Osmocom network in the box," 2022. [Online]. Available: www.osmocom.org
- [26] nuand, "Bladerf-xa5," 2022. [Online]. Available: www.nuand.com/product/bladeRF-xA5/
- [27] YateBTS, "Yatebts - lte gsm mobile network components for mno mvno." 2022. [Online]. Available: yatebts.com

Developments Pertaining to the Characteristics of the Sites of HIV Integration Highlighting its role in Clinical Research and its Future with AI: A Review

Minakshi Boruah

Computer Science and Engineering
National Institute of Technology Mizoram
Aizawl, India
minakshi.cse.phd@nitmz.ac.in

Dr. Ranjita Das

Computer Science and Engineering
National Institute of Technology Mizoram
Aizawl, India
rdas@nitmz.ac.in

Abstract—Human Immunodeficiency Virus (HIV) is one of the variants of the retrovirus that integrates into the human genome and affects the human immune system. HIV-1 is the common HIV that is studied. Previous studies have revealed that the HIV integration site plays a key role in the treatment of the fatal disease as it is critical in the latent viral reservoir formation process. This reservoir is the roadblock in the search of a cure to HIV. The whole life cycle of the virus depends on this site. In the clinical studies pertaining to the provirus integration sites, a wealth of data is collected from routine clinical care of HIV patients in the form of medical records to analyze HIV integration targeting to combat the immunity to antiretroviral therapy (ART). A novel futuristic review of the researches done on HIV, particularly on works done to analyse the integration sites, has been presented here. This study concludes that building upon previously collected data and progressing towards improvement by the incorporation of the latest technologies will unravel new information that can help in the betterment of the future prospectives of a breakthrough in the cure of HIV/AIDS.

Index Terms—Integration sites, HIV, ART, DNA sequence, LEDGF/p75

I. INTRODUCTION

HIV integration sites (IS) have a major play in the viral infection and the whole life cycle of the retrovirus [1] [2]. It determines the viral rebound after antiretroviral therapy (ART) is disturbed [3] [4] [5] [6]. Furthermore, the persistence of the infected cells is also determined by the HIV IS [7] [8].

Few in-depth laboratory researches have been done for the identification of HIV IS [5] and most researches have been done for finding out

other related activities after using the isolated IS in different types of cells as input i.e. to analyze the evolution of the virus after integration [9]. In fact, very less has been done for the study of the integration sites except for the integration targeting in ART [10]. The existing studies only statistically work towards integration targeting for ART that is to say they focus on the already infected cells and study the after-effects of the HIV integration pathway based on the analysis of various ‘other’ factors [4] [11]. Very few studies exist that solely study the exact sites of HIV integration and most take the ART into center stage. Clinical trials with the novel diagnostic technologies involving only experimental and biomarker epidemiology fall short in contributing to the goal of UNAIDS 95-95-95 to completely end the AIDS disease by 2030. The reason being the current rate at which new protein sequences are being generated is very fast. An improved automatic method [12] will serve the purpose in a better way. The scope of this survey was limited to HIV/AIDS medical care, clinical research and the scarcely available AI studies focused upon the analysis of genotype or phenotype factors for HIV integration site targeting. As there are very few studies involving AI, this review is devoted to addressing the following question: “What progresses have been made to predict the IS of a HIV provirus?” majorly based on the studies using statistical modeling or laboratory experiments as the mean. This addressing of the issue is eventually culminated with a discussion about ML (Machine learning) based approaches in the field.

II. IMPORTANCE OF PREDICTION OF INTEGRATION SITES OF A HIV PROVIRUS

Gary et al. [1] indicated the importance of integration sites with the computational prediction of nucleosomes using massively parallel pyrosequencing to find out the relation between major grooves in chromatin and sites of HIV integration and to show that the effects of histone modifications on the HIV integration process was partially independent of other genomic features incorporated during the integration process. The pyrosequencing and some methods that they called as the bioinformatic methods were used for the prediction of nucleosomes occupancy positions, annotation of each base pair with statistical tools for its likelihood of hosting integration for periodicity analysis of nucleosomes and also, correlation of integration sites with the ENCODE results. They did this to imply that investigating many diverse aspects of retroviral DNA integration is useful. They stressed the importance of the selection of acceptor integration target sites for both the host and the virus. Pyrosequencing is a process using the sequencing by synthesis of DNA polymerase along with light detection. The process is depicted in Fig 1. Prior to pyrosequencing, Jurkat T cells were infected with the HIV-based vector. Then the fact of DNA sequences guiding nucleosome positioning [13] was used and the placement of nucleosomes on chromosomal regions hosting integration events was mapped using the nucleosomes positioning prediction tool available at <http://genie.weizmann.ac.il/pubs/nucleosomes06/index.html>. Illustration of the nucleosome-DNA interaction model thus formed by the lab is presented in Fig 2¹.

They referred to the work by Segal et al. [14] for this. Pyrosequencing helps in clarifying the nucleosomal structure. Alignment of the integration site patterns relative to the nucleosome's symmetrical centre revealed that integration is favored at the backbone sites of phosphate at the edges of major grooves which face outwards. It was the first to study whether integration in host-cell chromosomal DNA takes place in a nucleosome-wrapped region which matches in-vivo DNA. As these are lab based methods, we outline how these are intensely based on new data and are expensive every time in comparison to learning it with ML-based techniques. The vector of viral stock was prepared by transfection with the HIV vector segment along with the packaging or helping viruses into the 293T cell line. The Viral supernatant was then harvested

for thirty-eight hours after transfection, filtered through 4.5 nanometers sized filters, treated with DNase I, concentrated and frozen. DNA from host-virus junctions was prepared using PCR. The PCR products were gel purified, dissolved in an enzyme pool, and dispatched to 454 Life Sciences for pyrosequencing. Integration sites were accepted for reporting only if the sequences beginning within the 3 bp of HIV LTR ends had greater than 98% sequence match; it was aligned for the best match to the human genome (hg17) using BLAT. BLAT interface²³ is presented in Fig 3. In addition a transcription-based profiling was performed using Affymetrix microarrays which use gene chip probe arrays which we are not concerned with while dealing with the integration sites. Analysis in the ENCODE Consortium Encyclopedia of integration site positions revealed the epigenetic associations of the integration sites. They finally inferred that nucleosome-wrapped chromosomal DNA is the in-vivo integration target, and other finding unrelated to IS (like targeting by PSIP1/LEDGF/p75 , HIV cure with LEDGINs will only affect some part of HIV-1 expression and not form a complete cure).

Debyser et al. [15] highlighted the Hurdles in the search for an HIV-1 Cure. They focused on patients undergoing combination antiretroviral therapy (cART). They mentioned the difficulty being the failure of current treatments to eradicate HIV as this infected DNA persists in long-lasting cellular reservoirs and inducing a viral rebound whenever the treatment is disturbed. In their review article, they focused on post-integration latency, where cells contain an integrated provirus that is latent as far as the generation of viral fragments is concerned. They analyzed the role of LEDGIN-mediated blocking-locking of epithelium-derived growth factor p75 (LEDGF/p75) based on previous studies. This use of LEDGIN as the antiviral to block the interaction of p75 and Integrase forms part of what is called retargeting. They pointed out that LEDGF/p75 determines HIV integration site selection more than any other protein does.

Debyser et al. in their other paper [16] mentioned that the retroviral family preferentially integrates near a unique genomic profile locality. They provided for validation of LEDGF/p75 as an antiviral target and highlighted the efforts undertaken to develop small-molecule-based inhibitors

¹Obtained from the web-link:
<http://genie.weizmann.ac.il/pubs/nucleosomes06/index.html>.
²<http://genome.ucsc.edu/>
³<https://www.ncbi.nlm.nih.gov/nuccore/284370810>

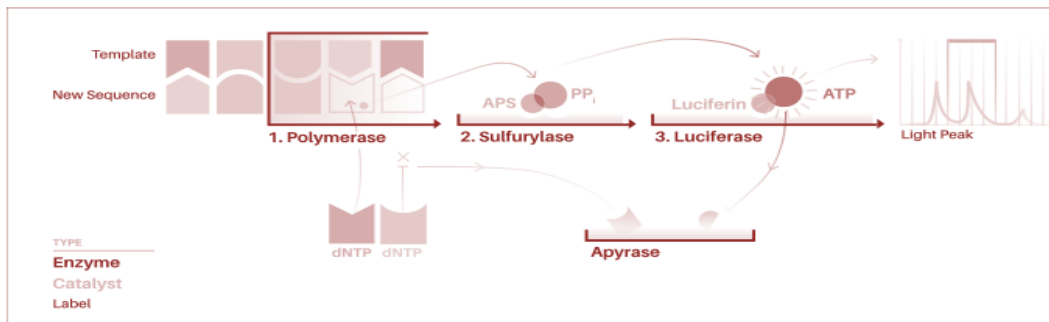


Fig. 1: The process involved in Pyrosequencing.

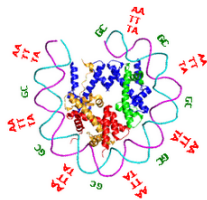


Fig. 2: Nucleosome-DNA interaction model predicted by the Segal Lab

of the interaction between HIV integrase (IN) and LEDGF/p75. During HIV replication it is mentioned that LEDGF/p75 has been identified as the exclusive tether of lentivirus integration with IN associated with the viral genome into the host chromatin. They were referring to some colocalization studies for this finding. Also this review paper mentioned LEDGF/p75 binds and protects IN from degradation and stimulates the catalytic activity of being in in-vivo as well as in-vitro. They mostly detailed the mechanism of targeting IN and its interaction with LEDGF/p75 (an antiviral tethering factor and cofactor target of many similar studies mentioned in the review paper) to the transcription region for HIV integration. They finally discussed some proposal for a safer MLV-derived vector for gene-therapeutic application. The HIV integration site was mentioned a few times for the emphasis of the role of LEDGF/p75 in facilitating of HIV integration.

Vansant et al. [17] confirmed the findings that the antivirals LEDGINs block the interaction between the co-factor LEDGF/p75 and the HIV-1 enzyme called integrase but they emphasized that closeness of silent proviruses to the epigenetic chromatin markers are associated with RNA expression and thus there is an association between provirus integration site and viral expression. It was a lab-

based research as to support their findings they did cell culture of SupT1 and Jurkat cells using the RPMI medium with 10% (v/v) fetal bovine serum and 0.01% (v/v) gentamicin. HEK293T cells were cultured in Eagle's medium modified with Dulbecco with five% (v/v) FBS (GIBCO) and 0.01% (v/v) gentamicin (GIBCO). For the viral Vector production Linear polyetylenimine (PEI) was used to cotransfect HEK293T cells with the packaging pHCC1. Other methods like Transduction, DNALibrary preparation, sequencing, Flow cytometry, cell sorting, qPCR(quantitative polymerase chain reaction), Bisulfite cytosine methylation analysis, RNA sequencing, Gene enrichment analysis, ChIP-sequencing and mRNA purification were incorporated along with LEDGIN CX014442 as the treatment agent for inhibiting LEDGF/p75. Chromatin immunoprecipitation (ChIP) is a precipitation method to study the interactions of proteins with the specific DNA regions. This requires the work by a specific antibody. Insertion sites were retrieved with more confidence using the transduction method of Barcoded HIV-ensembles (BHIVE). They found that the centrifugation with LEDGIN CX014442 during infection significantly changed and balanced off the chromosomal distribution as measured statistically in terms of the Chi-square test ($P < 0.0001$) and also increases the distance of integration sites to the chromatin H3K36me3. This work also mentions that HIV-1 integration is dependent on the LEDGF/p75 binding to H3K36me3 chromatin marker near the enhancer regions.

Patro et al. [6] narrowed down on the fact that the state of the persistent HIV provirus populations in the infected CD4+ T cells i.e. the structure of the proviral sequence in the latent reservoir along with its clonality has a link to the position of the integration sites. To reveal this link they used multiple-displacement amplification (MDA, see Fig 4) of cellular DNA which was already

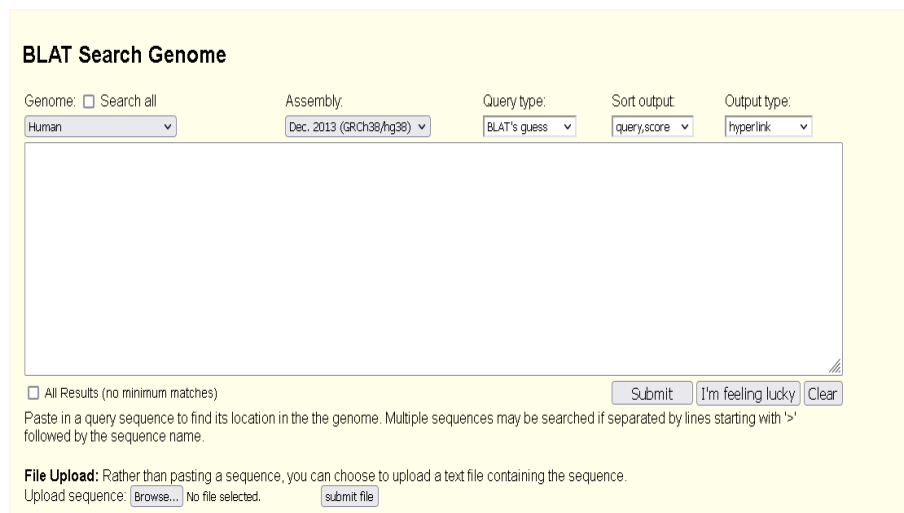


Fig. 3: Illustration of the BLAT interface.

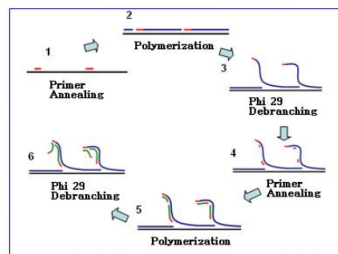


Fig. 4: MDA in progress.

diluted to a proviral endpoint so as to obtain the whole length of the proviral sequences for what is called “viral reconstruction” and their corresponding integration sites. The MDA was combined with Single-genome sequencing (SGS) and integration site analysis (ISA) to get the information regarding the structure of the persistent or clonal HIV. This was because the standard SGS has demerit of not being able to establish clonality of the provirus. So, they were finding out the origin of clonal proviral DNA And this has an implication for the removal of latent reservoir which is the obstacle to HIV treatment. They found that identical proviral sequences can result from both cellular expansion and viral genetic bottlenecks occurring prior to ART initiation and following ART failure. They applied the above mentioned methods starting with SGS to the lymph node and peripheral blood mononuclear cells of 5 ART undergoing donors to determine whether groups of identical sub-genomic proviral sequences are the result of

viral genetic bottlenecks or cellular expansion. SGS encompasses the region encoding P6, protease and reverse transcriptase (P6–PR–RT). They found that the identical sub-genomic proviral sequences can result from both viral genetic bottlenecks and cellular expansion occurring prior to ART initiation and following ART failure. A similar work was done by Brandt et. al. [13] who used the integration sites to predict the Clones of HIV-1-Infected Cell.

From the works discussed till now that targeting of integration sites with the help of various factors like LEDGIN mediated retargeting, analysis constrained in subtelomeric and pericentromeric regions, introducing new sequencing techniques etc the focus is not as highly constrained to the discovery of integration sites as is required. Given that we have discussed diverse experimental works, now we conclude briefly outlining what the experimental works that are similar in nature are doing. Most work target and characterize integration sites characteristics using the ART-based tools validating some phenomena as per preset constraints [18] [19] [20] [21] [22]. Given that a cure has not been obtained yet with the experimental methods some strong theory based automation may help to find some knowledge into the working of the HIV-1 virus. Due to the lack of studies based on DL or even ML focusing on IS, given the importance of IS and given the targets (UNAIDS 95-95-95) for the treatment of AIDs, there is need for some automation to speed up the processes involving much of already available large quantity of resources. So there is a need to redirect towards

ML based techniques.

III. CURRENT STATUS OF RESEARCH WORK DONE WITH ML FOR 'PREDICTION' OF INTEGRATION SITES OF AN HIV PROVIRUS

There is only one work solely dedicated to predict the integration sites of the HIV provirus [2]. Other related works are discussed below. Berry et al. [23] compared in a constrained way, his methodology with a machine learning strategy [24] that splits a chromosome again and again so as to get the gaps which may host many integration sites (IS) for the HIV-based vectors. He found Clumps by scan statistics based method (an ML based clustering method). His approaches found localized clusters of integration sites and provided the way forward for the comparison of the different clusters of integration sites with incorporated ART vectors.

In the work by Shen et al. [25], analysis based on the RF regression and KNN were performed on geno-phenotype data for HIV reverse transcriptase (RT) and protease (PR). This was done in order to predict the resistance against the HIV. Even though this paper used machine learning in explicit way it did not use it for the prediction of integration sites.

In their paper, Santoni et al. [26] used a few markers and identified markers with the highest F scores by combinatorial calculation of the markers and generated a supermarker. This was based on their derivation of efficiency of the F score in ranking and recognizing genetics involved with retrovirus integration site. This was done to accomplish the development of a method for displaying and detecting associations between chromatin and integration sites of the retrovirus. But these works were based on feature engineering-based methods. And as mentioned before there is a need for the machine learning to automate the process of classification of the enormous data related to the HIV integration sites to help in further treatment to eradicate HIV.

As discussed by Halin et. al. [2], who did a research based on deep learning and attention framework solely to extract the integration sites with Convolutional Neural Networks (CNNs), random forest-based model performed poorly even with extended input incorporating surrounding genomic attributes. The same was the case with the Score20 [23] the method which is a position weight matrix-based (PWM) conventional method using a lesser context. They [2] compared it with three baselines methods viz. logistic regression fitting [27], random forest construction [28] and gradient boosting decision tree (GBDT) construction [29]

[30] to come to such a conclusion. In their paper [2], Halin et. al. used a deep learning and attention-based framework. The main components used in their architecture is a dense neural network having the convolutional network.

As we can clearly observe that there is a dearth of ML-based work for the HIV integration sites identification and the datasets generated are huge after each experiment and are increasing the load of genomic consortium database like Gene Expression Omnibus. In addition taking into account the costs incurred in such experiments, there must be some studies done in order to combine the power of Deep Learning (DL) [9] [11] into the existing results and put to use the knowledge already available.

Deep learning architectures for HIV integration site learning and prediction have robust metrics already available, which are AUC-ROC, AUC-PR and F_β score [31] [2] for analysis of the results in comparison to the self-designed statistical tests like conservative estimate [6].

IV. CONCLUSION HIGHLIGHTING THE CONTRIBUTION OF THIS WORK

Keeping in view the fact that a very less amount of work exists with Deep Learning, we hope that some DL methods for the analysis of the integration sites will be developed very soon. So, the paper highlights that no work has been done to 'predict' sites with advanced DL frameworks.

So we can also summarily say that there is a problem in the detection of the integration sites selected by the HIV for integration and it has been done solely in lab-experiments [32]. However, AI especially in bioinformatics is really very effective tool to get faster, accurate and highly insightful results while saving time and money [9] [11] consumed in the procurement and synthesis and the overall experiment. Also, DL will use the pre-existing data to take comprehensive advantage of it rather than repetitively mining out new data from scratch. This repetitive discovery of new data to some extent also redundantly spans over many studies as we have seen in which were analyzing the role of LEDGIN, which need to be organized before proceeding with the next stock of experiments. AI techniques like DL can help by automatically detecting the useful features.

V. CONTRIBUTION

It is the first and the only work solely dedicated towards identifying the HIV integration sites while emphasizing the criticality of the role played by the sites. Also, this is the only paper trying to

target DL tools throwing light towards the path of the new-age hope of using the AI to its extreme benefit which is the true use of informatics. This paper further suggest that rather than beating around the bush stressing on the factors like PSIP1/LEDGF/p75, it is better to block the integration sites.

REFERENCES

[1] G. P. Wang, A. Ciuffi, J. Leipzig, C. C. Berry, and F. D. Bushman, "Hiv integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications," *Genome research*, vol. 17, no. 8, pp. 1186–1194, Aug 2007.

[2] H. Hu, A. Xiao, S. Zhang, Y. Li, X. S. T. Jiang, and L. Zhang, "Deephint: understanding hiv-1 integration via deep learning with attention," *Bioinformatics*, vol. 35, no. 10, pp. 1660–1667, September 2018. [Online]. Available: <https://doi.org/10.1093/bioinformatics/bty842>

[3] J. K. Wong *et al.*, "Recovery of replication-competent hiv despite prolonged suppression of plasma viremia," *Science*, vol. 278, no. 5341, pp. 1291–1295, 1997.

[4] A. S. Huang, V. Ramos, T. Y. Oliveira, C. Gaebler, M. Jankovic, M. C. Nussenzweig, and L. B. Cohn, "Integration features of intact latent HIV-1 in CD4+ T cell clones contribute to viral persistence," *Journal of Experimental Medicine*, vol. 218, no. 12, 10 2021, e20211427. [Online]. Available: <https://doi.org/10.1084/jem.20211427>

[5] Z. Debyser, G. Vansant, A. Bruggemans, J. Janssens, and F. Christ, "Insight in hiv integration site selection provides a block-and-lock strategy for a functional cure of hiv infection," *Viruses*, vol. 11, p. 12, 12 2018.

[6] S. C. Patro, L. D. Brandt, M. J. Bale, E. K. Halvas, K. W. Joseph, W. Shao, X. Wu, S. Guo, B. Murrell, A. Wiegand, J. Spindler, C. Raley, C. Hautman, M. Sobolewski, C. M. Fennessey, W.-S. Hu, B. Luke, J. M. Hasson, A. Niyongabo, A. A. Capoferri, B. F. Keele, J. Milush, R. Hoh, S. G. Deeks, F. Maldarelli, S. H. Hughes, J. M. Coffin, J. W. Rausch, J. W. Mellors, and M. F. Kearney, "Combined hiv-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors," *Proceedings of the National Academy of Sciences*, vol. 116, no. 51, pp. 25 891–25 899, 2019. [Online]. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.1910334116>

[7] F. Maldarelli *et al.*, "Specific hiv integration sites are linked to clonal expansion and persistence of infected cells," *Science*, vol. 345, no. 6193, pp. 179–183, 2014.

[8] T. A. Wagner *et al.*, "Proliferation of cells with hiv integrated into cancer genes contributes to persistent infection," *Science*, vol. 345, no. 6196, pp. 570–573, 2014.

[9] M. A. Younis, I. A. Khalil, and H. Harashima, "Gene therapy for hepatocellular carcinoma: Highlighting the journey from theory to clinical applications," *Advanced Therapeutics*, vol. 3, no. 11, p. 2000087.

[10] F. Spyarakis, M. Fornabaio, P. Cozzini, A. Mozzarelli, D. J. Abraham, and G. E. Kellogg, "Computational titration analysis of a multiprotic hiv-1 protease-ligand complex," *Journal of the American Chemical Society*, vol. 126, pp. 11 764–11 768, Sep 2004.

[11] A. Shukla, N.-G. P. Ramirez, and I. D'Orso, "Hiv-1 proviral transcription and latency in the new era," *Viruses*, vol. 12, no. 5, 2020. [Online]. Available: <https://www.mdpi.com/1999-4915/12/5/555>

[12] S. Makrodimitris *et al.*, "Automatic gene function prediction in the 2020's," *Genes*, vol. 11, no. 11, pp. 1264–1281, Sep 2020.

[13] L. D. Brandt, S. Guo, K. W. Joseph, J. L. Jacobs, A. Naqvi, J. M. Coffin, M. F. Kearney, E. K. Halvas, X. Wu, S. H. Hughes, and J. W. Mellors, "Tracking hiv-1-infected cell clones using integration site-specific qpcr," *Viruses*, vol. 13, no. 7, 2021. [Online]. Available: <https://www.mdpi.com/1999-4915/13/7/1235>

[14] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thrm, Y. Field, I. Moore, J.-P. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, pp. 772–8, 09 2006.

[15] Z. Debyser, G. Vansant, A. Bruggemans, J. Janssens, and F. Christ, "Insight in hiv integration site selection provides a block-and-lock strategy for a functional cure of hiv infection," *Viruses*, vol. 11, p. 12, 12 2018.

[16] Z. Debyser, F. Christ, J. De Rijck, and R. Gijssbers, "Host factors for retroviral integration site selection," *Trends Biochem. Sci.*, vol. 40, no. 2, pp. 108–116, Feb. 2015.

[17] G. Vansant, H.-C. Chen, E. Valera Zorita, K. Trejbalov, Mikl G. Filion, and Z. Debyser, "The chromatin landscape at the hiv-1 provirus integration site determines viral expression," *Nucleic Acids Research*, vol. 48, 06 2020.

[18] K. B. Einkauf, M. R. Osborn, C. Gao, W. Sun, X. Sun, X. Lian, E. M. Parsons, G. T. Gladkov, K. W. Seiger, J. E. Blackmer, C. Jiang, S. A. Yuki, E. S. Rosenberg, X. G. Yu, and M. Lichterfeld, "Parallel analysis of transcription, integration, and sequence of single hiv-1 proviruses," *Cell*, vol. 185, no. 2, pp. 266–282.e15, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0092867421014495>

[19] J. K. Yoon, J. R. Holloway, D. W. Wells, M. Kaku, D. Jetton, R. Brown, and J. M. Coffin, "Hiv proviral dna integration can drive t cell growth ex vivo," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 32 880–32 882, 2020. [Online]. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.2013194117>

[20] J. van Haasteren, A. M. Munis, D. R. Gill, and S. C. Hyde, "Genome-wide integration site detection using cas9 enriched amplification-free long-range sequencing," *Nucleic Acids Research*, vol. 49, no. 3, pp. e16–e16, 12 2020. [Online]. Available: <https://doi.org/10.1093/nar/gkaa1152>

[21] P. K. Singh, G. J. Bedwell, and A. N. Engelman, "Spatial and genomic correlates of hiv-1 integration site targeting," *Cells*, vol. 11, no. 4, 2022. [Online]. Available: <https://www.mdpi.com/2073-4409/11/4/655>

[22] Y.-H. J. Yeh, K. Yang, A. Razmi, and Y.-C. Ho, "The clonal expansion dynamics of the hiv-1 reservoir: Mechanisms of integration site-dependent proliferation and hiv-1 persistence," *Viruses*, vol. 13, no. 9, 2021. [Online]. Available: <https://www.mdpi.com/1999-4915/13/9/1858>

[23] C. Berry *et al.*, "Selection of target sites for mobile dna integration in the human genome." *PLoS Computational Biology*, vol. 2, no. 11, pp. e157–e170, Nov 2006.

[24] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based dna copy number data," *Biostatistics*, vol. 5, pp. 557–572, Oct 2004.

[25] C. Shen, X. Yu, R. W. Harrison, and I. T. Weber, "Automated prediction of hiv drug resistance from genotype data," *BMC Bioinformatics*, vol. 17, no. 8, pp. 557–572, Aug 2016.

[26] F. A. Santoni, O. Hartley, and J. Luban, "Deciphering the code for retroviral integration target site selection," *PLOS Computational Biology*, vol. 6, no. 11, pp. 1–20, Nov 2010. [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1001008>

[27] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression-A Self-Learning Text*.

- David G. Kleinbaum, Mitchel Klein, Department of Epidemiology, Emory University, Atlanta, GA 30333, USA: Springer-Verlag, 2002.
- [28] Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning*. Machine Learning Department, NEC Labs America: Springer, Boston, MA, 2012.
- [29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.
- [30] P. Xuan, C. Sun, T. Zhang, T. S. Yilin Ye, Y. Dongl *et al.*, "Gradient boosting decision tree-based method for predicting interactions between target genes and drugs," *Frontiers in Genetics*, vol. 10, pp. 459–469, May 2019. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fgene.2019.00459>
- [31] S. Farquhar, M. Osborne, and Y. Gal, "Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning," 2021.
- [32] M. Lusic and R. Siliciano, "Nuclear landscape of hiv-1 infection and integration," *Nature Reviews Microbiology*, vol. 15, no. 2, pp. 69–82, Feb 2017.

Design and Development of a smart garage door system

Mohamed Imran Mohamed Ariff
Department of Computer
Science, Faculty of Computer &
Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
moham588@uitm.edu.my

Farah Diyana Mohamad Fadzir
Department of Computer
Science, Faculty of Computer &
Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
farah.diyana.fadi@gmail.com

Noreen Izza Arshad
Department of Computer &
Information Sciences, Faculty of
Science and Information
Technology, Universiti
TEKNOLOGI PETRONAS, Seri
Iskandar, Perak, MALAYSIA
noreenizza@utp.edu.my

Samsiah Ahmad
Department of Computer
Science, Faculty of Computer &
Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus
MALAYSIA
samsi260@uitm.edu.my

Khairulliza Ahmad Salleh
Department of Computer
Science, Faculty of Computer &
Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
khair279@uitm.edu.my

Jufiza A. Wahab
Department of Mathematics
Faculty of Computer &
Mathematical Sciences
Universiti Teknologi MARA
Perak Branch, Tapah Campus,
MALAYSIA
jufiz279@uitm.edu.my

Abstract—This paper presents the design and development of a smart garage door system, which is operated by an Arduino microcontroller via the use of a mobile application and the Blynk cloud sever. Further, this mobile application allows the smart garage door to be controlled and accessed from any remote location via the use of the Blynk cloud server which is connected to the Internet using Wi-Fi or 3G/4G network. The operations of this smart garage door also function using the Google assistant voice command. Finally, this smart garage door application has been tested and it is able to successfully perform the basic operations of a smart garage door as proposed in the initial design and development stage.

Keywords—smart home, IoT, smart garage door, Arduino, mobile application

I. INTRODUCTION

Internet of Things (IoT) has become a major computing paradigm and the latest disruptive technology after the Cloud computing. The ability of IoT to connect with other systems and applications, from low to high level has prompted and promoted an unprecedented array of applications in all fields of human activity from science, engineering, business, health, leisure and everyday life, especially for home usage. In recent years, there has been a growing interest in smart home systems. The concept of smart home system is the ability to automate most of the household appliances with the use of network technology that enables users to perform tasks before arriving at their home with minimum human intervention [1, 2].

The development of smart home provides a remote interface for home appliances via network technology (i.e., wireless transmission, the internet and android application), to provide control and monitoring via the use of mobile applications [3, 4]. Furthermore, studies have also suggested the development of smart home system, enhanced power efficiency and improves the quality of living [5, 6]. In line with the development of smart homes, the term Internet of Things (IoT) appeared and began to spread widely [7]. The concept of IoT is defined as the interconnection network system between everyday household objects (e.g., house appliances, house doors, and garage doors) [8, 9]. The use of IoT in the smart home development has also promoted: (1) better security system and (2) convenience for home users. One of the popular IoT adoption in smart home project is the smart garage door system [10-12]. The development of this system help promotes the life safety and convenience of a home resident. The concept of the smart garage door allows users to operate the door remotely via the use of networking protocols and IoT sensors [13, 14]. Furthermore, the smart garage door mitigates the door operation, as the users can access the garage from anywhere with the help of mobile application installed inside smart phone devices. This in turn helps reduce the manual operation of the garage door opening and closing. The development of the smart garage door system also helps reducing as trespassing as the opening and closing of the door is controlled by the home owners via the mobile of a mobile application [15]. Furthermore, the use of IoT sensors has also made it possible for home owners, to remotely control the garage door from anywhere via a smartphone [16]. Although, previous literature has highlighted the benefits of the smart garage door system, previous literature has also stressed the development cost of this system is high and the energy usage of this system is inefficient, thus limiting the implementation of smart garage doors in residential area.

II. LITERATURE REVIEW

A. Garage door system

The use of IoT smart garage doors is an easy process to safeguard the security of residential premises [17]. In recent years, the popularization of the IoT technology has increased the: (1) safeguards, (2) security and (3) easy use of a garage gate. Furthermore, the use of IoT technology in the garage gate has promoted several advantages [18, 19]. Among the advantages highlighted in previous literature is: (1) door can be opened and closed from any remote places [20, 21], (2) promotes security as only authorized residents can enter the premises and (3) auto mode shut down if the garage door is left open for a certain period. Despite the above mentioned, advantages of incorporating the use of IoT technology in the garage door, there are 2 main disadvantages or problems reported in the literature. The problems are: (1) networking issues and (2) security and privacy issues. In terms of the networking issues, the network can be hacked by unauthorized personnel thus gaining access to the premises [15, 22], and in terms of the privacy issue, users or home residents are uninformed about any error that may occur while the garage door is being opened or closed [23, 24]. (e.g., the door has not closed properly) [25, 26]. Based on the disadvantages highlighted above, realization of the smart garage door needs a better design and development effort [27, 28].

B. Current issues and proposed solution

Previous literature highlights that the development of the smart garage door system is to meet the resident goals of comfort living, life safety, security and efficiency. Furthermore, this system is also a part of the smart home overall system, and can be controlled with the use of Internet (via mobile application). However, there are numerous issues relating to the usage, design and development cost and implementation of the smart garage door system. Among the issues presented in the literature, the design and development cost in the implementation of the smart garage door is frequently raised [29, 30]. Previous research also highlights the evolving state of IoT development has created several problems in the areas of: (1) user home device integrations, (2) high implementation cost and (3) customization of IoT device integration between different types of users [10, 15, 31].

Thus, this research project proposes a new and lower cost prototype solution to design and develop a smart garage door. Further, this research project is to develop a smart garage door - based on Android smartphones access through the use of a simple mobile application system. The aim of this research project is to design and develop an Android mobile application as an input to open and close garage doors using the Wi-Fi or 3G or 4G. The garage door may be controlled if there is a network signal in that location. This mobile application also utilizes the Google Assistant function as an additional feature in opening or closing the garage door. This paper is organized as follows; the next section will illustrate the methodology employed in the design and development of this research project. Each phase of the methodology will be briefly explained. Following that, is the results and discussion section. In this section, the appropriate tests will be highlighted and explained. Finally, a short conclusion and future areas for this research project will be presented.

III. METHODOLOGY

The following subsections will briefly illustrate the methodology process of this research project.

1) IOT development

This research project adopts a modified version of the methodology proposed by Fahmideh and Zowghi [32]. Further, this methodology has been in favour for various IoT projects [31, 33]. Figure 1 (refer below) illustrates in brief the overall flow of the methodology adopted in this research project.

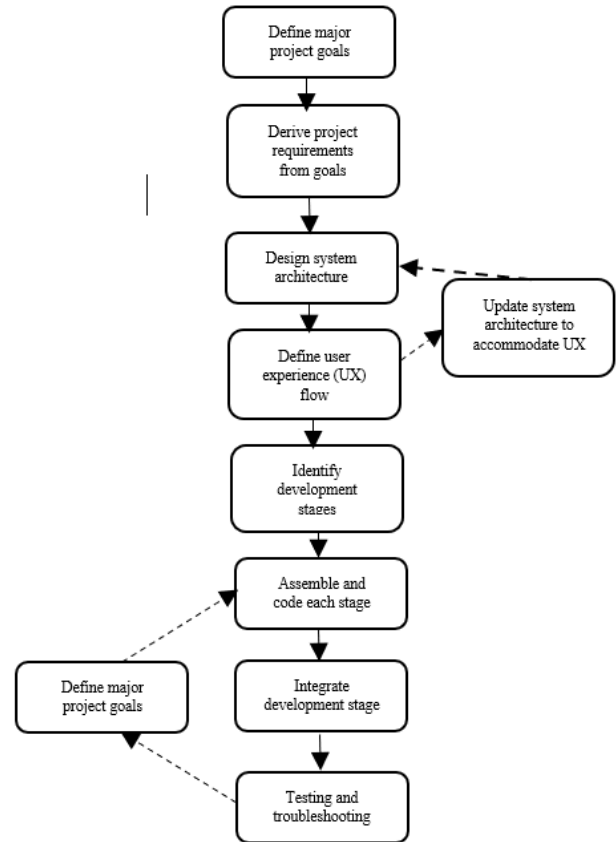


Fig. 1. IoT project development approach

This research project first begins by defining the main objectives that needs to be accomplished. Following that, the research project is broken down into specific requirements (i.e., software, hardware, and all other important tasks). The hardware and software requirement for this project is highlighted in figures 2 and 3, respectively. Next, the overall research project architecture is constructed. The IoT system architecture is comprised of various functional layers to enable IoT applications, including a sensing layer, network layer, and application layer (refer figure 3). Following that, appropriate user interfaces are created as this system is a mobile application. Each user interface is well thought out, to meet each specific requirement in the design and development of the smart garage door project. Examples of the user interface is shown in figures 4 and 5, respectively. The next stage is prototype development stage.

This stage encapsulates the overall components (i.e.: hardware, software, networking, and their integration) along with their interactions within the research project. In this stage, the chosen hardware and software will then be assembled and coded accordingly. The codes will also be linked together with the user interface designed prior to the prototype development. This smart garage door runs on the Arduino IDE, integrated with Blynk (refer figure 6).

No	Hardware	Specification
1	Laptop HP	Processor: Intel® Core™ i5-1035G1 CPU @ 1.00GHz 1.19 GHz RAM: 8.00 GB
2	Honor 50 Lite	CPU: Octa-core (4x2.0 GHz Kryo 260 Gold & 4x1.8 GHz Kryo 260 Silver) Chipset: Qualcomm SM6115 Snapdragon 662 (11 nm)
3	Things board	Esp32
4	Garage door	Servo MG 995
5	RFID Scanner	RFID RC-552
6	LED light	Green and red

Fig. 2. Hardware requirement

No	Software	Details
1	Android 11	The operating system for mobile phone
2	Microsoft Word	As documentation to report the project
3	Windows 10	The operating system for laptop
4	Arduino	Software design using C++ and Java
5	Blynk	The mobile application for the project

Fig. 3. Software requirement



Fig. 4. Smart garage door interface

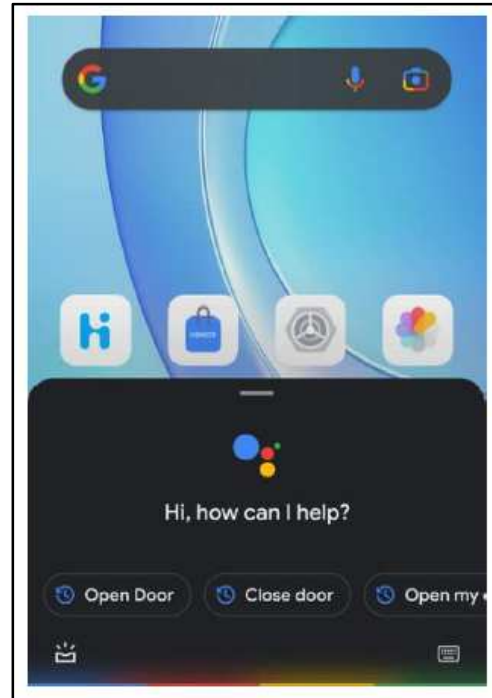


Fig. 5. Smart garage door interface (Google assistance)

2) IoT system architecture

The designing of an IoT system architecture should conform to the most needed systems requirements regarding horizontal scalability when the system is used multiple times. Furthermore, the designing of a IoT system architecture should have the following features (if possible) [34]:

1. Flexibility – solution must ensure better user experience and interface simplicity
2. Evaluation and prediction – access data will be stored and eventually used for better resource management and further improvement of balance between individual comfort and overall work efficiency
3. Energy savings – better utilization of power consumption resources

Base of the features above, this research project adopts the IoT architecture (refer figure 2) by [35] which comprises of: (1) the perception layer (2) the network layer and (3) the application layer. Each layer is then linked to a specific source code embedded inside Arduino IDE connected to the cloud server (i.e., Blynk), and the mobile application.

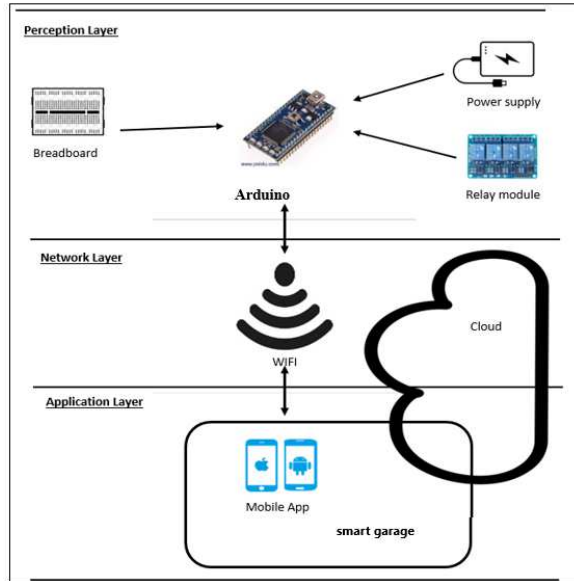


Fig. 6. IoT system architecture

3) Prototype design

The prototype design (refer figure 7) is a physical layout showing the overall functionality of the smart garage door system. There are two sections in this prototype, the: (1) hardware structure and (2) IoTs (sensors). This prototype design will show the actual flow of the mobile application, the IoT and the micro controller. This prototype is designed merely as a proof of concept in highlighting the design and development of this research project.



Fig. 7. Prototype design

IV. RESULTS AND DISCUSSION

This smart garage door was tested to demonstrate its overall feasibility and effectiveness. The smart garage door underwent three types of testing: (1) user interface, (2) user expectation and (3) overall satisfaction. Each of the test outcome is presented in the subsection below.

A. User interface

The purpose of this test is to get the testers (participants) feedback or their satisfaction level when using the smart garage mobile application - interface. The result is presented in Figure 8.

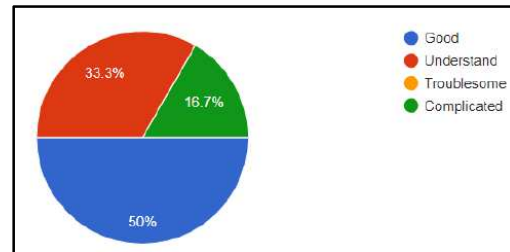


Fig. 8. User interface testing result

From Figure 8 above, it shows that the majority of respondents are satisfied with the design of the smart garage mobile application.

B. User expectation

The purpose of this test is to assess whether output presented by the smart garage mobile application is well understood by the respondents [38]. The user expectation also provides a clearer picture, as to highlight the design and development of the system. Based on the results (refer figure 9, 83% of the respondents can understand the output of the smart garage mobile application.

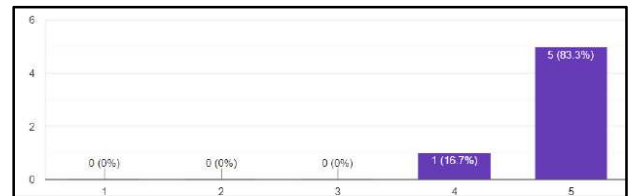


Fig. 9. User expectation

C. Overall satisfaction test

The overall satisfaction test presents the overall user satisfaction towards the smart garage mobile application. Based on the results (refer Figure 10) it indicates that users are very satisfied (score 5) with the research project. The results also indicate that this smart garage mobile application can help the users in overcoming their problems as highlighted in previous literature.

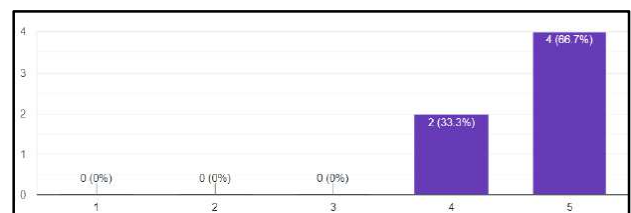


Fig. 10. Overall satisfaction test

V. CONCLUSION

In conclusion, the main aim of this research project is to design and develop a smart garage door via the use of a mobile application. Further, this project used the Arduino micro controller to control the system via IoT wireless connection. This research project will make accessing the garage door more convenient and faster with the help of the mobile application.

This research project contributes to homeowners by providing them with an alternative simple solution to design and development their own smart garage door system, as this research project provides homeowners with security, energy efficiently (low operating cost) and convenience. This research project also contributes to the literature as the design and development of the smart garage door mobile application system utilizes the Google Assistant feature. This feature is a particularly important feature when homeowners have limited motion access to mobile phone usage.

For future work, the security level criteria should be considered in designing the smart garage door system. The implementation of face detection, fingerprint, voice recognition should also be implemented together with the Internet of Things (IoT). The use of mobile application in this system should also be improved as the mobile application is the visible frontier of the system to end

ACKNOWLEDGMENT

This work is supported by the Department of Computer Science, Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Perak Branch under the final year project (FYP) 2021 / 2022.

REFERENCES

[1] M. S. Hossain, M. Rahman, M. T. Sarker, M. E. Haque, and A. Jahid, "A smart IoT based system for monitoring and controlling the sub-station equipment," *Internet of things*, vol. 7, p. 100085, 2019.

[2] T. Alam, A. A. Salem, A. O. Alsharif, and A. M. Alhejaili, "Smart home automation towards the development of smart cities," *APTİKOM Journal on Computer Science and Information Technologies*, vol. 5, no. 1, pp. 152-159, 2020.

[3] A. Van Berlo, "Smart home technology: Have older people paved the way," *Gerontechnology*, vol. 2, no. 1, pp. 77-87, 2002.

[4] P. Sethi and S. R. Sarangi, "Internet of things: architectures, protocols, and applications," *Journal of Electrical and Computer Engineering*, vol. 2017, 2017.

[5] M. A. E.-L. Mowad, A. Fathy, and A. Hafez, "Smart home automated control system using android application and microcontroller," *International Journal of Scientific & Engineering Research*, vol. 5, no. 5, pp. 935-939, 2014.

[6] J. Dutta and S. Roy, "IoT-fog-cloud based architecture for smart city: Prototype of a smart building," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, 2017: IEEE, pp. 237-242.

[7] R. J. Robles, T.-h. Kim, D. Cook, and S. Das, "A review on security in smart home development," *International Journal of Advanced Science and Technology*, vol. 15, 2010.

[8] T. Alam, "Cloud Computing and its role in the Information Technology," *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, vol. 1, no. 2, pp. 108-115, 2020.

[9] T. Alam, "Cmi computing: A cloud, manet and internet of things integration for future internet," *Tanweer Alam. CMI Computing: A Cloud, MANET and Internet of Things Integration for Future Internet.*, *JAMBURA JOURNAL OF INFORMATICS*, vol. 2, no. 1, 2020.

[10] Y. Jie, J. Y. Pei, L. Jun, G. Yun, and X. Wei, "Smart home system based on iot technologies," in *2013 International conference on computational and information sciences*, 2013: IEEE, pp. 1789-1791.

[11] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A survey on industrial Internet of Things: A cyber-physical systems perspective," *Ieee access*, vol. 6, pp. 78238-78259, 2018.

[12] D. Mocrii, Y. Chen, and P. Musilek, "IoT-based smart homes: A review of system architecture, software, communications, privacy and security," *Internet of Things*, vol. 1, pp. 81-98, 2018.

[13] H. Landaluce, L. Arjona, A. Perallos, F. Falcone, I. Angulo, and F. Muralter, "A review of IoT sensing applications and challenges using RFID and wireless sensor networks," *Sensors*, vol. 20, no. 9, p. 2495, 2020.

[14] S. Singh and N. Singh, "Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce," in *2015 International conference on green computing and internet of things (ICGCIoT)*, 2015: Ieee, pp. 1577-1581.

[15] B. Sanjay, N. Kaushik, V. Srinivasan, S. Prabhakar, and K. Jayavel, "Design and implementation of smart garage-An IoT perspective," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017: IEEE, pp. 2712-2716.

[16] G. S. U. Al-Mamun, "Analysis Of Sensors Used to Make Smart Homes," *LC International Journal of STEM (ISSN: 2708-7123)*, vol. 2, no. 4, pp. 1-6, 2021.

[17] M. Rouse and I. Wigmore, "Definition. Mobile App," Retrieved February, vol. 12, p. 2017, 2013.

[18] A. A. A. Sen and M. Yamin, "Advantages of using fog in IoT applications," *International Journal of Information Technology*, vol. 13, no. 3, pp. 829-837, 2021.

[19] T. L. N. Dang and M. S. Nguyen, "An approach to data privacy in smart home using blockchain technology," in *2018 International Conference on Advanced Computing and Applications (ACOMP)*, 2018: IEEE, pp. 58-64.

[20] D. Pavithra and R. Balakrishnan, "IoT based monitoring and control system for home automation," in *2015 global conference on communication technologies (GCCT)*, 2015: IEEE, pp. 169-173.

[21] S. Hong and H.-J. Sin, "Analysis of the Vulnerability of the IoT by the Scenario," *Journal of the Korea Convergence Society*, vol. 8, no. 9, pp. 1-7, 2017.

[22] X. Zhang, P. Huang, L. Guo, and Y. Fang, "Hide and seek: Waveform emulation attack and defense in cross-technology communication," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019: IEEE, pp. 1117-1126.

[23] A. Rahmati, E. Fernandes, K. Eykholt, and A. Prakash, "Tyche: A risk-based permission model for smart homes," in *2018 IEEE Cybersecurity Development (SecDev)*, 2018: IEEE, pp. 29-36.

[24] W. He, J. Martinez, R. Padhi, L. Zhang, and B. Ur, "When smart devices are stupid: negative experiences using home smart devices," in *2019 IEEE Security and Privacy Workshops (SPW)*, 2019: IEEE, pp. 150-155.

[25] E. Alihodzic and E. Sokic, "Development of a Wi-Fi based car gate remote control and supervision system," in *2020 28th Telecommunications Forum (TELFOR)*, 2020: IEEE, pp. 1-4.

[26] A. Hasibuan, R. Rosdiana, and D. S. Tambunan, "Design and Development of An Automatic Door Gate Based on Internet of Things Using Arduino Uno," *Bulletin of Computer Science and Electrical Engineering*, vol. 2, no. 1, pp. 17-27, 2021.

[27] M. Kim, B. Hilton, Z. Burks, and J. Reyes, "Integrating blockchain, smart contract-tokens, and IoT to design a food traceability solution," in *2018 IEEE 9th annual information technology, electronics and mobile communication conference (IEMCON)*, 2018: IEEE, pp. 335-340.

[28] F. Shawki, M. Dessouki, A. Elbasiouny, A. Almazroui, and F. Albeladi, "Microcontroller based smart home with security using GSM technology," *IJRET: International Journal of Research in Engineering and Technology*, vol. 4, no. 06, 2015.

[29] P. Mtshali and F. Khubisa, "A smart home appliance control system for physically disabled people," in *2019 Conference on Information Communications Technology and Society (ICTAS)*, 2019: IEEE, pp. 1-5.

- [30] R. K. Kodali, J. John, and L. Boppana, "IoT Monitoring System for Grain Storage," in 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020: IEEE, pp. 1-6.
- [31] W. Rafique, X. Zhao, S. Yu, I. Yaqoob, M. Imran, and W. Dou, "An application development framework for Internet-of-Things service orchestration," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4543-4556, 2020.
- [32] M. Fahmideh and D. Zowghi, "An exploration of IoT platform development," Information Systems, vol. 87, p. 101409, 2020.
- [33] M. Lin, C. Huang, Z. Xu, and R. Chen, "Evaluating IoT platforms using integrated probabilistic linguistic MCDM method," IEEE Internet of Things Journal, vol. 7, no. 11, pp. 11195-11208, 2020.
- [34] M. Pavelić, Z. Lončarić, M. Vuković, and M. Kušek, "Internet of things cyber security: Smart door lock system," in 2018 international conference on smart systems and technologies (SST), 2018: IEEE, pp. 227-232.
- [35] M. Wu, T.-J. Lu, F.-Y. Ling, J. Sun, and H.-Y. Du, "Research on the architecture of Internet of Things," in 2010 3rd international conference on advanced computer theory and engineering (ICACTE), 2010, vol. 5: IEEE, pp. V5-484-V5-487.

IronMan: An Android-Web Based Application for Laundry Services

Mohammad Moshfique Uddin, Rohit Roy, Saima Alam Miduri, Rashedur M. Rahman

*Department of Electrical and Computer Engineering
North South University*

Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh
{ moshfique.uddin, rohit.roy, saima.miduri, rashedur.rahman}@northsouth.edu

Abstract— The majority of people despise the chore of ironing clothes but enjoy wearing nice, crisp, wrinkle-free clothes. IronMan provides steam ironing services right to your door. IronMan is a service that supplies ironing and laundry services online. In this work, the system allows to create an innovative Internet of Things (IoT)-based Laundry Services E-commerce business model. This work will bring laundry service right to your door, in the palm of your hand. This work will include a proper Android app for accessing nearby laundry shop services from home and a related website to this android app. In this application, a machine learning model was used to create intelligent logistic management systems. Using this application, people can easily access laundry services while staying at home. Nowadays people are more inclined to market from online shops rather than going outside. The current initiative follows the same pattern for laundry service providers and consumers. In most laundry shops, they do not keep records of their customers. So there is a possibility that the clothes of the customers may mix up with one another. This system helps users to not only track their orders but also physically scan their items using image classification into their list with an image proof to avoid any mismatch and help customers claim if a mix-up happens. This application has a location tracker for riders to accept the nearest order so that the customer gets a smooth experience. It also has a website for users to create their account and check their services display and customers can order from both web and phone.

Keywords—*Android-Web, Internet of Things (IoT), Image Classification, Machine Learning.*

I. INTRODUCTION

Trying to imagine life without technology is nearly impossible. People have become so reliant on technology that it has become an inseparable part of our lives. The internet of things or IoT [1] influences people's lifestyle from how people react to how people behave from air conditioners that all people can control with smartphones, from using smartphones to get area addresses, ordering food using smartphones, etc. IoT method collects data from different uses of different software and devices and uses that data to make another software or device work better. The method IoT in recent days was implemented on other E-commerce platforms to have the best outcomes for the business. For this application, this method is used and an E-commerce platform [2] is built named Ironman which will provide laundry services from nearby accessible

shops. An application that acts as a go-between all the iron-laundry shops and people. IronMan allows you to order from anywhere at any time without having to pause everything. Scroll through the app to find your favorite shop and tap to order. This project includes a customer application, as well as a shop owner and delivery application. Alongside the project also have a website relevantly. In Bangladesh, approximately 50 million people use the internet on a daily basis.

People are increasingly inclined to use various technologies and applications. The objective of this work is to generate a smart iron-laundry application that will allow people to order from anywhere at any time. Delivery partners will arrive in less than 20 minutes to pick up your clothes. The entire washing, ironing, and folding process are completed quickly and efficiently. IronMan assignees are equipped with an app that indicates successful pickups and deliveries.

An Android application will be used to access the services. Using the shop owners' application, shop owners can create shops and add items to the database. Users can access it through the customer app; Riders will use the rider application. There is an image classification model [3] which can classify customer clothes to track the clothes' identity. There has been implemented a logistic regression machine learning model [4], for which a set of data was collected from people who were thinking about this project via a survey. This model can predict the sustainability of IronMan.

II. LITERATURE REVIEW

According to the most recent studies and research, many familiar applications and research journals are already related to this research work. However, those studies had some gaps that are hoped to be filled with the current initiative. Extensive research has been done and a few research articles are here to compare this work.

The authors in [5] describe Laundry Applications, and the primary goal of developing Laundry Applications is to create employment. The intended application is also helpful for working people, students, and people who live outside their homes and do not have time to wash their clothes. So, in essence, they are attempting to connect people (customers and

Dhobis) so that they can benefit from one another. There is a decent amount of services available in this application, so the customer can easily select the services, date of service, time of service, etc. They are creating their laundry shop and intend to help unemployed people.

They targeted people who are willing to order laundry services online. They are the administrator as well as the shop owner. Consequently, the customer does not have the option of selecting a specific shop from their list of options. They tried to be available everywhere, but they were not adaptable.

Here IronMan solved the problem with the shops. Because it is an open platform, and the primary goal is to act as a middleman between laundry shops and people. As a by-product, the app allows multiple shop owners to register their businesses. As a necessary consequence, people can select and order from any shop. IronMan also provides image classification, so people do not have to fear losing their clothes. It is a machine learning model that can detect images from the database and verify the owner of the cloth.

B. Project Background Related Application

According to our most recent observation on research journals, there are already a plethora of well-known applications are gettable on Google Play and the Apple App Store.

Here are a few examples of popular and similar applications:

- Sheba.xyz [6]
- Laundry Vai [7]

In Bangladesh, many more related applications are available. Those applications have a lot in common, but they also have some differences and limitations. For instance, in the Sheba.xyz app, one can only communicate with one store, and there is no other laundry shop. Laundry Vai and Sheba.xyz are nearly identical applications to ours, but both have some limitations. These applications only contain information about their own services, only available within Dhaka.

On the other hand, IronMan is a fantastic platform for laundry services, with information on all local laundry shops throughout Bangladesh. The earlier apps also do not allow customers to track their orders or provide convenient payment gateway services. The plan is to develop an application that would give the same benefits across the country.

C. Project Background Related Literature

The theoretical literature review of this project reviewed what other researchers had written related to our topic to compile, categorize, and evaluate it.

The authors in [8] discuss big data analytics, smart logistics management, and machine learning techniques used on an IoT platform for laundry services based on e-commerce. Data capture, storage, analysis, search, sharing, migration, visualization, tracking, updating, data privacy, and data sources are significant data analysis challenges. Data capture, storage,

analysis, search, sharing, migration, visualization, tracking, updating, data privacy, and data sources are all big data analysis challenges. Using this data to make the program better is the central concept. Intelligent logistics [9] in real-time value networks enable customers to accurately navigate the entire life cycle of the transportation process, including purchases and contracts, shipment planning, execution and tracking, yard management, appointment schedules, and financial and claims settlement - all individually, cloud-based platform. A single connection to a real-time value network can connect to the thousands of companies already signed on, including partners and carriers. This is the network's worth. Other advantages include lower annual freight costs, better customer service, streamlined transportation processes, and reduced network inventory. In the work [10], the authors implemented big data analytics, intelligent logistics management, and machine learning to make the laundry service into an innovative IoT based e-commerce platform. They used Dijkstra's algorithm to find the shortest path with the least amount of traffic and the least amount of distance. They collected the data and then analyzed it using big data analytics. An idea was taken and tried to put into action in this work. This application also used machine learning techniques to test the sustainability of this idea by using people's thoughts about it as datasets.

The second paper was "Food panda: Changing the Way Bangladeshis Eat Meals" [11]. We selected this paper to understand the marketing better and promoting strategies and how to gather traffic and audiences so that our work "IronMan" can turn into a successful project. The following paper covers the part of the relationship between early business practices and companies' medium and long-term challenges and opportunities. It can give us an insight into our future challenges in building an e-commerce app, IronMan. The paper also discusses the food order process and delivery system. So we can use those ideas on this app to have a better delivery process so that this app can be sustainable and durable.

The third paper was "Customers' Use Intentions of Using Online Laundry Service" [12]. The journal depicts the customer psychology of choosing an online laundry service. The following study aims to understand customers' intentions to use online laundry services through one-on-one interviews as the qualitative research method approach. The findings revealed that customers' preferences to use online laundry services were primarily influenced by the effectiveness and efficiency of such services. It is simple to use and save time. As a result, an outcome of customers' behavior and attitude towards accepting this project was found.

The authors in [13] researched to develop Android-based laundry services that are more efficient in terms of time to process laundry pickup, records of incoming and outgoing garments, and information on their own laundry OXY. The method employed includes a literature review, interviews, questionnaires, design, and testing. The paper assisted in providing information such as price, type of service, and

preferred laundry branch. It also aided in the monitoring of applications and operations.

The fifth research paper is "Laundry Services Application" [14]. It shows a smart application that enhances the stipulation of a service that every home requires, mainly washing clothes and household items in specialized laundry centers, as their application facilitates linking the user to the nearest laundry center. It would be easy for them to communicate with each other and provide the laundry service online via an online request submitted by the user. To accomplish this, they created an Android application that works on smartphones connected to the Internet.

III. METHODOLOGY

This application has three types of mobile applications to help each user to complete those tasks efficiently, which are basically for shop owners, riders, and customers. The Android Studio is used as an android framework for development, and fire-base is used for the database framework. And the Google Cloud Platform to integrate Google Maps into the application.

A. Android Application

Shop Owner App: After logging into the system, the shop owners have the luxury too:

- add their products
- remove any product
- modify the products
- edit the price list

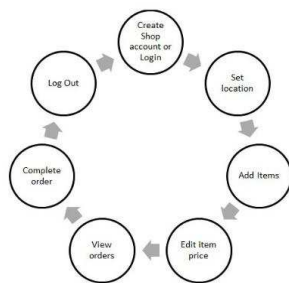


Fig. 1. Shop Owner App Diagram

Customer App: Customers must first create an account and log in to the system using their email or phone number. Customers can see a variety of shops and their names after logging into the system. Customers can edit their profiles by clicking my profile bar if they desire. They can then choose any payment method, such as card, mobile banking, or cash on delivery. Customers can view the order status after confirming their order.

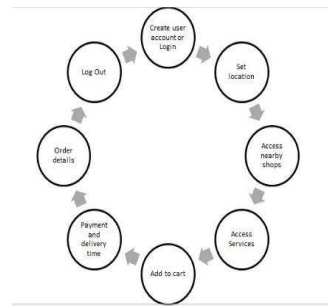


Fig. 2. Customer App Diagram

Rider App: When a customer orders, a notification will appear for the rider closest to that customer. The rider can then accept the pickup request. After picking up the products from the customer, the rider must deliver them to the shop chosen by the customer. After finishing the laundry, the shop owner can direct contact with the rider to provide the product to the

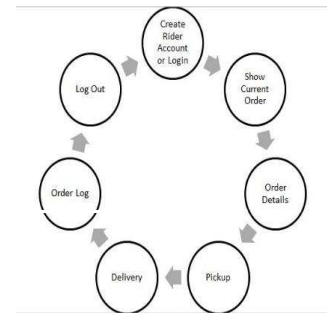


Fig. 3. Rider App Diagram

customer.

The details of the flow of different applications are given in Fig.1-Fig.3.

B. Web Based Application

The IronMan website will include a user panel with product showcasing, login and signup application, and a search of laundry shops nearby. On the other hand, there will be a Shop user panel where owners can decorate their shop by creating an account and adding all their services. These sections will have an administrator who will connect these users and enable the E-laundry service.

It all started with HTML and CSS as web development tools. Then React JS was used for the back-end operation. And Django as a framework. So that the HTML page of the website can easily create and stored the database in the Django administration, from there admin can control the whole website and do have access to read and write and modify anything. DialogFlow and Kommunicate is used to implement Chatbot.

For the machine learning model dataset, a survey was created with questions about people's opinions on the idea of an E-commerce-based laundry service [15]. Then the responses

were converted into CSV files and used as datasets. We convert all the categorical data into numeric data [16]. For Yes and No, 1 and 0 was set. Score set calculation is given in Fig.4

```

a=0
b=0
c=0
result = []
for score in finalScore:
    if score>8 :
        result.append('Sustainable')
    elif score>= 6 and score <= 8:
        result.append('Preferable')
    elif score<6:
        result.append('Not Sustainable')
    
```

Fig. 4. Score Set Code

The logistic regression model [17] was trained. 70 percent data were given into the train set and 30 percent data into the test set for training. The results from Logistic Regression with no regularization gives a test accuracy 86%. Fig.5 presents the confusion matrix.

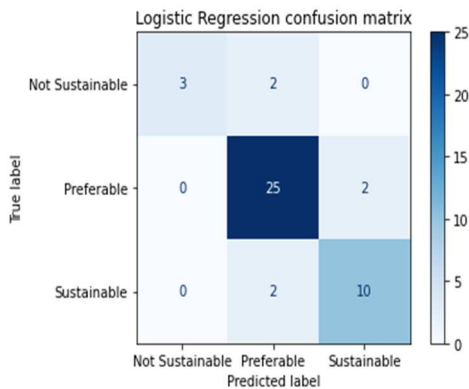


Fig. 5. Confusion Matrix

TensorFlow Lite [18] and a teachable machine were used to implement image classification. In TensorFlow Lite, some models were deployed to classify the images. It is one kind of machine learning process to object detection. In the models, we have deployed some images related to a shirt, pants, t-shirt, etc. Then we converted the model, optimized it, and deployed that model in the android studio project. The datasets of images and imported the model were trained as .tflite in android studio. DialogFlow [19] is used to understand natural language that makes it easy to create and integrate a conversational user interface into a mobile app, web app

IV. RESULTS

It was discovered that our current work almost performs flawlessly after individually analyzing our system. It successfully implemented the prototype of the entire system, as

planned initially. We divided into two parts. The final application demonstration will be used for the result, to analyze this project, and also to conduct some qualitative research.

A. Analysis of Results of Web Development

For the website, the user has to create a signup page at first. Then after registering, the customer can sign in to the system.

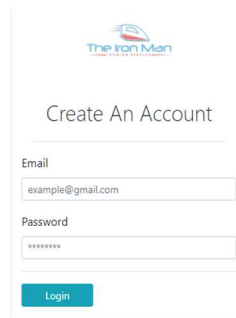


Fig. 6. Sign Up Page

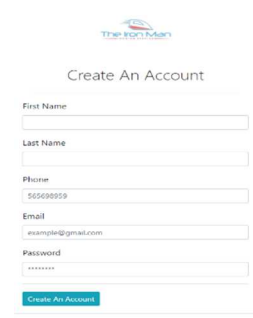


Fig. 7. Login Page

After logging into the website, customers can view the whole dashboard. There are different products from different custom laundry and iron shops.

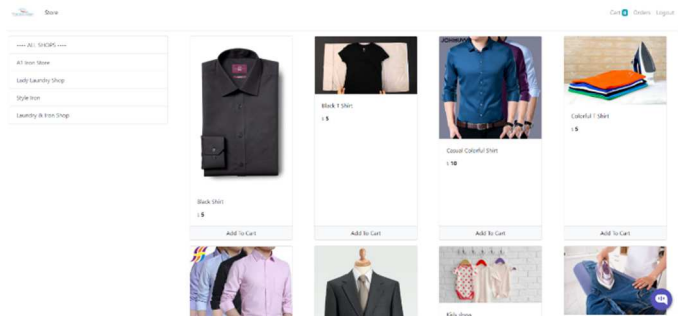


Fig. 8. Home Dashboard

Then the customer can select any products from any shops they like to order. They can increase or decrease the order. After ordering the products, they view them in the cart. The system will show the details like price, quantity, and total amount.

Sno.	Image	Product	Price	Quantity	Total
1		Black T-Shirt	₹ 5	2	₹ 10
2		Jeans	₹ 5	2	₹ 10
3		Shirt	₹ 15	2	₹ 30
Total					₹ 50

[Check out](#)

Fig. 9. Cart

After that, the customer can check out by giving the related information like their address and phone number.

Check Out Form

Address
Dhaka

Phone
01712981821

Check out

Fig. 10. Check Out Form

There is an order menu to view the orders or check the orders. From there, customers can view the products they ordered. Also, they can check the order status.

Your Orders

Sno.	Image	Product	Date	Price	Quantity	Total	Status
1		Black T Shirt	April 23, 2022	₳ 5	2	₳ 10	Pending
2		Jeans	April 23, 2022	₳ 5	2	₳ 10	Pending
3		Saree	April 23, 2022	₳ 15	2	₳ 30	Pending

Fig. 11. Order Status Pending

There is also a system where the admin can control the order status. If the order is done, the page will show the completed status.

Your Orders

Sno.	Image	Product	Date	Price	Quantity	Total	Status
1		Black T Shirt	April 23, 2022	₳ 5	2	₳ 10	Completed
2		Jeans	April 23, 2022	₳ 5	2	₳ 10	Completed
3		Saree	April 23, 2022	₳ 15	2	₳ 30	Completed

Fig. 12. Order Status Complete

All the steps from signup and order are described vividly through figures 6-12.

This application also has a Chabot where people can ask their queries (Fig.13).

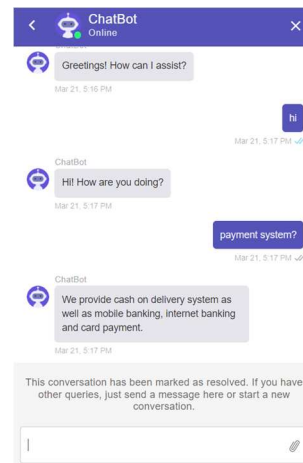


Fig. 13. Chatbot

B. Analysis of Result of Android Development

• Shop Owner App:

For Android development, the user has to create the shop owner app first. Any shop owner can register their shops by filling out the signing up page.

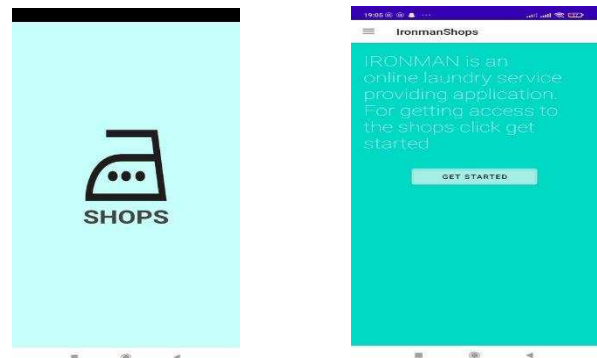


Fig. 14. Shop Owner App Interface

Shop owners can view their profiles as well as the orders. They can update the products and prices also. After updating, it will show the updated details on the product page. These are described in Fig14-15.

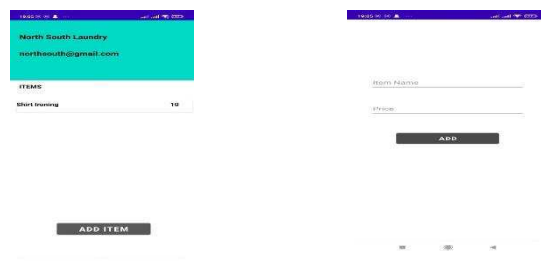


Fig. 15. Shop Owner Customizing Product

When the shop owner clicks on the 'Get Started' button, he can see the new page where he can update the products and their price. After filling those items, he must click on the 'Add' button. Then the product and its price will be updated.

• **Pricing of The Products:**

Initially, the shop owners can update the prices of the products according to their preferences. However, there has been made a template for updating the name of the products and prices. After registering in the app, shop owners have the authority to edit them. A survey was made asking different questions to different shop owners. They answered that they had set different prices for other products. They have asked us whether they could update the pricing as per their liking or not. This system assured them that that feature is available and flexible in the app.

• **Rider App:**

The next android application is for the rider. The rider will pick up the products from the customer's home, and after completing the ironing at the shops, he will deliver them to the customer. But firstly, the rider has to sign in to the app. Then he can see the orders that have come from the customers. After that, he can see the order details and prices through orders.

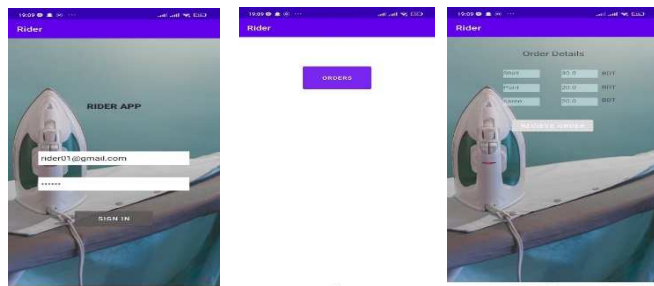


Fig. 16. Rider App Interfaces

The rider also has access to the image classification tool. He can check the clothes through image classification. It will assure the customer that their products will not be lost. Image classification will show the percentage of confidence whether that is the correct product.

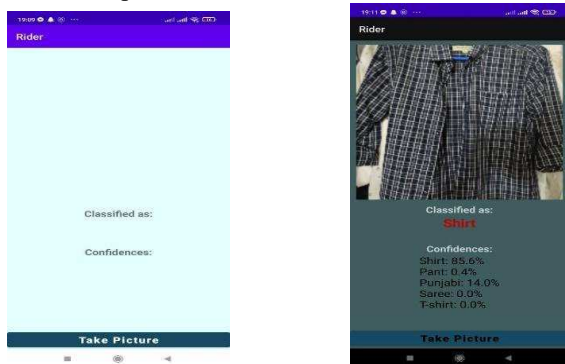


Fig. 17. Image Classification for Rider App

• **Customer App:**

There is also an app for the customer. Customers need to register through their mail, and then they can access their profile.

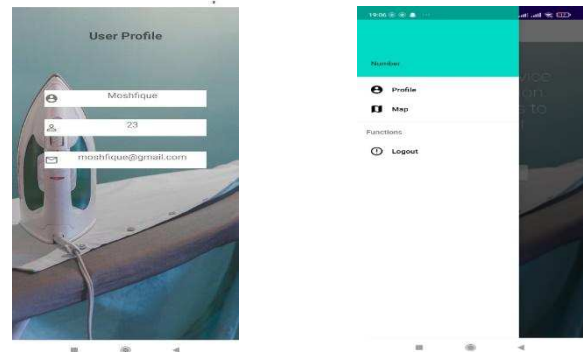


Fig. 18. Customer Profile

Customers can have access to Google Maps, where they will be able to see the current location. Also, they see different laundry shops that are registered through the app. Then they can choose any shop and order from that shop (Fig16-20).

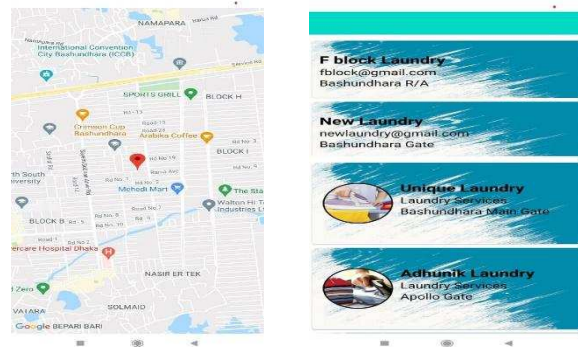


Fig. 19. Current Location and Registered Shops

After ordering from their particular liking shop, customers can view the cart and total amount of money.

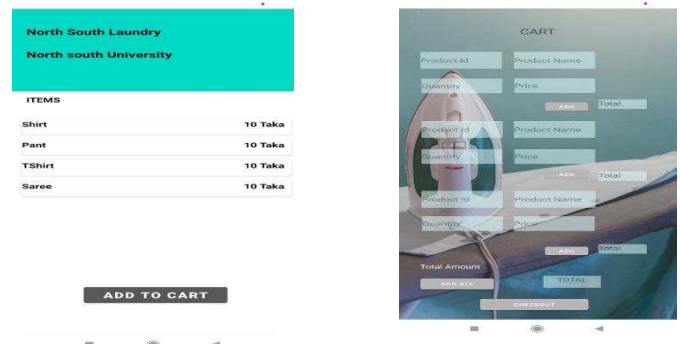


Fig. 20. Customer Cart

• **Image Detection:**

TensorFlow Lite is used for image classification. Our idea was to keep an image classifier in the rider app for instantly taking an image in real-time when picking order and adding the Image to the user's database for a future claim. The application has used machine learning to train the model by gathering images of categories like shirts, T-shirts, saree, Punjabi, etc. And then used TensorFlow Lite to add it to the Rider app and save the data to individual customer's databases using Firebase.

Image classification is an essential concern in which a set of target classes is defined, and a model is trained to identify them using labeled example photos. The system used numpy, Opencv, and PIL (Python Imaging Library) libraries for training.

Training the model requires the following steps[26]

1. First, it needs to provide the model with some training data.
2. Then the model tries to immaculate the provided images and labels.
3. After that, the model predicts the test sets.
4. Lastly, the model verifies whether the predictions match the labels or not.

The output of The Rider App is given in Fig.21.



Fig. 21. Image Classification

As shown in the above Fig. 21, after the rider takes a picture of items, the model can recognize the Shirt with 80.7% confidence, Punjabi with 100%, and Pant with 86.1%. The accuracy of the model will increase as more data gets added.

When the rider captures the images after taking the clothes from the customer, the photos will be saved to our database. When the rider is about to deliver the product back to the customer, he can check the product. The rider will click the images, and then the app will search the photos from the database. There will be the owner's name saved by their clothes images in the database. After that, the app will show the name of the owner of the clothes. Some snapshot of the image detection is given in Fig.22.

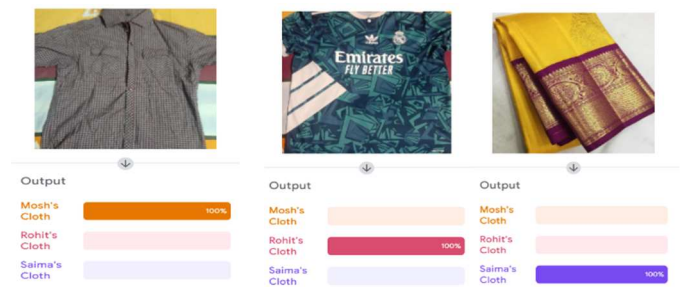


Fig. 22. Clothes Owner Image Detection

As shown in Fig. 22, the app will detect the owner's name through the database, then it will return the name of the owner matched by the Image of the clothes. After completing the ironing, the shop owner can check the owner of the clothes by using the feature of image detection from the database. Thus, every clothe will be scanned and will reach its actual owner, and there will not be any situation of mixing up any clothes.

V. QUALITATIVE SURVEY

The survey explains the view of the local laundry shop owners regarding the project 'IronMan.' We showed our demonstration of the application to them and asked them how they felt about our app. They were also asked about the sustainability of the project idea and if they are convenient and comfortable enough to use the mobile application.

After gathering the information and data from the different shop owners from other places, we realized that they were initially not sure or were not clear enough about the project process. They thought it would be an extra burden for them to appoint a delivery man, or some thought it would negatively impact their business. But when the method of the project was explained thoroughly, that is when they seemed clear about the project. So the majority of the shop owners are willing to participate in this project and appreciate the ideas. In short, the small shop owners see this as an opportunity to grow their business and set up a solid customer base. Whereas middle and large shop owners appreciate our project, they are currently not ready to be involved in our project. From the survey, an idea was found to appoint a go-between or delivery man, most of the shops are willing to be involved instantly. Overall, our current initiative seems sustainable to them. It is an excellent possibility that this project might be successful in the near future.

VI. CONCLUSION AND FUTURE WORK

Iron-Laundry Services is a service that is required regularly. People increasingly rely on e-commerce and internet-based jobs to help them manage their lives. This project aims to make laundry services more accessible to the general public. There has been developed e-commerce laundry services and brought all local laundry shops together on a single platform. In addition, the project contains Android and web-based applications. And to improve the user experience, it has been planned to add features related to big data analytics, intelligent

logistics management, and machine learning models. We have already taken people opinions on this project idea as a survey and turned them into datasets, which were used to train our logistic regression model and determine whether or not this work will be successful. To launch this as a pilot project, it needed to conduct qualitative research among local laundry shops to see if they are interested in such an idea or are ready to take their businesses online. This application needs to do fieldwork before launching as a pilot project. The main question, in the beginning, will be feasibility; will need to teach local shop owners how this project will work, its benefits and drawbacks, and so on. We hope that ecommerce based laundry can make a considerable change in the e-commerce industry.

REFERENCES

- [1] L. D. Xu, W. He and S. Li, "Internet of Things in Industries: A Survey," in *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233-2243, Nov. 2014.
- [2] S. Singh and N. Singh, "Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce," 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015, pp. 1577-1581.
- [3] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017, "Machine Learning Models that Remember Too Much. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)". Association for Computing Machinery, New York, NY, USA, 587–601.
- [4] Duncan McFarlane, Vaggelis Giannikas, Wenrong Lu, Intelligent logistics: Involving the customer, *Computers in Industry*, Volume 81, 2016, Pages 105-115, ISSN 0166-3615.
- [5] Gupta. Akanksha. et al, "Development of Mobile Application for Laundry Services Using Android Studio". Volume 13, Number 12 (2018) pp.10623-10626.
- [6] Sheba Platform Limited, "Sheba.xyz," 2020. [Online]. Available:Sheba.Xyz. [Accessed 16 February 2022].
- [7] Inc, Griho, "Laundry Bhai - Apps on Google Play". Available:Laundry Bhai. play.google.com, 8 Nov 2020.
- [8] Chang Liu, Yongfu Feng, Dongtao Lin, Liang Wu & Min Guo (2020) Iot based laundry services: an application of big data analytics, intelligent logistics management, and machine learning techniques, *International Journal of Production Research*, 58:17, 5113-5131.
- [9] "Transportation Management System TMS". Consumer-Driven Digital Supply Chain Management, www.onenetwork.com, 13 Oct 2020.
- [10] A Fitriansyah et al 2019 *J. Phys.: Conf. Ser.* 1338 012044
- [11] Akter, Mohinur & Disha, Nadia. (2021). Exploring Consumer Behavior for App-based Food Delivery in Bangladesh During COVID-19. *Bangladesh Journal of Integrated Thought*
- [12] Afifah, Hana, and Norlaile binti Salleh . "Customers' Use Intentions of Using Online Laundry Service." *International Journal of Business and Management*, 0 0 2018
- [13] Primawaty, Christine, and Sufa atin. "LAUNDRY SERVICE APPLICATION DEVELOPMENT ANDROID BASED."
- [14] Saati, Ibrahim. (2020). Laundry Services Application
- [15] "https://forms.gle/98DETKx5yLg8k8xc8".
- [16] Al Aghbari, Zaher, "Classification of Categorical and Numerical Data on Selected Subset of Features" .2010.
- [17] "Logistic Regression | Building an End-to-End Logistic Regression Model."
- [18] "Image Classification | TensorFlow Core."
- [19] Pykes, Kurtis, "How To Create A Conversational Agent with Dialogflow by Kurtis Pykes Towards Data Science". 5 May 2021. Medium, towardsdatascience.com.

A Brief Overview on Security Challenges and Protocols in Internet of Things Application

Gajjala Savithri^{1,2}, Bhabendu Kumar Mohanta³, Mohan Kumar Dehury⁴

¹Department of Animation, Dr. YSR Architecture and Fine Arts University, Kadapa, AP, India

²Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

³Department of CSE, GITAM University, Visakhapatnam, Andhra Pradesh, India.

⁴Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

Email: savithrigreddy@gmail.com¹, bhabendukumar@gmail.com³, mohankdehury@gmail.com⁴

Abstract—Internet of Things(IoT) is one of the emerging technology which connect billion of smart devices with minimum human interaction. Over the last decade, the Internet of Things has been increasingly used in applications like smart city, smart home, smart transportation, intelligent healthcare system, agriculture and smart grid. Traditional applications are becoming smarter and more automated as a result of the Internet of Things and some other emerging technologies. Due to resource constraint IoT devices are more vulnerable to the different attacks. To begin, this article identifies security problems in terms of design, data storage, and computation. Several researchers have already made significant contributions to this field of study. The authors of this study analyse the security challenges and available solutions to these challenges over the last five years. Finally, this article addressed the research problems and future technologies that would need to be integrated in order to meet the complex security risks associated with IoT applications.

Index Terms—Internet of Things, Security, Protocols, Privacy, Blockchain.

I. INTRODUCTION

Internet of Things (IoT) was first invented by Kevin Ashton, the Executive Director of Auto-ID Labs at MIT, in the year 1999 [1]. During his presentation for Procter Gamble supply chain system, he predicted that sooner than later computers or things will be able to collect information without human intervention. The Internet of Things (IoT) has emerged as one of the most promising research disciplines in the recent decade. Because of the smartness and real-time monitoring without human intervention, the IoT services have come as a huge relief to human life. IoT has a lot of applications such as smart home, smart transportation, intelligent lighting system, smart agriculture, supply chain system, smart metering and smart grids, smart healthcare system, industrial automation, smart retail: etc. IoT applications are frequently employed because they offer a high level of comfort, system automation, and efficiency. The large number of IoT devices and users are connected to the network which generates huge volumes of data. But it is also prone to basic security issues such as confidentiality, integrity and availability. To make successful

use of the ever-growing IoT applications, security, privacy, and trust must be addressed in order to secure IoT devices and user privacy from attackers.

Eight key research topics are identified by authors in [2], these are massive scaling, architecture and dependencies, creating knowledge and big data, robustness, openness, security, privacy, and human-in-the-loop for the future IoT. Similarly, security features of IoT are confidentiality, integrity, availability, identification authentication, privacy and trust are studied in [3], which are need to be addressed. When user information is shared over untrusted network, trust among users, and data privacy, as well as data confidentiality, are important factors. The traditional security mechanisms are not suitable for IoT environments. The number of devices connected are increased exponentially which rises to scalability issues. In IoT applications, devices are accessed remotely. The protocols used to connect smart devices are ZigBee, Bluetooth, 6LoWPAN, Message Queuing Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), Extensible Messaging and Presence Protocol (XMPP), Near Field Communication (NFC), Routing protocol (RPL) Sigfox, LoRaWAN, or WiFi [9]. Each of these protocol is vulnerable to different types of attacks. The attacks are focuses on different type in IoT protocols like attacks on packets, attacks on the protocol and attacks on the whole system [10]. As most of the IoT architecture smart devices are connected through wireless manner, there is high risk of security vulnerabilities. The different attacks on wireless protocols are explained in [11] by the authors. The IoT is making the traditional applications into smart applications using the smart devices where processing and computation is done real time using fog computing or edge computing. Some of the applications like healthcare system, home automation system, smart city where sensitive information are share among different users if the attackers success to do any attacks then whole system become failed [12]. The IoT devices are vulnerable to various types of attacks due to lack of security protocols implementation [13]. Security

TABLE I
RECENT RELATED SURVEY ON IOT SECURITY

References	Year	Focus Point	Contribution
Y.Yang.et al. [4],	2017	Security and Privacy	The authors concentrate on the limitations of IoT devices and various attacks model. The authentication and access control mechanism of IoT applications are explained in details. The different layers of security vulnerabilities are also identified.
I. Stellios.et al. [5]	2018	Attacks on critical IoT infrastructure	Authors in this survey identify IoT-enabled cyber threats , verified IoT-enabled attacks in IoT applications. The solution approach of different direct and indirect attacks against important tagrets smart devices is explained. Finally, the authors discuss numerous problems associated with essential infrastructure and services in terms of IoT applications.
F.Meneghello.et al. [6]	2019	Security Vulnerabilities	The purpose of this article is to provide a high-level overview of the security dangers associated with the Internet of Things and to examine some potential countermeasures. To this goal, authors cover the specific security procedures used by the most prevalent IoT communication protocols after a fundamental introduction to security in the IoT area. Authors made comparison of the IoT technologies under consideration in terms of a set of qualifying security qualities, including integrity, anonymity, confidentiality, privacy, access control, authentication, authorization, resilience, and self-organization.
I. Butun.et al. [7]	2020	Vulnerabilities, Attacks, and Countermeasures	This paper provides a comprehensive overview of security attacks against WSNs and IoT, as well as approaches for preventing, detecting, and mitigating those attacks. In this article, attacks are classified and studied primarily in two categories: "Passive Attacks" and "Active Attacks." Understanding these threats and the security mechanisms connected with them will assist in constructing a secure road toward the spread and public adoption of IoT technology.
Euijong Lee.et al. [8]	2021	Interoperability and Security	Authors in this article work on interoperability and security in IoT applications. Additionally, authors looked at international standards bodies that have been producing standards for the Internet of Things in term of security and protocols used in IoT applications.

challenge like malicious behavior of smart device detection [14] or identifying the unrecognized information send from untrust devices need to be address. To make IoT applications secure and trustful where users integrity and privacy are taken care. So security challenges are need to be address properly. This paper’s summary is as follows:

Section II provides context for the study process. Section III discussed the security risks associated with IoT applications and the techniques employed to address them. The section IV highlighted and discussed the research that has already been conducted to address security concerns through the use of Blockchain Technology. The paper conclusion and future scope of research mentioned in Section V.

II. BACKGROUND STUDY OF RESEARCH METHOD

The initial study of the research method is shown in figure 1. Four major research database name IEEE Xplore, ScienceDirect, Springer and ACM digital are consider to collect the articles.

The relevant query was used to look for publications published between 2017 and 2021, which corresponds to the last five years’ databases. Along with the specified keywords, some filtering techniques are used to ensure that only English language journals with applications in computer science are evaluated. The total number of documents first downloaded is

2130. Following that, screening processes such as removing duplicate articles, deleting articles after reading the abstract and title, and finally reviewing the complete text are used to choose 60 articles for the survey.

III. SECURITY RISK ASSOCIATES IN IOT APPLICATION

Since the inception of IoT lots of applications are develop to monitor in real-time. The technology like machine learning, artificial intelligent and blockchain use to implement IoT applications [15]. As the number of smart devices are develop and deploy in numerous applications. These smart devices generate huge volume of data to be process, compute and store.

The layerwise attacks on IoT are shown in figure 2. Each of three layer in IoT have some attacks associated with it. The requirement to enhance the security of IoT systems is increasing. Users may choose not to use such systems if they become aware that adequate security measures are not in place [16].Recent advances in IoT security research have been facilitated by the availability of simulation tools, IoT modellers, and analysis platforms [17]. The security attacks, privacy issue, secure architecture design challenges, and protocols for secure communication are some of the important challenges exist in IoT applications are mentioned in paper [18]–[21]. The enormous privacy and security attacks associated with Internet-of-Things (IoT) devices, it appears desirable to encourage

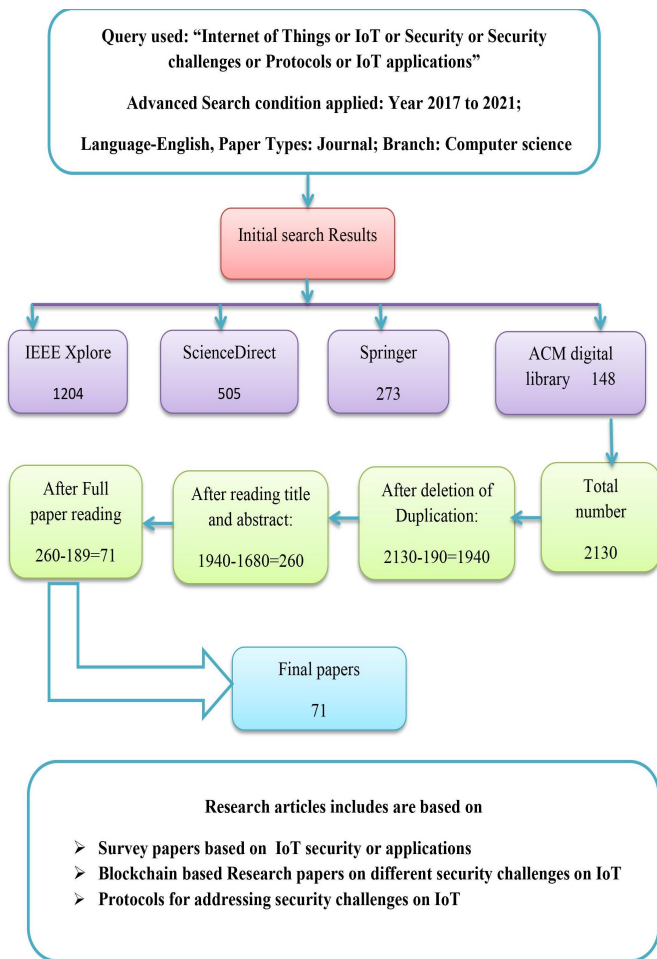


Fig. 1. Search criteria and final paper selection process

consumers to select more secure equipment and to factor privacy into their purchasing decisions [22]. The end user must build the trust to share his/her personal information to access different service provided by the IoT applications. In smart city applications there are many services which are available for the citizen. The citizen must have trust to share and access the service provided by the authority. Other wise the smart city ambition will not be full fill. To make the any applications usable to every user proper security measure must be build which will make the system reliable, secure and user friendly. When user information is shared over untrusted network, trust among users, and data privacy, as well as data confidentiality, are important factors. The traditional security mechanisms are not suitable for the IoT environment. The number of devices connected are increased significantly which rise to scalability issue, therefore lightweight, efficient and flexible protocols are needed to deal with the security attacks in IoT applications.

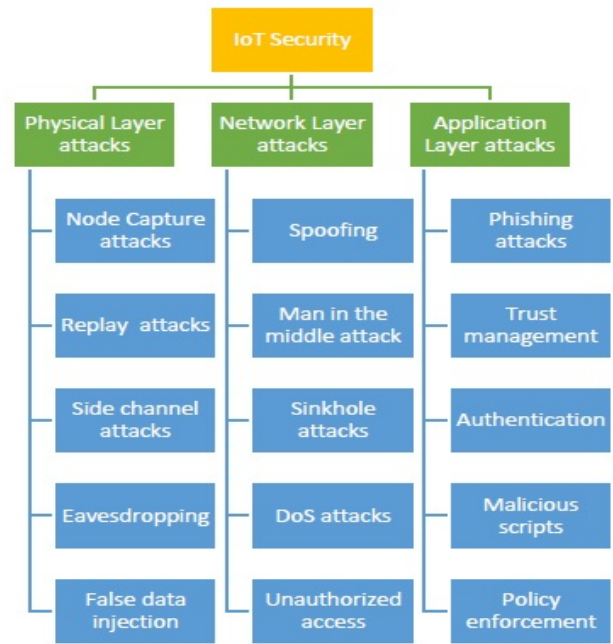


Fig. 2. Different layerwise security attacks in IoT [23]

Secure protocols in IoT applications is essential for rapid IoT adoption by the society. The security challenges in IoT may be active or passive type. Some of the security attacks and existing solution approach by different researcher are explained in Table II.

IV. BLOCKCHAIN FOR IoT SECURITY

In previous section that is section III, we have identified various potential risk and security attacks in IoT system. The emerging technology like Artificial intelligence (AI), Machine Learning(ML) and Blockchain are used to address different security and privacy challenges in IoT applications [15]. Blockchain technology consider as one of the secure platform because of it has some inherent security properties like immutable, distributed architecture. By utilising a peer-to-peer network, a Blockchain is totally decentralised. To be more exact, each node in the network keeps a copy of the ledger in order to avoid a single point of failure. All copies are simultaneously updated and confirmed. Blockchains can be private (permissioned) or public (non-permissioned) (public). The first category imposes limits on the contributors to the consensus. Only selected trustworthy actors are granted the authority to validate transactions. It does not involve much deliberation to reach a consensus, and thus is neither time- nor energy-consuming. Finally, it ensures the privacy of transactions by restricting access to them to authorised parties. The second type (public blockchains) makes use of an infinite amount of

TABLE II
SECURITY ATTACKS AND CORRESPONDING SOLUTION APPROACH

Attack Name	Reference	Solution approach/Methods Used
DOS	[24]–[27]	The DoS attacks is one of the significant security issues in IoT applications. The availability of services is one of primary aspect of intelligent gadgets as they monitor the items in real time. In IoT applications need dependable and effective data transmission with data delivery to the desire node. The numerous lightweight authentication communication protocols are developed to solve the misbehaviour of users’ or identify any anomaly in end to end communication.
MITM	[28]	Through various intermediary devices such as switches and controllers, smart devices are connected and deployed in diverse applications. Additionally, the majority of intelligent gadgets are connected via wireless communication protocols. If any of the network’s devices are compromised or an attacker obtains knowledge via an insecure channel. The Bloom lightweight countermeasure using Bloom filters protocol that is used to monitor the modification of packets in order to detect MITM attacks.
Trust Management	[29]–[31]	The heterogeneous nature of smart devices is a significant challenge for IoT design. Due to the variety of attacks on smart devices and the vulnerability of the majority of services to various types of attacks, trust management is a difficult issue for users’. Detection and prevention of malicious threats through the use of secure protocols builds user group confidence. The Blockchain-based secure platform addresses the issue of trust management in Internet of Things applications.
Malicious attacks	[32]–[39]	Malicious code is unrecognized software to gain the credential or information which will damage the device. IoT devices being the resource constraint are more vulnerable to the malicious attacks. The malicious node detection using different methods is essential to make the system security proof. The conventional security solutions are not enough to address all challenges like when detection of abnormality in the network and alert to the security administrator for handle the challenge are not suitable for IoT devices. Recently, researchers presented many schemes to handle malicious node detection in IoT applications, including index-based data provenance, a multi-agent and multi-layered game protocol, and a privacy-preserving game-based strategy.
routing attacks	[40]–[42]	Communication between smart devices might be decentralised, centralised, or dynamic in nature, depending on the routing protocols used in IoT applications. Security at the network level is a challenge in IoT networks due to the higher energy consumption of standard cryptography. Any IoT application is said to be using dense architecture since it makes use of a large number of smart devices to collect data and monitor the environment. The researcher addresses routing attacks by presenting a lightweight model based on machine learning algorithms.

anonymous nodes. Each actor can communicate safely using cryptography. Each node is identified by a pair of public/private keys. Any actor on the blockchain can read, write, and validate transactions. The blockchain is secure, and network consensus is achieved when at least 51% of nodes are honest. Typically, permissionless blockchains are energy and time intensive, as they require a certain amount of computing to ensure the system’s security. Blocks are used to group and store transactions based on their timestamps. These blocks are then connected (or chained) to form a chain of blocks, or blockchain. The blockchain employs elliptic curve cryptography (ECC) and the SHA-2 hashing algorithm to ensure the authenticity and integrity of data. The block data is fundamentally a list of all transactions and a hash to the preceding block. The blockchain maintains a complete record of all transactions and enables global distributed trust across borders. Trusted Third Parties (TTPs) or centralised authority and services can be infiltrated, hacked, or interrupted. They may also behave inappropriately and become corrupt in the future, even though they are currently trustworthy. Each transaction on the distributed public ledger is confirmed in blockchain by a majority consensus of miner nodes that are actively engaging in checking and validating transactions. Ethereum’s blockchain technology enables the storage of data and, more significantly, the execution of smart contracts. Nick Szabo originated the term smart contracts in 1994. A smart contract is, in essence, a computerised transaction protocol that carries out the contract’s terms. In a nutshell, smart contracts are user-created programmes that are uploaded and executed on the blockchain. Solidity is the scripting or programming language for smart contracts. It is similar to JavaScript. The Ethereum Blockchain enables the use of EVMs (Ethereum Virtual Machines), which are essentially miner nodes. These nodes are capable of enabling trustworthy execution and enforcement of these programmes or contracts in a cryptographically tamper-proof manner .

Design/architecture : Through protocols and gateways, the Internet of Things (IoT) architecture layers are utilised to track a system’s consistency. Due to the heterogeneity of IoT devices deployed in a variety of locations, managing connection, mobility, and interoperability within devices is challenging. As a result, certain attacks such as denial of service (DoS) attacks and malware attacks are a cause of worry. The authors of the publications identified the design issues and also discussed solution strategy [43]–[47].

Lightweight Protocols: The most of the IoT devices are resource constraint devices, so existing security protocols are not suitable to implement in IoT application. The lightweight computation algorithms are need to be develop which will be suitable for the smart devices. some of the research work already done on this regard [48]–[52].

Data Privacy: The data collected from different smart

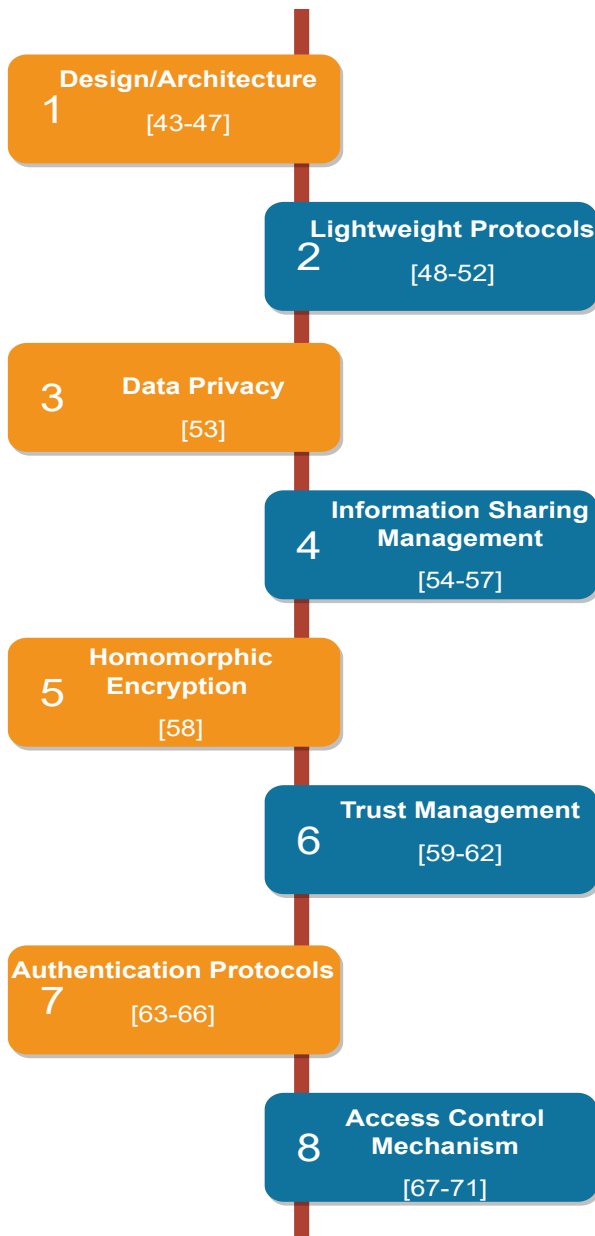


Fig. 3. Blockchain based solutions for various security issues of IoT.

devices are store in a centralized system for further processing in a IoT application. The end user access these data through various services to gain knowledge where user share lot of personal information to the network.If single point of failure occur or central server get compromise then privacy is a big concern as it is challenging to validate to IoT devices in untrusted IoT Network [53].

Information Sharing Management: Information security

management (ISM) establishes and manages the policies that an organisation must put in place to ensure that the confidentiality, integrity, and availability of information are protected from threats and vulnerabilities in a sensible manner.The system uses a double-chain approach that combines data and transaction blockchains, with distributed storage and tamper-proof data in the data blockchain and an upgraded practical Byzantine fault-tolerant (PBFT) mechanism consensus algorithm in the transaction blockchain. Through improved algorithms based on partial blind signature algorithms, data registration efficiency; resource and data transactions in the transaction blockchain are improved transaction efficiency and privacy protection [54]–[57].

Homomorphic Encryption: Homomorphic encryption is a type of encryption that allows you to do numerous operations on data while it's encrypted without compromising its security [58].A blockchain-based data aggregation framework and a cloud computing-based data aggregation framework are intended to improve security. A blockchain is used to record the transaction in the cloud. The cloud is also utilised to offload computations for smart metres with limited processing capability.

Trust Management: Tamper-proof data, more reliable trust information integrity verification, and improved privacy and availability during sharing and storage are all advantages of blockchain-based trust management [59]–[62]. The distributed network store record securely using cryptography protocols. Authentication, data encryption, digital signature and hashing ensure security. Once the security and privacy are ensure, end user build the trust to use different services in IoT application.

Authentication Protocols: Identity of users and smart devices in IoT applications is important aspect. Authentication is the process by which restriction can be put into the unauthorized users. The various cryptography authentication protocols are proposed by the researcher which are suitable for the lightweight devices. Blockchain based authentication protocols using one way hashing, multi server authentication, mutual authentication and smart contract in decentralized environment for smart devices [63]–[66].

Access Control Mechanism: It is a process designed to identify and prevent unwanted access while allowing authorised access to an information system or physical location. The attribute based encryption(ABE), Blockchain enable control, three factor based user access control scheme and ABE integration with Blockchain technique are different methods proposed by the researchers to make the system secure [67]–[71].

Some of the research challenges are still need to be address to make the IoT applications more usability. Scalability, localization of smart devices, deployment, smart device management, malware detection, and encryption of collected

information from smart sensors are various challenges need to be address to make the system security proof.

V. CONCLUSION

The authors of this paper begin mentioning the growth of IoT devices and development of various applications. The security risk associated with IoT layer wise are identified. Some of the attacks like DOS, MITM, Trust management, Malicious and Routing attacks and corresponding solution are identified. According to the finding, considerable research has already been conducted to solve the security challenges associated with Blockchain technology. Eight different challenges of IoT Security corresponding solution address by Blockchain technology are categorically explained. Lastly paper conclude with critical issue that must be addressed.

REFERENCES

[1] K. Ashton *et al.*, "That 'internet of things' thing," *RFID journal*, vol. 22, no. 7, pp. 97–114, 2009.

[2] J. A. Stankovic, "Research directions for the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, 2014.

[3] I.-C. Lin and T.-C. Liao, "A survey of blockchain security issues and challenges," *Int. J. Netw. Secur.*, vol. 19, no. 5, pp. 653–659, 2017.

[4] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in internet-of-things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, 2017.

[5] I. Stellos, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3453–3495, 2018.

[6] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, "Iot: Internet of threats? a survey of practical security vulnerabilities in real iot devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8182–8201, 2019.

[7] I. Butun, P. Österberg, and H. Song, "Security of the internet of things: Vulnerabilities, attacks, and countermeasures," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 616–644, 2020.

[8] E. Lee, Y.-D. Seo, S.-R. Oh, and Y.-G. Kim, "A survey on standards for interoperability and security in the internet of things," *IEEE Communications Surveys Tutorials*, vol. 23, no. 2, pp. 1020–1047, 2021.

[9] R. Yugha and S. Chithra, "A survey on technologies and security protocols: Reference for future generation iot," *Journal of Network and Computer Applications*, p. 102763, 2020.

[10] J. Tournier, F. Lesueur, F. Le Mouël, L. Guyon, and H. Ben-Hassine, "A survey of iot protocols and their security issues through the lens of a generic iot stack," *Internet of Things*, vol. 16, p. 100264, 2021.

[11] I. Tomić and J. A. McCann, "A survey of potential security issues in existing wireless sensor network protocols," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1910–1923, 2017.

[12] W. Iqbal, H. Abbas, M. Daneshmand, B. Rauf, and Y. A. Bangash, "An in-depth analysis of iot security requirements, challenges, and their countermeasures via software-defined security," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10250–10276, 2020.

[13] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, and N. Ghani, "Demystifying iot security: An exhaustive survey on iot vulnerabilities and a first empirical look on internet-scale iot exploitations," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2702–2733, 2019.

[14] J. Wang, S. Hao, R. Wen, B. Zhang, L. Zhang, H. Hu, and R. Lu, "Iot-praetor: Undesired behaviors detection for iot devices," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 927–940, 2021.

[15] B. K. Mohanta, D. Jena, U. Satapathy, and S. Patnaik, "Survey on iot security: Challenges and solution using machine learning, artificial intelligence and blockchain technology," *Internet of Things*, vol. 11, p. 100227, 2020.

[16] F. A. A. Lins and M. Vieira, "Security requirements and solutions for iot gateways: A comprehensive study," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8667–8679, 2021.

[17] W. H. Hassan *et al.*, "Current research on internet of things (iot) security: A survey," *Computer networks*, vol. 148, pp. 283–294, 2019.

[18] V. Sharma, I. You, K. Andersson, F. Palmieri, M. H. Rehmani, and J. Lim, "Security, privacy and trust for smart mobile- internet of things (m-iot): A survey," *IEEE Access*, vol. 8, pp. 167123–167163, 2020.

[19] B. K. Mohanta, U. Satapathy, S. S. Panda, and D. Jena, "A novel approach to solve security and privacy issues for iot applications using blockchain," in *2019 International Conference on Information Technology (ICIT)*, 2019, pp. 394–399.

[20] S. Patel, A. Sahoo, B. K. Mohanta, S. S. Panda, and D. Jena, "Dauth: A decentralized web authentication system using ethereum based blockchain," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1–5.

[21] C.-S. Park and H.-M. Nam, "Security architecture and protocols for secure mqtt-sn," *IEEE Access*, vol. 8, pp. 226422–226436, 2020.

[22] N. Ho-Sam-Sooi, W. Pieters, and M. Kroesen, "Investigating the effect of security and privacy on iot device purchase behaviour," *computers & security*, vol. 102, p. 102132, 2021.

[23] B. K. Mohanta, D. Jena, S. Ramasubbarreddy, M. Daneshmand, and A. H. Gandomi, "Addressing security and privacy issues of iot using blockchain technology," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 881–888, 2021.

[24] M. Ghahramani, R. Javidan, M. Shojafar, R. Taheri, M. Alazab, and R. Tafazolli, "Rss: An energy-efficient approach for securing iot service protocols against the dos attack," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3619–3635, 2021.

[25] M. Stute, P. Agarwal, A. Kumar, A. Asadi, and M. Hollick, "Lidor: A lightweight dos-resilient communication protocol for safety-critical iot systems," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6802–6816, 2020.

[26] C. Lyu, X. Zhang, Z. Liu, and C.-H. Chi, "Selective authentication based geographic opportunistic routing in wireless sensor networks for internet of things against dos attacks," *IEEE Access*, vol. 7, pp. 31068–31082, 2019.

[27] C. Pu, "Sybil attack in rpl-based internet of things: Analysis and defenses," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4937–4949, 2020.

[28] C. Li, Z. Qin, E. Novak, and Q. Li, "Securing sdn infrastructure of iot-fog networks from mitm attacks," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1156–1164, 2017.

[29] M. D. Alshehri and F. K. Hussain, "A fuzzy security protocol for trust management in the internet of things (fuzzy-iot)," *Computing*, vol. 101, no. 7, pp. 791–818, 2019.

[30] S. T. Mehedi, A. A. M. Shamim, and M. B. A. Miah, "Blockchain-based security management of iot infrastructure with ethereum transactions," *Iran Journal of Computer Science*, vol. 2, no. 3, pp. 189–195, 2019.

[31] W. She, Q. Liu, Z. Tian, J.-S. Chen, B. Wang, and W. Liu, "Blockchain trust model for malicious node detection in wireless sensor networks," *IEEE Access*, vol. 7, pp. 38947–38956, 2019.

[32] M. Hossain and J. Xie, "Third eye: Context-aware detection for hidden terminal emulation attacks in cognitive radio-enabled iot networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 214–228, 2020.

[33] S. Shen, L. Huang, H. Zhou, S. Yu, E. Fan, and Q. Cao, "Multistage signaling game-based optimal detection strategies for suppressing malware diffusion in fog-cloud-based iot networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1043–1054, 2018.

[34] M. R. Dey, U. Satapathy, P. Bhanshe, B. K. Mohanta, and D. Jena, "Magtrack: Detecting road surface condition using smartphone sensors

- and machine learning,” in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 2485–2489.
- [35] Z. Liu and Y. Wu, “An index-based provenance compression scheme for identifying malicious nodes in multihop iot network,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4061–4071, 2020.
- [36] A. Agiollo, M. Conti, P. Kaliyar, T.-N. Lin, and L. Pajola, “Detonar: Detection of routing attacks in rpl-based iot,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1178–1190, 2021.
- [37] Z. Hassan, A. Mehmood, C. Maple, M. A. Khan, and A. Aldegheshem, “Intelligent detection of black hole attacks for secure communication in autonomous and connected vehicles,” *IEEE Access*, vol. 8, pp. 199 618–199 628, 2020.
- [38] C. Marche and M. Nitti, “Trust-related attacks and their detection: A trust management model for the social iot,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 3297–3308, 2021.
- [39] A. Azmoodeh, A. Dehghantaha, and K.-K. R. Choo, “Robust malware detection for internet of (battlefield) things devices using deep eigenspace learning,” *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 88–95, 2019.
- [40] K. Prathapchandran and T. Janani, “A trust aware security mechanism to detect sinkhole attack in rpl-based iot environment using random forest-trust,” *Computer Networks*, vol. 198, p. 108413, 2021.
- [41] N. Pavlović, M. Šarac, S. Adamović, M. Saračević, K. Ahmad, N. Maček, and D. K. Sharma, “An approach to adding simple interface as security gateway architecture for iot device,” *Multimedia Tools and Applications*, pp. 1–16, 2021.
- [42] B. Maram, J. Gnanasekar, G. Manogaran, and M. Balaanand, “Intelligent security algorithm for unicode data privacy and security in iot,” *Service Oriented Computing and Applications*, vol. 13, no. 1, pp. 3–15, 2019.
- [43] B. K. Mohanta, D. Jena, S. S. Panda, and S. Sobhanayak, “Blockchain technology: A survey on applications and security privacy challenges,” *Internet of Things*, vol. 8, p. 100107, 2019.
- [44] K. Seyhan, T. N. Nguyen, S. Akleylek, K. Cengiz, and S. H. Islam, “Bi-gis ke: Modified key exchange protocol with reusable keys for iot security,” *Journal of Information Security and Applications*, vol. 58, p. 102788, 2021.
- [45] S. Brotsis, K. Limniotis, G. Bendiab, N. Kolokotronis, and S. Shiaeles, “On the suitability of blockchain platforms for iot applications: Architectures, security, privacy, and performance,” *Computer Networks*, vol. 191, p. 108005, 2021.
- [46] R. Kamal, E. E.-D. Hemdan, and N. El-Fishway, “A review study on blockchain-based iot security and forensics,” *Multimedia Tools and Applications*, pp. 1–32, 2021.
- [47] A. A. Alfa, J. K. Alhassan, O. M. Olaniyi, and M. Olalere, “Blockchain technology in iot systems: current trends, methodology, problems, applications, and future directions,” *Journal of Reliable Intelligent Environments*, vol. 7, no. 2, pp. 115–143, 2021.
- [48] A. Bakshi, L. Chen, K. Srinivasan, C. E. Koksai, and A. Eryilmaz, “Emit: An efficient mac paradigm for the internet of things,” *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1572–1583, 2019.
- [49] L. Zhou, C. Su, and K.-H. Yeh, “A lightweight cryptographic protocol with certificateless signature for the internet of things,” *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 18, no. 3, pp. 1–10, 2019.
- [50] A. Cherif, M. Belkadi, and D. Sauveron, “A lightweight and secure data collection serverless protocol demonstrated in an active rfids scenario,” *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 18, no. 3, pp. 1–27, 2019.
- [51] S. N. Mohanty, K. Ramya, S. S. Rani, D. Gupta, K. Shankar, S. Lakshmanaprabu, and A. Khanna, “An efficient lightweight integrated blockchain (elib) model for iot security and privacy,” *Future Generation Computer Systems*, vol. 102, pp. 1027–1037, 2020.
- [52] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, “Lsb: A lightweight scalable blockchain for iot security and anonymity,” *Journal of Parallel and Distributed Computing*, vol. 134, pp. 180–197, 2019.
- [53] F. Loukil, C. Ghedira-Guegan, K. Boukadi, A.-N. Benharkat, and E. Benkhelifa, “Data privacy based on iot device behavior control using blockchain,” *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 1, pp. 1–20, 2021.
- [54] H. Si, C. Sun, Y. Li, H. Qiao, and L. Shi, “Iot information sharing security mechanism based on blockchain technology,” *Future Generation Computer Systems*, vol. 101, pp. 1028–1040, 2019.
- [55] A. Lohachab and B. Karambir, “Critical analysis of ddos—an emerging security threat over iot networks,” *Journal of Communications and Information Networks*, vol. 3, no. 3, pp. 57–78, 2018.
- [56] H. Wang, D. He, J. Yu, N. N. Xiong, and B. Wu, “Rdic: A blockchain-based remote data integrity checking scheme for iot in 5g networks,” *Journal of Parallel and Distributed Computing*, vol. 152, pp. 1–10, 2021.
- [57] N. Miloslavskaya and A. Tolstoy, “Iotblocksiem for information security incident management in the internet of things ecosystem,” *Cluster Computing*, vol. 23, no. 3, pp. 1911–1925, 2020.
- [58] S. Gupta, R. Garg, N. Gupta, W. S. Alnumay, U. Ghosh, and P. K. Sharma, “Energy-efficient dynamic homomorphic security scheme for fog computing in iot networks,” *Journal of Information Security and Applications*, vol. 58, p. 102768, 2021.
- [59] S. Hameed, S. A. Shah, Q. S. Saeed, S. Siddiqui, I. Ali, A. Vedeshin, and D. Draheim, “A scalable key and trust management solution for iot sensors using sdn and blockchain technology,” *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8716–8733, 2021.
- [60] B. Shala, U. Trick, A. Lehmann, B. Ghita, and S. Shiaeles, “Blockchain and trust for secure, end-user-based and decentralized iot service provision,” *IEEE Access*, vol. 8, pp. 119 961–119 979, 2020.
- [61] E. M. Abou-Nassar, A. M. Iiyasu, P. M. El-Kafrawy, O.-Y. Song, A. K. Bashir, and A. A. A. El-Latif, “Ditrust chain: Towards blockchain-based trust models for sustainable healthcare iot systems,” *IEEE Access*, vol. 8, pp. 111 223–111 238, 2020.
- [62] Y. Zhang, X. Xu, A. Liu, Q. Lu, L. Xu, and F. Tao, “Blockchain-based trust mechanism for iot-based smart manufacturing system,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1386–1394, 2019.
- [63] Y. Zhang, B. Li, B. Liu, Y. Hu, and H. Zheng, “A privacy-aware pufs-based multiserver authentication protocol in cloud-edge iot systems using blockchain,” *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13 958–13 974, 2021.
- [64] S. S. Panda, D. Jena, B. K. Mohanta, S. Ramasubbareddy, M. Daneshmand, and A. H. Gandomi, “Authentication and key management in distributed iot using blockchain technology,” *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 947–12 954, 2021.
- [65] A. Vangala, A. K. Sutrala, A. K. Das, and M. Jo, “Smart contract-based blockchain-envisioned authentication scheme for smart farming,” *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 792–10 806, 2021.
- [66] G. Ali, N. Ahmad, Y. Cao, S. Khan, H. Cruickshank, E. A. Qazi, and A. Ali, “xldbauth: Blockchain based cross domain authentication and authorization framework for internet of things,” *IEEE Access*, vol. 8, pp. 58 800–58 816, 2020.
- [67] J. Zhang, Y. Xin, Y. Gao, X. Lei, and Y. Yang, “Secure abe scheme for access management in blockchain-based iot,” *IEEE Access*, vol. 9, pp. 54 840–54 849, 2021.
- [68] B. Bera, S. Saha, A. K. Das, and A. V. Vasilakos, “Designing blockchain-based access control protocol in iot-enabled smart-grid system,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5744–5761, 2021.
- [69] S. Mandal, B. Bera, A. K. Sutrala, A. K. Das, K.-K. R. Choo, and Y. Park, “Certificateless-signcryption-based three-factor user access control scheme for iot environment,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3184–3197, 2020.
- [70] W. Ren, Y. Sun, H. Luo, and M. Guizani, “Siledger: A blockchain and abe-based access control for applications in sdn-iot networks,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4406–4419, 2021.
- [71] S. Ding, J. Cao, C. Li, K. Fan, and H. Li, “A novel attribute-based access control scheme using blockchain for iot,” *IEEE Access*, vol. 7, pp. 38 431–38 441, 2019.

Brain Waves Pattern Recognition Using LSTM-RNN for Internet of Brain-Controlled Things (IoBCT) Applications

Mokhles M. Abdulghani
University of North Dakota
Fargo, North Dakota
mukhlisalrawi@ieee.org

Olivier Franza
Intel Corporation
Hudson, Massachusetts
Olivier.franza@intel.com

Farah Fargo
Intel Corporation
Hudson, Massachusetts
farah.e.fargo@intel.com

Haider Raad
Xavier University
Cincinnati, Ohio
raadhd@xavier.edu

Abstract— *A great number of individuals suffer from spinal cord injuries which lead to drastic limitations to their daily life functions. Specifically, suffering from paraplegia and quadriplegia will deem the patient unable to move two limbs or all four limbs respectively. However, even if the spine is damaged, this does not stop the brain from functioning, and it will still have the ability to distinguish objects. Further, the implementation of Brain-Machine Interface (BMI)-based IoT system suffers from several challenges such as the issues of accurately translating the user intention. This paper presents a method for brain waves recognition using deep learning (DL) based on shapes and colors for use in merging concepts of the internet of things (IoT) and the brain computer interface (BCI) which we wish to call “Internet of Brain Controlled Things or (IoBCT) in short. The results showed an acceptable accuracy of 0.93 in the brain waves pattern’s recognition which opens the way towards designing a reliable IoBCT.*

Keywords: *Deep Learning, DL, EEG, IoT, Brain-Computer Interface, BCI, Brain-Machine Interface, BMI, IoBCT, LSTM.*

I. INTRODUCTION

The number of connected devices to the Internet is growing exponentially with over a 50 billion devices connected through Edge and Cloud computing. As a result, the Internet of Things (IoT) architecture is necessary to enable the connected devices to work efficiently and positively impact different domains such as healthcare, home automation system, transportation and other industries [1]. IoT provided internet for individual devices to control, connect and exchange data using smartphones, wearable devices or voice and gestures applications.

One of the methods used to support the interactions of individuals and electronic devices via brain waves is the concept of Brain-Machine Interface (BMI). BMI opened the door to control objects such as robots, prosthetics, and smart home appliances using human thoughts. BMI enables the interaction between machine and human through the establishment of a communication between an external device and the human brain [2]. Moreover, modern studies revealed that BMI is the way to translate human thinking

into physical actions such as Brain-controlled wheelchairs and IoT-enabled appliances [3], [4], and [5].

BMI provides many advantages such as brainwaves observation and human interactions with brain signals on runtime. One of the biggest challenges in BMI is signal accuracy. For example, the brain signals can be observed thorough several methods such as Electroencephalogram (EEG) [2], Functional Near-Infrared Spectroscopy (fNIR) [18], and Magnetoencephalography (MEG) [19]. Those methods can be negatively affected by several factors such as noise, environmental changes, and high brain concentration which makes the brain signal suffer from low signal to noise ratio and lack of temporal-spatial locality [6].

As presented in our previous work [6], the number of IoT devices is increasing rapidly at the edge of the network to be computed at the data centers, which is pushing network bandwidth requirements to the limit. We have also reported that the combination of BMI with the IoT results in a new concept coined as (IoBCT). IoBCT can offer great potentials to improve the quality of life of the disabled people who form about %15 of the world population [7]. Many applications can be operated using the concept of IoBCT such as controlling a smart home [8].

II. RELATED WORK

Deep learning is a fascinating tool used to analyze, process, and classify EEG data. The main obstacle in obtaining acceptable EEG dataset is the type of sensors or the EEG headsets used in the data acquisition stage, which usually are quite expensive and difficult to use [11]. L. Vezard et al. [12] have achieved 71.6% accuracy in a binary alertness states (BAS) prediction by applying the common spatial pattern (CSP) to extract the pattern feature. The methods in [13] and [14] have achieved accuracies of only 54.6 % and 56.76 % o, respectively, through applying multi-stage (CSP) for the EEG dataset feature extraction. While [15] took the power of deep learning networks and employed the recurrent neural network (RNN) to classify the EEG dataset and extract the features over time. The achieved accuracy was 85%.

Long-short term memory (LSTM) network has been used in [16], [17]. In [16] 84% accuracy of brainwave classification has been achieved. EEG dataset has been obtained while providing visual stimuli. High cost EEG headset has been used in [17] and therefore an accuracy of 98% has been achieved in classifying EEG dataset to learn human intentions to move eyes, feet and hands.

In this paper, low-cost EEG headset has been used to process and classify human brainwaves while looking at different images. LSTM-RNN has been selected to classify the brainwaves and extract temporal features from the time-series input signals with different frequencies and length. LSTM-RNN has improved the accuracy of EEG brainwave data classification. Moreover, the overfitting problem in the learning process has been avoided through analyzing and possessing the EEG dataset prior to getting them to the learning stage.

III. THE PROPOSED IoBCT SYSTEM

The brain produced signals (neural oscillations) have been recorded using the EPOC+ headset in this work which is manufactured by Emotiv company. Figure 1 shows the EPOC+ headset. Then it can be analyzed as an electroencephalogram or EEG. High beta, responsible for sensorimotor function and gamma, responsible for processing information from the senses, were the focus. During the EEG signals recording, an individual was looking to a PowerPoint slides containing black image which is the reset and control, and several successive slides of a Rorschach and a fractal images [9], [10]. The black image was presented before and after the Rorschach and a fractal images as a rest to start new recording or to stop the previous one. In addition to that, whenever the two images were presented, the 14 channel EEG signals were recorded using the Emotiv EPOC+ headset. Figure 2 and Figure 3 shows the Rorschach and a fractal images respectively.



Fig. 1: The Emotive EPOC+ EEG headset [5].

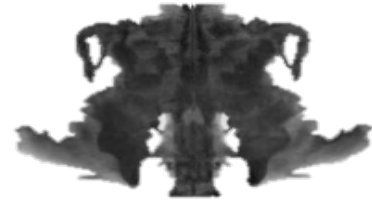


Figure 2: Rorschach image [9].

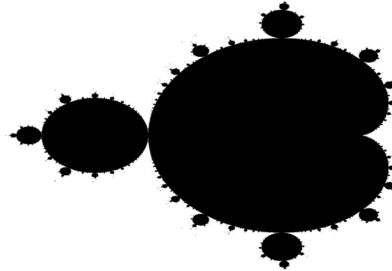


Figure 3: Fractal image [10].

A. Autonomic IoT- EEG based Framework

The framework is based on leveraging the classified brain waves which will be used as a control commands. Figure.4 shows the IoBCT framework, where we have different brain activities collected through IoT sensors. After the data is computed and processed, our framework will retrieve the output with the required action into the IoT destination such as Smart cities, smart hospitals, or smart homes.

Deep learning algorithms will be training and discovering the most important features from the brain signal that has high correlation. All the data will be trained during the offline model to discover the most sensitive features to latency, which then will be analyzed and stored in the database. On runtime, the online data would be analyzed and compared with the offline trained data to identify the higher priority features from the brain signal to optimize the latency.

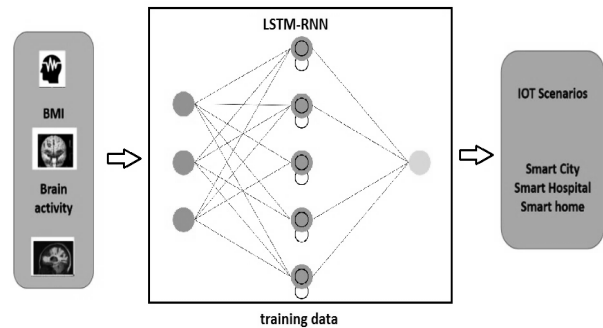


Figure 4: IOBCT framework

B. Deep learning based BMI

An LSTM network is a type of recurrent neural network (RNN) that can learn long-term dependencies between time steps of sequence data by memorizing the input features in RNN-cell memory. LSTM-RNN algorithm updates, forgets and outputs the RNN-cell memory information using multiple gated recurrent units (GRUs). The core components of an LSTM network are a sequence input layer and an LSTM layer. A sequence input layer inputs sequence or time series data into the network. An LSTM layer learns long-term dependencies between time steps of sequence data. The network starts with a sequence input layer followed by an LSTM layer. To predict class labels, the network ends with a fully connected layer, a soft-max layer, and a classification output layer. Figure 5 illustrate the chosen LSTM architecture.

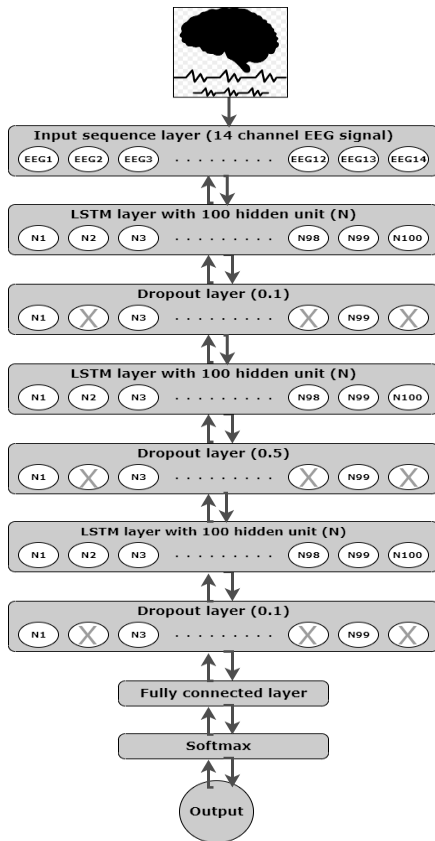


Figure 5: The architecture of LSTM network used to classify the EEG signals

The brain waves classification has been done using deep as follows:

• **Recording the EEG data:** Fourteen channel signals have been recorded using Emotive EPOC+ head seat during 25 sessions of looking to two different images (25 session for each image). The sessions of recording have been implemented by a 21 years old male while looking at Fractal

and Rorschach pictures. Ten seconds for each session with 10 seconds looking at black background image between each of them. Figure 6 depicts the recorded raw EEG signals set for the first session while looking to Rorschach picture.

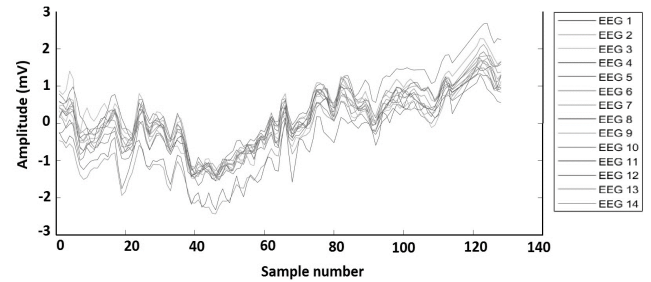


Figure 6: Raw EEG signal set

• **Processing the EEG data:** Before training the LSTM network, the raw EEG signals data still need features extraction to enhance the accuracy of the output classification. The data has been divided to training and testing data (70% for training and 30% for testing). The EEG data has been recorded from 14 EEG sensor and it is consist of different frequencies with different amplitude range. Therefore, it was so important to normalize the EEG data to help to speed up the training process and obtain as much accurate results as possible. The training and testing data have been normalized by calculating the pre-feature mean and standard deviation for all the input sequences. Then, the mean value has been subtracted from each of the training and testing observation and the results for both of them have been divided by the standard deviation as follows:

$$EEG_{Normalised} = \frac{x - \mu}{\sigma}$$

Where (x) is the raw EEG signal, (μ) is the calculated mean value and (σ) is the calculated standard deviation for the recorded EEG dataset. After the normalization process, the data is now ready for the training process. Figure 7 shows the normalized 14 channel EEG signals for the recorded first session while looking to Rorschach picture.

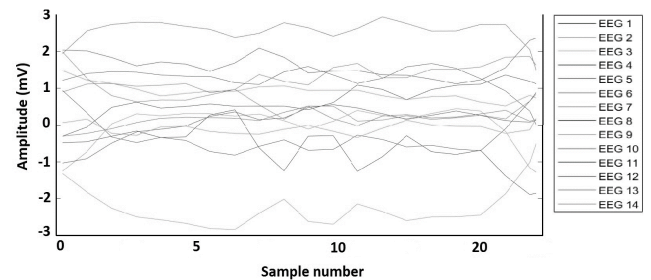


Figure 7, normalized EEG signal set

•**Training LSTM network:** The processed data has been used later to train a deep learning model using Long Short-Term Memory (LSTM) network on MATLAB 2019a. Since we have a sequence of inputs (14 time series signals) and a vector output (vector of zeros for the Factual picture and vector of ones for the Rorschach picture), the LTMS network has been designed as sequence input to vector output and structured as follows: Sequence input layer with number of features equal to the number of channels of the recorded EEG data (14 channels), 3 LSTM layers with 100 hidden unit for each. To prevent overfitting, a regularization technique has been implemented by separating each of the 3 LSTM layers with a dropout layer. The dropout layers will randomly set 10%, 50% and 10% of the training parameters to zero in the first, second and third LSTM layer respectively (dropout ratios 0.1, 0.5 and 0.1). The network has been ended with a fully connected, softmax and a classification output layers with the number of class labels equal to the desired number of the outputs (2 outputs).

IV. RESULTS AND ANALYSIS

An accuracy of 0.93 has been achieved when testing the resulted LSTM network with the remaining 30% of the normalized EEG dataset. These results have been obtained with the help of the adaptive moment estimation (Adam) algorithm. Adam algorithm has been used as an optimization algorithm to tune the hyper-parameters of the LTSM-RNN model. After training the LSTM-RNN model on 70% of the recorded EEG dataset with a 100 max Epochs and 25 for mini batch size the above accuracy has been obtained. Figure 8 shows the model accuracy in the training stage. With the designed LSTM network, it was possible to achieve almost 93% accuracy to classify a normalized EEG dataset for two different activities while looking at two different pictures using low cost EEG headset.

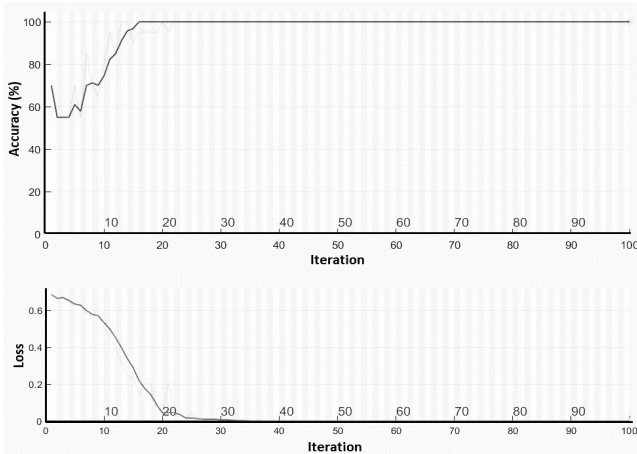


Figure 8, LSTM accuracy after the training process

Table 1. Shows the predicted outputs for the designed LSTM-RNN network to classify the recorded EEG dataset (the test data) while looking to the Rorschach and Fractal images.

Table 1. The predicted outputs of the designed LSTM-RNN network

Input (14 channel EEG signal)	Desired Output	Predicted Output
Rorschach image	1	1
Rorschach image	1	1
Rorschach image	1	1
Rorschach image	1	1
Rorschach image	1	1
Rorschach image	1	1
Rorschach image	1	1
Fractal image	0	1
Fractal image	0	0
Fractal image	0	0
Fractal image	0	0
Fractal image	0	0
Fractal image	0	0
Fractal image	0	0
Fractal image	0	0

V. CONCLUSION AND FUTURE WORK

Although better accuracy has been achieved by other research work to classify the human brainwaves, this work proves the possibility of using a low cost EEG headset (such as Emotiv EPOC+) with the concept of deep learning to process and classify the human thoughts. With the concept of IoT, the classified brain thoughts can be used as a control command to help the disabled people around the world to improve the quality of their life. However, due to a lack in the interface software between the MATLAB and the Emotive EPOC+, all the testing and training stages have been implemented offline without an online execution. Further work will be conducted in the future using different EEG headset with the possibility of having online brain waves testing on MATLAB to assess the classification accuracy of the implemented LSTM model online. Moreover, more than one task will be implemented over the internet using the concept of IoT.

REFERENCES:

- [1]. L. Yao, Q. Z. Sheng, and S. Dustdar, "Web-based management of the internet of things," *IEEE Internet Computing*, vol. 19, no. 4, pp. 60–67, 2015.
- [2]. A. Vallabhaneni, T. Wang, and B. He, "Brain—computer interface," in *Neural engineering*. Springer, 2005, pp. 85–121.
- [3]. A. Teles, M. Cagy, F. Silva, M. Endler, V. Bastos, and S. Teixeira, "Using brain-computer interface and internet of things to improve healthcare for wheelchair users." *The Eleventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies 2017*, ISBN: 978-1-61208-598-2.
- [4] Jagadish, B.; Kiran, M. P. R. S.; Rajalakshmi, P. "A novel system architecture for brain controlled IoT enabled environments", 19th International Conference on e-Health Networking, Applications and Services (Healthcom), 1–5. doi:10.1109/HealthCom.2017.8210814.
- [5] Mokhles M.Abdulghani; Al-Aubidy, Kasim M. "Wheelchair Neuro Fuzzy Control Using Brain Computer Interface", [IEEE 2019 12th International Conference on Developments in eSystems Engineering (DeSE) - Kazan, Russia, 640–645. doi:10.1109/DeSE.2019.00120.
- [6] H. Raad, F. Fargo, O. Franza, "Autonomic Architectural Framework for Internet of Brain Controlled Things (IoBCT)". *ISNCC 2021*.
- [7]. W. H. Organization, "World report on disability 2011," World Health Organization, Valletta, Malta2011.
- [8]. Nafea, Marwan; Hisham, Amirah 'Aisha Badrul; Abdul-Kadir, Nurul Ashikin; Che Harun, Fauzan Khairi "Brainwave-Controlled System for Smart Home Applications", [IEEE 2018, 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS) - Malaysia , 75–80. doi:10.1109/ICBAPS.2018.8527397.
- [9]. Abbott, Alison "Fractal secrets of Rorschach's famed ink blots revealed" *Nature*, nature.2017.21473–. doi:10.1038/nature.2017.21473.
- [10]. Vicsek, Tamás (1992). *Fractal growth phenomena*. Singapore/New Jersey: World Scientific. pp. 31, 139–146. ISBN 978-981-02-0668-0.
- [11]. Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *J. Neural Eng.*, vol. 16, no. 5, p. 051001, Aug. 2019.
- [12]. L. Vezard, P. Legrand, M. Chavent, F. Faïta-Aïnseba, and L. Trujillo, *Eeg classification for the detection of mental states, Applied Soft Computing*, (2015).
- [13]. H. Meisheri, N. Ramrao, and S. K. Mitra, *Multiclass common spatial pattern with artifacts removal methodology for eeg signals*, in *ISCBI, IEEE*, 2016.
- [14]. T. Shiratori, H. Tsubakida, A. Ishiyama, and Y. Ono, *Three-class classification of motor imagery eeg data including rest state using filter-bank multi-class common spatial pattern*, in *BCI, IEEE*, 2015.
- [15]. P. Bashivan, I. Rish, M. Yeasin, and N. Codella, *Learning representations from eeg with deep recurrent-convolutional neural networks*, arXiv, (2015).
- [16] J. J. Bird, D. R. Faria, L. J. Manso, A. Ekárt, and C. D. Buckingham, "A Deep Evolutionary Approach to Bioinspired Classifier Optimisation for BrainMachine Interaction," *Complexity*, vol. 2019, pp. 1–14, Mar. 2019.
- [17] W. Chen et al., "EEG-based Motion Intention Recognition via Multi-task RNNs," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018, pp. 279–287.
- [18] M. A. Rahman and M. Ahmad, "Evaluating the connectivity of motor area with prefrontal cortex by fnir spectroscopy," in *ECCE. IEEE*, 2017, pp. 296–300.
- [19] M. Iijima and N. Nishitani, "Cortical dynamics during simple calculation processes: a magnetoencephalography study," *Clinical Neurophysiology Practice*, vol. 2, pp. 54–61, 2017.

Low-Power and High Speed SRAM for Ultra Low Power Applications

Neha Meshram, Govind Prasad
Dept. of ECE
 IIT Naya Raipur
 Naya Raipur, India
 neha20301@iitnr.edu.in

Divaker Sharma
Dept. of ECE
 Jamia Millia Islamia
 Delhi, India
 divakersharma17@gmail.com

Bipin Chandra Mandi
Dept. of ECE
 IIT Naya Raipur
 Raipur, India
 bipin0087@gmail.com

Abstract—The rapid development of battery-powered gadgets has made low-power design a priority in recent years. In addition, integrated SRAM units in contemporary soCs have become an essential component. The increased number of transistors in SRAM units and the increased leakage in scaled technology of the MOS transistors have turned the SRAM unit into a power block from dynamic and static perspectives. This memory circuitry consumes many chips and determines the system's overall power consumption. Typically, the primary 6T SRAM cell gives more power loss and delay. In this paper, various SRAM transistor cells have been built and analyzed from different topologies. A proposed low-power 9T SRAM cell area has improved reading and writing access time. As anticipated from the modeling findings, experimental results show a significant overall power decrease compared to traditional and previously published.

Index Terms—Dynamic Power, high speed, process analysis, SRAM, stability, static power.

I. INTRODUCTION

SRAM is intended to work in close collaboration with the central processor unit (CPU) and consumes less power to conserve battery life due to the fast proliferation of portable devices [1]- [5]. The speed and power parameters are critical in boosting the capabilities of electronic devices used in research and commercial facilities. With the growing need for low-power SRAMs, low-power design methodologies are concentrating their efforts on a few specific sources of power consumption [6]-[8]. Low-power methods of decreasing the SRAM cell's leakage current are necessary because of the increased leakage current associated with scaled technology. The high capacitance of the long interconnecting wires also contributes to the SRAM unit's large dynamic power consumption. Thus, the electronic industry's need for tiny and portable electronic gadgets with low power consumption and fast speed has developed significantly [9]-[13]. Therefore, designing an SRAM-based memory with optimum power and stability without sacrificing performance is tricky.

The traditional 6T cell design is relatively simple in various applications [14]-[18]. It has excellent conduction and performs well in terms of delay and power. The power dissipation of 6T, Comparatively high. The numerous types of SRAM cells, such as 7T SRAM [26], 8T SRAM, and 9T SRAM [26], are overviewed to the power consumption problem of the 6T SRAM. High static power is provided by the 6T SRAM cell [19]-[22]. An 8T is chosen and implemented to achieve low static power. The 8T SRAM cell is more stable and uses less power than the 6T SRAM cell. A new 9T SRAM cell has been proposed to overcome these limitations. The suggested SRAM has a lower power usage than a 6T SRAM [20]. Compared to a standard 6T cell, the proposed cell includes less outgoing current. Several techniques are discussed for reducing the SRAM cell's outgoing currents [23]-[25]. The power dissipation of the SRAM cell is particularly targeted.

The rest of the paper is structured as follows: The conventional SRAM cells are explained in Section II. Section III describes the proposed SRAM cell. Section IV presents and discusses the experimental results using the Cadence tool. The paper concludes with Section V.

II. STANDARD SRAM CELLS

SRAM is made up of two cross-coupled inverters in such a fashion that it store data and provide positive feedback. When powered, it functions as a volatile memory, storing data. It accelerates 'read' and 'write' operation using BL and BLB bit lines.

A. 6T SRAM Cell

The CMOS standard 6T SRAM cell has two PMOSs, two NMOSs, and two NMOSs as access transistors. Two data-storage inverters are manufactured and cross-coupled so that positive feedback is generated. The two inverters are P0, N0, and P1, N1. It also has vertically aligned Bit Lines

(BL and BLB) and a horizontally oriented Word Line (WL) [20]. SRAM has three modes of operation: standby, read, and write. Vdd is provided to BL during standby mode, and WL is switched off, causing the transistors connected to it to turn off as well.

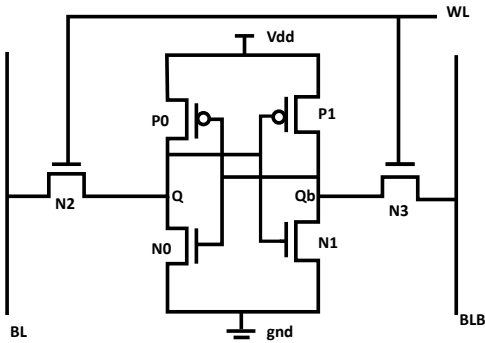


Fig. 1: Schematic diagram of 6T SRAM cell.

The two cross-coupled inverters provide positive feedback since $BL = V_{dd}$, allowing data storage while power is supplied. Only read and write operations are permitted on the WL line, which controls the state of the access transistors.

Sense Amplifier: A sense amplifier is required for memory circuits to attain high performance, endurance, and usefulness. It is a part of the read circuitry responsible for reading data from memory. Its purpose is to monitor low-power signals from a bit line and convert the tiny voltage difference to full logic voltage, representing the data bit stored in a memory cell, thus significantly reducing read time. It includes voltage sensing, low-voltage amplification, delay reduction, power consumption [4] reduction, and signal restoration capabilities. They are mainly used to amplify the differential voltage across complementary bit lines during 'read' operations without flipping the stored cell data, enabling the data to be adequately handled by the memory's output circuitry. To function correctly, the SRAM cell must have a Static noise margin (SNM) since this determines the dependability of sensing data from the chosen cell [5]. Since the driving transistors do not need to discharge the bit lines completely, the inclusion of a sensing amplifier reduces the size of the memory cell. The sensing amplifier's design significantly impacts the read speed and power consumption.

It has seven transistors, two of which serve as differential input devices (N0 and N1). They function as driver transistors and accept inputs, while transistors P0 and P1 work as active current mirror loads, transistors N3 are current sources, and transistors P2 and N4 are inverter transistors (amplifier). The control circuit generates a sense enable

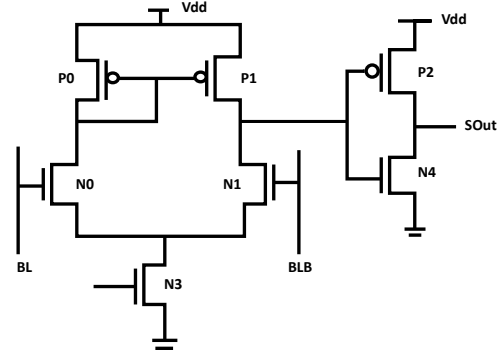


Fig. 2: Schematic diagram of Sense Amplifier.

signal (SE) to activate the sense amplifier. When this enabled signal is high, the sensing amplifier increases the differential voltage of the bit lines (active). Amplification of differential voltage results in a single-ended digital output reflecting one bit read from the SRAM array.

Pre-charge circuits: It is used to quickly and painlessly charge both bit-lines voltages to VDD before each read and write operation. The pre-charge circuit is shown. It comprises two pull-up PMOS transistors and an equalization that balances the voltages on both bit lines [4]. The pull-up transistors are controlled by the PR signal, which controls the transistor P3, an equalization signal that balances the voltage on both bit lines. The term "Vdd pre-charge" refers to the pre-charge. PMOS transistors are utilized in the pre-charge circuit because they are the most efficient at transmitting the Vdd level voltage to the bit lines.

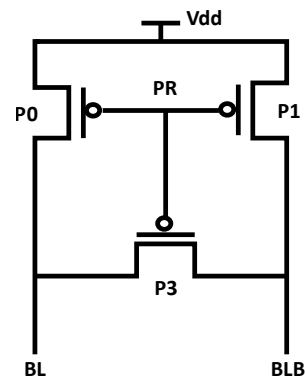


Fig. 3: Schematic diagram of Pre-charge circuit.

When the word lines are activated, the memory cell nodes must raise the voltage on the bit lines to a level adequate to read the correct data value from memory during a read operation. Due to the high load capacitance of the bit lines in large memory arrays, charging and discharging them takes

a lengthy time. Waiting for the bit-line capacitance to drain before performing the next read operation may be an issue since it adds time to the 'read' process. Additionally, an error may result if the subsequent read operation occurs before the bit lines have been discharged. Consequently, a pre-charge circuit is required to provide a high current to bit lines to charge them fast. To do this, a pre-charge circuit is used to charge both bit-lines to a steady voltage Vdd. Since the voltage on both bit lines is known and steady, the voltage difference may be easily detected.

Write access time: The time interval between the positive edge of the word line reaching 50% Vdd and the junction of the two storage nodes Q and Qb are defined as the memory cell's write access time.

Read access time: The period between the positive edge of the word line hitting 100% Vdd and the sensing amplifier output reaching 100% of the stored value is the read access time of a memory cell.

B. 7T SRAM Cell

Before write operation, The 7T SRAM cell disables the feedback link between the two inverters inv1 and inv2. An additional NMOS transistor N5 handles feedback connection and disconnection, and the cell is completely reliant on the BL bar for a write operation.

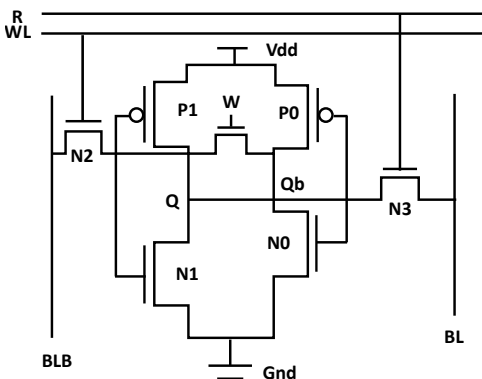


Fig. 4: Schematic diagram of 7T SRAM cell.

C. 8T SRAM Cell

The architecture of the 8T SRAM cell is depicted in the figure. It features two extra transistors N4 and N5 as compared to a standard 6T SRAM cell. Precharging BL is the first step in the read process. The reading is then initiated by setting RD to Vdd [3]. If bit 1 is stored in the cell, Q is at Vdd and Qb is at the ground. During the read process, the precharged BL begins to discharge via N4 and N5. Discharging should be fast enough to keep up with the

rate at which BL leaks through all of the unaccessed SRAM cells connected to it. As a consequence, the read circuit can correctly identify the data.

If bit 0 is stored in the cell, BL should not discharge and should stay at or near Vdd. In this case, the read circuit should calculate the value of BL before leakage reduces it considerably. To assist retain the value one in the bit line, we connected a weak pull-up to the read circuit. Because short bit lines have low capacitance, they facilitate the charging and discharging of SRAM cells and peripheral circuits [10]. As a consequence, transistor sizes are reduced, leakage is reduced, and reliability is enhanced.

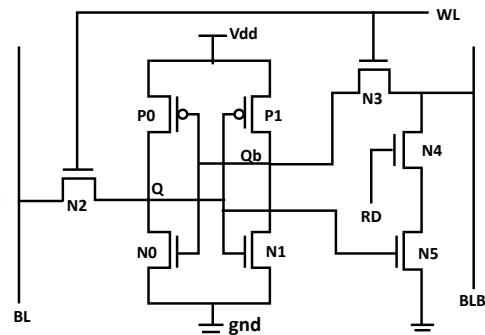


Fig. 5: Schematic diagram of 8T SRAM cell.

D. 9T SRAM Cell

The 9T SRAM's top circuit is identical to the 6T SRAM's but with fewer transistors. The two access transistors are controlled by the word line (W), and the data is stored in this subcircuit. A separate read signal (R) controls the bottom circuit, which consists of a single read access transistor [26].

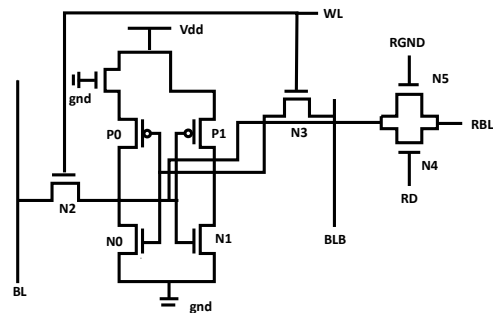


Fig. 6: Schematic diagram of 9T SRAM cell.

III. THE PROPOSED 9T SRAM CELL

A new 9T SRAM cell is being developed. Compared to a standard 6T SRAM cell, it has three more transistors, N4, N5, and N6. Figure 7 shows a single-sided read circuit. It

takes up less space and does not rely on the perfect timing of control signals, as the differential read of a typical sense amplifier circuit does. Figure 7 depicts the write circuit. One of these circuits is for BL, and the other receives the data bit and its inversion as input. Because PMOS transistors are larger than NMOS transistors, their pullup and pulldown currents are symmetrical, allowing the write circuit to drive both 0 and 1 to the bit lines. The reliability and static power

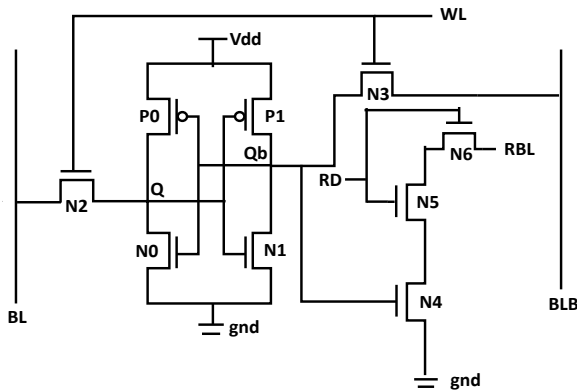


Fig. 7: Schematic diagram of Proposed 9T SRAM cell.

consumption of transistors are the primary determinants of transistor size. The dimensions used are listed in Table I. P0 and P1 have the narrowest widths allowed by the technology. Vdd leak currents are reduced as a result. In 6T SRAM, the pull-down NMOS transistors N0 and N1 are wider than the access transistors N2 and N3. This improves cell stability during read operations when 6T SRAM is most susceptible to undesirable state change. Because N2 and N3 are not used during read operations, the widths of the pull-down and access transistors in the proposed 9T SRAM have been flipped, making write operations faster and more reliable. The width of the N4, N5, and N6 of 9T SRAM decrease leakage.

IV. SIMULATION AND RESULT

The proposed 9T and other comparative SRAM cells are simulated at a 45-nm technology node in this work. The cadence virtuoso simulation tool is used for simulations.

A. Transient Response and Power Consumption

Power savings are significant in the proposed 9T cell structure, regardless of the stored value, i.e., whether the SRAM cell stores 0 or 1. When a bit cell in the 7T structure stores a 1, the PMOS is enabled, leading to significant power consumption in standby mode. When the bit cell stores a 0, the voltage difference between the drain and source raises the sub-threshold leakage via the PMOS. This grows

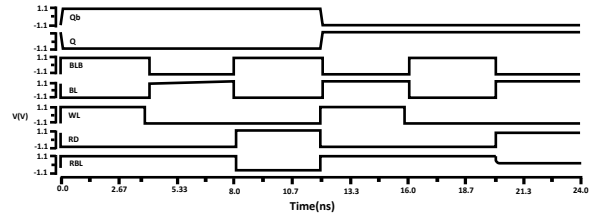


Fig. 8: Transient Responses of the proposed 9T SRAM cell.

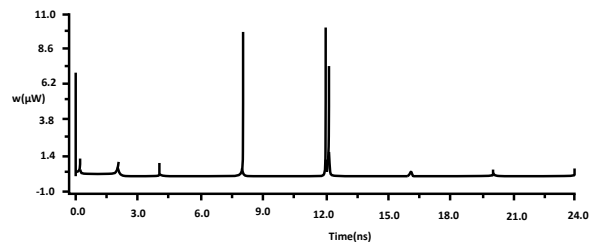


Fig. 9: Power consumption of the Proposed 9T SRAM cell.

exponentially as the threshold voltage decreases, indicating a significant leakage current component in the off-state. The average standby power consumption of the 7T structure in Figure 4, the 8T structure in Figure 5, and the 9T structure in Figure 6, the proposed 9T structure in Figure 7 when the SRAM cell stores a 0 is shown in Figure for various power supply voltages, the proposed 9T cell achieve minimum power savings in each bit cell without sacrificing performance.

B. Analysis of Stability Using the N-Curve Method

The static noise margin (SNM) is a metric for the SRAM cell's stability. The read, write and hold modes of this SNM are used to get the read SNM, write SNM, and hold SNM, respectively. Generally, the butterfly method is employed to determine the SNM in various modes [26]. However, the disadvantage of the butterfly technique is that it only provides information about SNM; deriving SINM (Static Current Noise margin) requires mathematical calculations. Two distinct circuits must be examined to achieve read and write SNM. Another technique for determining SNM is the N-curve method, which provides both voltage and current information concurrently. The cell's read stability and write capabilities are also determined directly from the N-curve. The N-curves for all SRAM and planned SRAM cells are shown in fig. below. The values for SVNM, SINM, WTV, and WTI are as given in Table I.

C. Analysis of Stability Using the Butterfly Method

The SRAM cell's ability to store data against noise is measured in SNM. The SNM of SRAM is defined as

TABLE I: Comparative Analysis of SRAM cells

Parameters	Proposed 9T	6T [3]	7T [6]	8T [8]	9T [25]
Total Power Dissipation (μW)	0.09041	25.48	25.26	6.101	0.1154
Read Time Access (ns)	68.938	191.74	141.85	155.177	69
Write Time Access (ns)	78	92.562	70.5028	79.058	78
Static Voltage Noise Margin (mV)	423.32	369.094	371.581	359.41	393.51
Static Current Noise Margin (μA)	74	36.094	33.394	55.535	51.650
Write Trip Voltage (mV)	513.79	522.57	540.86	496.99	511.96
Write Trip Current (μA)	-12.208	-8.6711	-9.1072	-9.4437	-8.6705
Static Power Noise Margin (μW)	31.308	13.35	12.40	19.9598	20.325
Write Trip Power (μW)	-6.2723	-4.531	-4.9257	-4.6934	-4.439

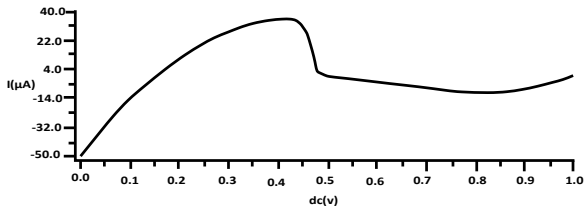


Fig. 10: N-curve Analysis of Proposed 9T SRAM.

the minimum amount of noise voltage required to flip the state of a cell on the SRAM’s storing nodes. The static voltage transfer characteristics of the SRAM cell inverters are used in the graphical technique to determine the SNM. It multiplies one cell inverter’s voltage transfer characteristic (VTC) by the inverse VTC of the other inverter [14].

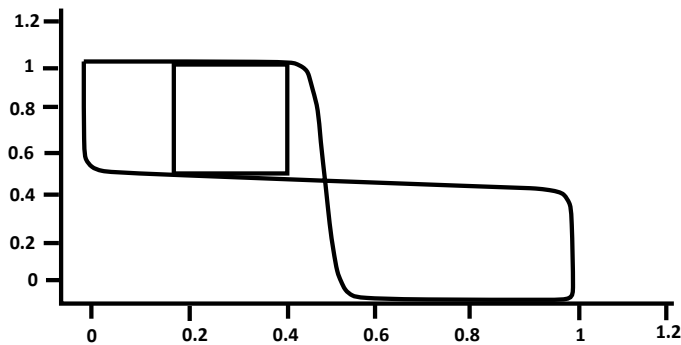


Fig. 11: Butterfly Curve Analysis of Proposed 9T SRAM cell.

The greater the SNM value, the better the read stability of the SRAM cell. The read stability of a cell with a lower RSNM is poor.

V. CONCLUSION

The performance of three SRAM cell topologies in terms of stability has been demonstrated. SRAM speeds will increase as process technologies advance, but devices will be more prone to mismatches, reducing the static noise margin of SRAM cells. The 6T SRAM’s RNM is relatively low. The width of the pull-down transistor must be raised to get a larger RSNM in 6T SRAM cells. However, this increases the area of the SRAM, which increases leakage currents. The read noise margin of an 8T SRAM cell is significantly higher. An 8T SRAM cell may fail during a write operation due to the asymmetric cell structure. Compared to 6T and 8T SRAM cells, 9T SRAM cells have stronger RSNM and WSNM, resulting in improved stability performance.

REFERENCES

- [1] S. Tamimi, Z. Ebrahimi, B. Khaleghi and H. Asadi, "An Efficient SRAM-Based Reconfigurable Architecture for Embedded Processors," in IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 38, no. 3, pp. 466-479, March 2019.
- [2] R. Giterman, L. Atias, and A. Teman, "Area and Energy-Efficient Complementary Dual-Modular Redundancy Dynamic Memory for Space Applications," in IEEE Trans. Very Large Scale Integration (VLSI) Systems, vol. 25, no. 2, pp. 502-509, Feb. 2017.
- [3] S. Pal and A. Islam, "Variation Tolerant Differential 8T SRAM Cell for Ultralow Power Applications," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 35, no. 4, pp. 549-558, April 2016.
- [4] C. -C. Wang, D. -S. Wang, C. -H. Liao and S. -Y. Chen, "A Leakage Compensation Design for Low Supply Voltage SRAM," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 5, pp. 1761-1769, May 2016.
- [5] A. Do, Z. Kong, K. Yeo, and J. Y. S. Low, "Design and Sensitivity Analysis of a New Current-Mode Sense Amplifier for Low-Power SRAM," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 19, no. 2, pp. 196-204, Feb. 2011.
- [6] G. Prasad, D. Tandon, Isha, B. C. Mandi, and M. Ali, "Process Variation Analysis of 10T SRAM Cell for Low Power, High-Speed Cache Memory for IoT Applications," 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), 2020.

- [7] R. Bishnoi, F. Oboril and M. B. Tahoori, "Design of Defect and Fault-Tolerant Nonvolatile Spintronic Flip-Flops," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1421-1432, April 2017 .
- [8] G. Razavipour, A. Afzali-Kusha and M. Pedram, "Design and Analysis of Two Low-Power SRAM Cell Structures," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 10, pp. 1551-1555, Oct. 2009.
- [9] F. Hameed, A. A. Khan and J. Castrillon, "Performance and Energy-Efficient Design of STT-RAM Last-Level Cache," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 6, pp. 1059-1072, June 2018.
- [10] Ruchi and S. Dasgupta, "Compact Analytical Model to Extract Write Static Noise Margin (WSNM) for SRAM Cell at 45-nm and 65-nm Nodes," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 1, pp. 136-143, Feb. 2018.
- [11] G. Prasad and A. Anand, "Statistical analysis of low power SRAM cell structure," *Analog Integrated Circuit and Signal Processing (Springer)*, vol. 82, no. 01, pp. 349-358, Jan. 2015.
- [12] D. Nayak et al., "A high stable 8T-SRAM with bit interleaving capability for minimization of soft error rate," *Microelectronics Journal (Elsevier)*, vol. 73, pp. 43-51, Mar. 2018.
- [13] K. Dhanunjaya et al., "Cell stability analysis of conventional 6T dynamic 8T SRAM cell in 45nm technology," *Inter. J. VLSICS*, vol.3, no.2, Apr. 2012.
- [14] P. Elakkumanan et al., "NC-SRAM a low-leakage memory circuit for ultra-deep submicron designs" in *proc. IEEE Inter. SOC Conf.*, pp. 3-6, Sep. 2003.
- [15] G. Razavipour et al., "Design and Analysis of Two Low-Power SRAM Cell Structures," *IEEE Trans. VLSI*, vol. 17, No. 10, Oct. 2009.
- [16] G. Prasad, and R. Kusuma, "Statistical (MC) and static noise margin analysis of the SRAM cells," in *proc. IEEE Students Conference on Engineering and Systems (SCES)*, pp. 1-5, Apr. 2013.
- [17] G. Chen et al., "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," in *proc. IEEE Inter. Solid-State Circuits Conf.*, pp. 288-289, Feb. 2010.
- [18] D. P. Wang et al. "A 45nm dual-port SRAM with write and read capability enhancement at low voltage," in *proc. IEEE Inter. SOC Conf.*, pp. 211-214, Sep. 2007.
- [19] G. Prasad et al. "Novel low power 10T SRAM cell on 90nm CMOS," in *proc. IEEE Inter. Conf. on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 109-114, Feb. 2016.
- [20] G. Prasad, N. Kumari, B. C. Mandi, and M. Ali, "Design and Statistical Analysis of Low Power and High Speed 10T SRAM Cell," *International Journal of Circuit Theory and Applications*, Wiley, vol. 48, no. 8, pp.1319-1328, May 2020.
- [21] G. Prasad, B. C. Mandi, and M. Ali, "Power optimized SRAM cell with High Radiation Hardened for Aerospace Applications", *Microelectronics Journal, Elsevier*, vol. 103, 104843, Sep. 2020.
- [22] G. Prasad, B. C. Mandi, and M. Ali, "Low Power and Write-Enhancement RHBD 12T SRAM Cell for Aerospace Applications", *Analog Integrated Circuit and Signal Processing (2021)*, vol. 107, No. 1, pp. 377-388, Jan. 2021.
- [23] G. Prasad, B. C. Mandi, and M. Ali, "Soft-Error-Aware SRAM for Terrestrial Applications", *IEEE Trans. Device and Material Reliability*, vol. 21, no. 4, pp. 658-660, Dec. 2021.
- [24] G. Prasad, D. Sahu, B. C. Mandi, and M. Ali "Novel Low Power and High Stable Memory Cell Design using Hybrid CMOS and MTJ", *International Journal of Circuit Theory and Applications*, Wiley, 10 Dec. 2021.
- [25] Choudhari, Shruti H., and P. Jayakrishnan. "Structural Analysis of Low Power and Leakage Power Reduction of Different Types of SRAM Cell Topologies." 2019 *Innovations in Power and Advanced Computing Technologies (i-PACT)*. Vol. 1. IEEE, 2019.
- [26] Anitha, D., K. Manjunatha Chari, and P. Satish Kumar. "N-Curve Analysis of Low power SRAM Cell." 2018 *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018.

Smart Home Automation IoT System for Disabled and Elderly

Nesreen Alsbou and Naveen Mohan Thirunilath
 nalsbou@uco.edu, nmohan@uco.edu
 University of Central Oklahoma

Abstract— This paper provides a simulation based home automation system. The system is designed to be automated and controlled remotely through a cloud system. The proposed system enables the user to monitor and control the home environment including temperature, humidity, and detection of intruders. The system sends alerts via email and text messages to the user's phone for continuous monitoring. The cloud feature helps the user to read the data and control the functionality of different appliances at a home using mobile phone. This affordable system is best suited to remotely monitor and take care of disabled and elderly that are limited in movements.

Keywords – Internet of Things (IoT), Smart Home, Packet Tracer, Disabled, Elderly, and IoT Systems.

I. INTRODUCTION

Smart homes use home automation systems to control home appliances and provide automatic remote control inside and outside the home. Despite the comfort and ease of use of remote control, specific key problems must be addressed and improved, such as how to provide an intuitive and user-friendly remote-control scheme in IoT-based smart homes and can be accessed away from home [1]. In this system, using an MKR1000, connected devices are controlled, and data has been read and sent to the cloud in real-time. Various sensors that are kept in different areas of the house will read the values and send them into an Arduino. The Arduino is programmed and set up using the Arduino IoT cloud, which uses these values from the sensor to process and do necessary actions. An OLED display was also connected to the Arduino for the user to see real-time values from the sensors if the user does not want to check the phone.

This system can also be used for a COVID19 patient, which helps the patient to stay in a controlled environment and still would be able to control everything in fingertips. Values like room temperature, lawn soil humidity, light intensity, intruder detection, and appliances control are focused on in this project. In the proposed study, we are building a system with five sensors for data collection and

Two LEDs, a fan, and a buzzer for feedback. The temperature sensor will sense the temperature of the room, and after a certain limit on temperature value, a buzzer will go ON. The humidity sensor can be used to detect the humidity, and if it detects a value below a certain number, a fan will turn ON, which can be used as a lawn sprinkler. A rotary sensor is used to adjust the brightness of a LED. An ON/OFF switch is also added to the cloud dashboard to control a LED remotely. All the desired changes can be received as a notification via email and SMS text messages.

II. LITERATURE REVIEW

The topic of Internet of Things has been studied in various papers, and home automation is easy and popular among them. The main objective of IoT is to manage and control physical objects around us in a more intelligent and meaningful manner and improve quality of life by providing cost-effective living, including safety, security, and entertainment. Mandula introduced a simple Bluetooth-based home automation system in 2015. An Arduino ATMEGA238 was used for the study to control home appliances from user's mobile phones. The mobile is connected using a Bluetooth module HC-05, limited to a 10-meter range [2].

In 2013, Home automation using Zigbee stack had gone to get a faster communication to mobile phones. Their study investigates several methods to equip an Android device with a dongle capable of ZigBee communication. They proposed a scalable architecture with three abstraction layers that scales over multiple communication channels, such as the TCP channel for communication with the gateway, and the USB channel, for direct communication with devices through the dongle [3]. A similar study by B. Balaji in 2020 introduced a domestic home automation system in which an Arduino UNO is used to control home appliances. In the proposed study, an external Wi-Fi module had connected to get the Arduino connected to the Internet [4]. The system is designed again that only if the user is connected to the home network Wi-Fi the access to operate the system is granted.

III. PROPOSED SYSTEM

The proposed system's IoT architecture block diagram is shown in Figure 1. The smart thing we used in the system was Arduino MKR1000, and the sensors were connected to different sensors and actuators from the Arduino. The IoT gateway used in the study is Wi-Fi. The Arduino is connected to the cloud and Webhook via a Wi-Fi gateway. The Arduino IoT cloud is used to program the Arduino and to connect the Webhook Email client to the code. The smartphone will receive text messages and display the values of all connected devices in the dashboard through any available internet connection. The system block diagram is shown in Figure 1.

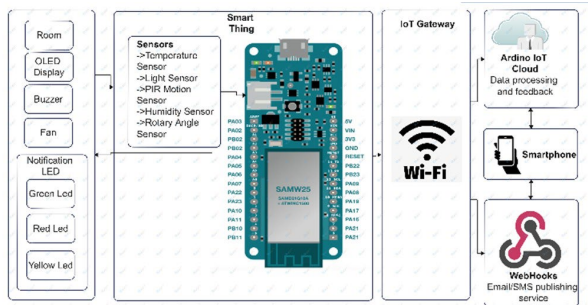


Fig. 1. System Block Diagram

IV. HARDWARE DESIGN

IoT is a structure in which objects, people are provided with an exclusive identity and the ability to move data through a network without requiring two-way inter-human to the human source, i.e., the destination or human interaction to the computer. IoT is a very promising development to optimize life based on intelligent sensors and smart appliances that work together over the Internet [1].

Arduino: Arduino MKR1000, as shown in Figure 1, is the main controller used in the project. The main advantage of this Arduino is the easy network connectivity and cross-platform advantage. The Arduino is open-source with extensible software support in different cloud platforms. It has a low-power ARM microcontroller which would be perfect for a long-life project like this. The device is equipped with 8 I/O pins and 12 PWM Pins which runs on an operating voltage of 3.3V. This will be perfect for connecting enough appliances and working for a relay for the fan in the proposed study. Arduino MKR1000 Wi-Fi is the board that has been chosen because this board has been designed especially for IoT. This board has a more powerful processor than Arduino Uno or Mega. Besides, it has a Wi-Fi WINC15000 chip to connect to the Internet, which allows sending data to devices from the Internet [5]. Different Sensors are used in the proposed system as

described below. Temperature Sensor: A groove temperature sensor that uses a thermistor resistance that varies according to temperature is used. The sensor is shown in Figure 2 and it is connected to the Analog pin A0 of the Arduino.

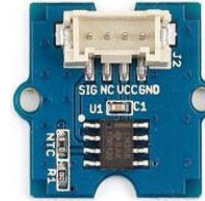


Fig. 2. Temperature Sensor

Light Sensor: The light sensor will detect the light intensity in the room and can be used to display the value on the dashboard for the Arduino IoT cloud, the sensor is shown in Figure 3.

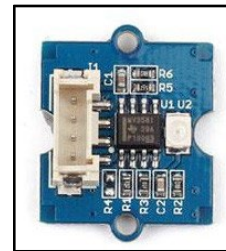


Fig. 3. Light Sensor

Rotary Angle Sensor: The sensor will detect an angle from 0 to 300 degrees. This sensor is shown in Figure 4 and it is known as a potentiometer which is connected to Arduino to vary the light intensity of a LED.

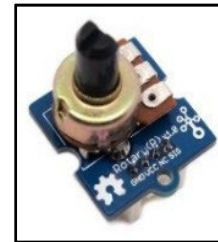


Fig. 4. Rotary Angle Sensor

PIR Motion Sensor: The PIR (Passive Infrared) motion sensor will detect human movements in its range. Once the motion is detected, the sensor will output HIGH and will send that to the Arduino. The sensor is shown in Figure 5.



Fig. 5. PIR motion Sensor

Humidity Sensor: We are using a DHT11 humidity sensor that uses a capacitive humidity sensor to send Analog output to the Arduino. The sensor can be kept outside the house on the lawn to detect soil humidity. The sensor is shown in Figure 6.

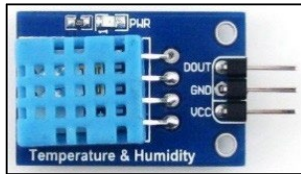


Fig. 6. DHT11 Humidity Sensor

Buzzer: The buzzer is composed of a piezo buffer which emits a tone when the output pin is high. The buzzer will make a sound when the temperature is above a specific threshold value. The buzzer used is shown in Figure 7.



Fig. 7. Buzzer

OLED: Organic Light-emitting diode which is a grayscale 128x128 pixel, it is used to display the temperature and humidity values from the sensor as shown in Figure 8.

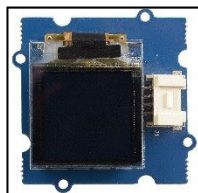


Fig. 8. OLED

V. SOFTWARE DESIGN

Smart Home is an application of a combination of technology and services devoted to the home environment with specific functions aimed at improving the safety, efficiency, and comfort of its inhabitants. The smart home system usually consists of monitoring tools, control devices, and automatic actuators. Several devices can be

accessed using a computer or smartphone connected to the Internet network [6-8]. For the proposed study, a simulation using packet tracer is done. All the appliances are placed on a two- bedroom apartment layout, as shown in Figure 9.

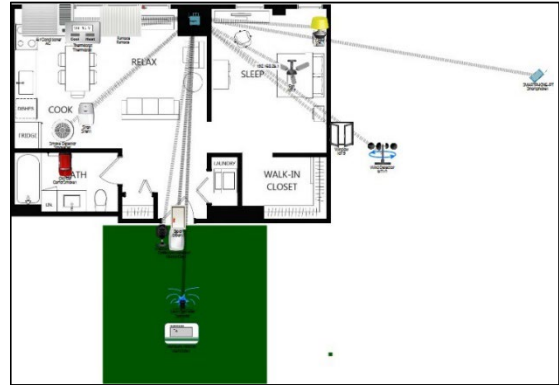


Fig. 9. Cisco Packet Tracer Simulation

Sensors are kept at the door to detect motion and to access the door. A humidity sensor is kept on the lawn to detect humidity and turn the sprinkler when needed. Light ON/OFF switch available on the smartphone. The wind detector will sense the wind and open/close the windows as per the conditions mentioned below in Table 1.

Table 1: Conditions on Cisco Packet Tracer

actions	Enabled	Name	Condition	Actions
E:R: Remote	Yes	Close Window 1	IoT18 Wind is true	Set IoT8 On to true
E:R: Remote	Yes	Door Open	MotionDet1 On is true	Set Door1 Lock to Unlock
E:R: Remote	Yes	Door Close	MotionDet1 On is false	Set Door1 Lock to Lock
E:R: Remote	Yes	Open Window	IoT18 Wind is false	Set IoT8 On to false
E:R: Remote	Yes	SprinklerOn	Humidity1 Number >= 55	Set Sprinkler Status to true
E:R: Remote	Yes	SprinklerOff	Humidity1 Number <= 55	Set Sprinkler Status to false
E:R: Remote	Yes	Motion Cam	MotionDet1 On is true	Set Cam On to true
E:R: Remote	Yes	MotionCam2	MotionDet2 On is false	Set Cam On to false
E:R: Remote	Yes	ScreenOn	SmartDet Level <= 0.52	Set Smart1 On to true
E:R: Remote	Yes	ScreenOff	SmartDet Level <= 0.52	Set Smart1 On to false
E:R: Remote	Yes	ACOn	Thermostat Temperature >= 19.0 °C	Set AC On to true
E:R: Remote	Yes	ACOff	Thermostat Temperature <= 19.0 °C	Set Furnace On to true
E:R: Remote	Yes	ACOff	Thermostat Temperature <= 19.0 °C	Set AC On to false
E:R: Remote	Yes	FurnaceOn	Thermostat Temperature >= 19.0 °C	Set Furnace On to false

The simulation also has a smoke sensor that detects smoke and will turn ON/OFF a siren after a specific value of smoke. Additionally, a Webcam is also placed at the door to detect people coming at the door. The simulation is a demo representation of how the home automation system will work in real life. On the packet tracer, all the components are added to the IoT home gateway network. This helps the user to access the devices from the smartphone.

VI. SECURITY FIREWALL

The goal of adding a IPv4 or IPv6 firewall to the network is to have every device assigned an IP address so that connections can be established directly between them instead of being behind on firewall. Adding a Cisco ASA (Adaptive Security Appliance) firewall to router will

improve the security and it stops threats before they spread through the network. This setup works best for a system which works on a cloud instead of adding just an IPv4 firewall for every device on the Cisco packet tracer. For the proposed study we are using an ASA 5505 on-board accelerator (Version 8.4(2)). The ASA is configured with domain ccnasecurity.com which is provided for the Cisco security network devices. The Cisco Packet Tracer Simulation for ASA firewall is shown in Figure 10.

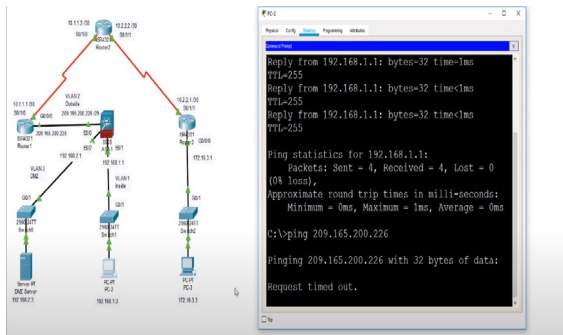


Fig. 10. Cisco Packet Tracer Simulation for ASA firewall

We setup VLAN 1 on the inside with the highest security level of 100 and VLAN 2 on the outside with the security level zero. This will provide the system interaction from lower security level and thereby securing system by any attacks outside from the network. On the figure we set up a demo system in which the ASA firewall will be connected to the router and the ping request from the lower security level is timed out when requested as shown in the figure.

VII. ARDUINO IOT CLOUD

The Web Editor platform on the Arduino IoT cloud was made use of for the programming of Arduino. This helps the user to have direct access to the cloud to make changes and reprogram it remotely. Arduino Create Agent client installed on the computer, which helps to communicate the hardware to the cloud and upload sketches. Things option on the IoT Arduino cloud dashboard is used to define the variables and to determine how they change according to the input. Necessary libraries are added during the coding for the different components added to the Arduino. IFTTT offers Webhook services that are used for email notification. Once the email is received, Gmail is set up to send a text notification to the desired number.

VIII. CONNECTION DIAGRAM

The overall circuit diagram for the proposed system is given in Figure 11. The pin diagram is connected in the circuit according to the program code in the IoT Arduino Cloud. The Arduino pins are soldered and placed on a

breadboard. The sensors and actuators are connected to the Arduino using connection wires.

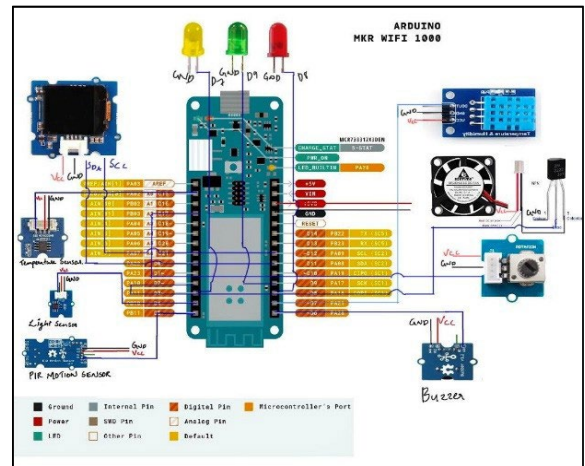


Fig. 11. Circuit Diagram and PIN Layout of MKR1000

IX. HARDWARE IMPLEMENTATION

Once the simulation is completed. The next step was to build the system and check its functionality. The goal was to implement the wireless sensors which can be connected to the Arduino to help the user easily place the sensors anywhere at their house with minimum to no wires. The final completed circuit of the system is shown in Figure 12 and the dashboard showing values detected and control buttons to monitor and control the home environment remotely as needed is shown in Figure 13 and Figure 14 shows text notifications of different events detected by the system.

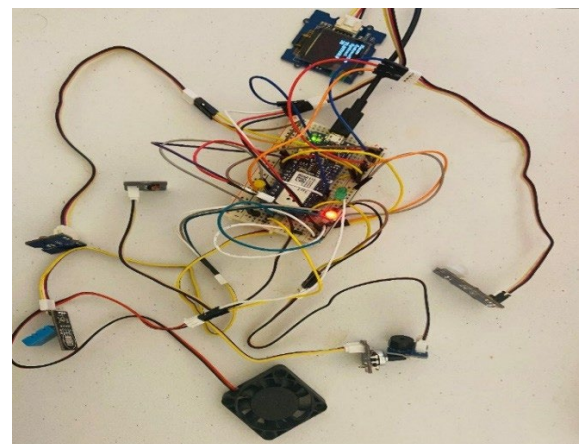


Fig. 12. Final Circuit

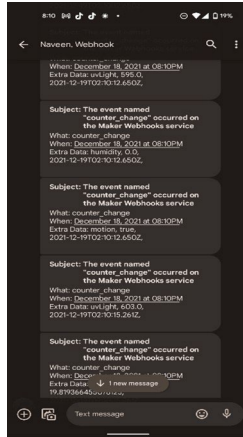


Fig. 13. IoT System Dashboard

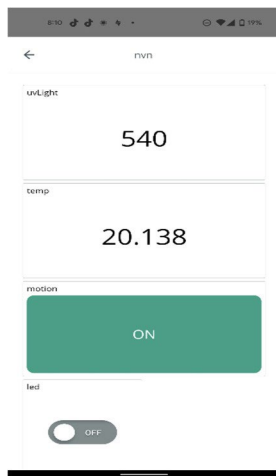


Fig. 14. IoT System Text Notification

X. CONCLUSION

This paper shows the simulation and implementation of an automated home to help the elderly using a low-powered home automation system that is user-accessible remotely. The projects enable the user to control appliances, read sensor values, and detect intruders from their mobile phones. The motion detector will detect any irregular motion in the house, the temperature sensor and humidity sensor will read and display the temperature and humidity values, respectively. The system is ready to use and needs to be connected to the home modem Wi-Fi. Once the necessary connections are completed, the home automation system is complete. Internet of things is a booming technology that brings all the technology together to make home automation possible in a simple and affordable way.

XI. REFERENCES

1. Dawood, M., S. Hossein, and N. Fatemeh, Design and implementation of a low-power active RFID for container tracking at 2.4 GHz frequency. *Advances in Internet of Things*, 2012.
2. Mandula, K., et al. Mobile based home automation using Internet of Things (IoT). in *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICT)*. 2015. IEEE.
3. Olteanu, A.-C., et al. Enabling mobile devices for home automation using ZigBee. in *2013 19th international conference on control systems and computer science*. 2013. IEEE.
4. Balaji, B., R. Priya, and R. Revathy. Domestic Automation System Using Internet of Things and Arduino. in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*. 2020. IEEE.
5. Fernández-Pacheco, A., S. Martin, and M. Castro. Implementation of an Arduino remote laboratory with raspberry Pi. in *2019 IEEE Global Engineering Education Conference (EDUCON)*. 2019. IEEE.
6. Evangelos A, K., T. Nikolaos D, and B. Anthony C, Integrating RFIDs and smart objects into a Unified Internet of Things architecture. *Advances in Internet of Things*, 2011.
7. Sulayman, I. I. A., Almalki, S. H., Soliman, M. S., & Dwairi, M. O. (2017, May). Designing and Implementation of Home Automation System Based on Remote Sensing Technique with Arduino Uno Microcontroller. In *2017 9th IEEE-GCC Conference and Exhibition (GCCCE)* (pp. 1-9). IEEE.
8. Ali, M., Nazim, Z., Haroon, M., Azeem, W., Javed, K., Tariq, M., & Hussain, A. (2020). An IoT based Approach for Efficient Home Automation with ThingSpeak system. *dgd*

IoT-Based Smart Hospital using Cisco Packet Tracer Analysis

Nesreen Alsbou and Dakota Price
 nalsbou@uco.edu, dprice16@uco.edu
 University of Central Oklahoma

Abstract – Today's hospitals are outfitted with cutting-edge technology in order to save as many lives as possible. However, the hospital is missing one thing: IoT-based connectivity between vehicles such as ambulances and surrounding hospitals. WiFi or cellular data will be used to connect to the internet. An Arduino MKR1000 will be used in this IoT system, with sensors attached to the OBD II connector. These sensors will be used by the smart hospital to communicate live data collected from the patient on board to many hospitals at the same time. EMTs could examine the damage and speed up the time it takes for doctors to start treating patients once they arrive.

Keywords – Internet of Things (IoT), Smart Hospital, Packet Tracer, VANET, Vehicle-to-vehicle, and Security.

I. INTRODUCTION

The Internet of Things (IoT) is a network of interconnected, internet-connected items that can gather and transmit data without the need for human interaction across a wireless network [1]. Businesses are currently inspired by IoT and the potential for increased revenue, lower operational costs, and improved efficiency. Regulatory compliance is also a driving force for businesses. IoT device installations give the data and insights required to optimize workflows, visualize use trends, automate operations, satisfy regulatory needs, and compete more effectively in a changing business environment, regardless of the reasons. Physical things can exchange and gather data with minimum human interaction thanks to low-cost computers, the cloud, big data, analytics, and mobile technologies. Digital systems can record, monitor, and alter each interaction between linked items in today's hyper connected environment.

Packet Tracer is a visual cross-platform tool developed by Cisco Systems that employs simulations to assist users understand computer networking fundamentals. It can run on Linux, Windows, Android, and macOS. Many sorts of network (including IoT) devices may be added or removed using a simple drag-and-drop approach. It also enables students to conduct simulations in a basic manner using a variety of programming languages, including Javascript, Python, and Blockly [2].

IoT-based technologies are all over the world right now in many different applications. Smartphones are predicted to expand at a rate of 3% per year, whereas industry and smart

homes are likely to increase at a rate of 20% per year. The rapid expansion of this sector may result in fewer house invasions, vehicle accidents, and the time required for physicians to review arriving patients at emergency departments. Ambulances are anticipated to arrive within 15 minutes on average in the United States to treat persons with life-threatening diseases or injuries. Before arriving at the hospital, this time will be extended by another 15 to 30 minutes. This time frame might be utilized to gather all data from blood pressure, heart rate, respiratory rate, and pulse oximetry in 15 to 30 minutes. In this paper, we will be using multiple sensors included in a BeagleBone development kit, which also includes connections. Some of the kit's sensors and actuators will be used as external components for the vehicle. The information collected will be sent to the hospital so the physicians can keep track of it and plan the next step prior to arrival at the hospital and going to the emergency room.

II. METHODOLOGY

A. Design of Network Topology

A Local Region Network (LAN) – a small collection of sensor-equipped devices connected by wired (e.g., Ethernet) or wireless communication inside a constrained area is the smart hospital's suggested network design. Small-scale network applications are best served by LANs. The hospital will be linked to this, as well as a small onboard gadget on an ambulance. The Cisco Packet Tracer simulation was developed to demonstrate how this process would operate. You can add or delete a variety of network (including IoT) devices using a simple drag-and-drop approach. The diagram below depicts how the smart hospital connects to the internet before connecting to the server. The server establishes a connection with the mobile tower, and then connect with the smartphone.

B. Security Protocol Configuration

A network security device will be fitted with ICMP for increased security in the smart hospital. This limits the computer's ability to connect with other devices to merely browsing the internet. An IP address firewall is the other step. These guard the smart hospital against outside threats and let it connect to other IP addresses on the network. Figure 1 shows the server cluster with the low-security firewall configured on the router, PC, and switch. [1,3]

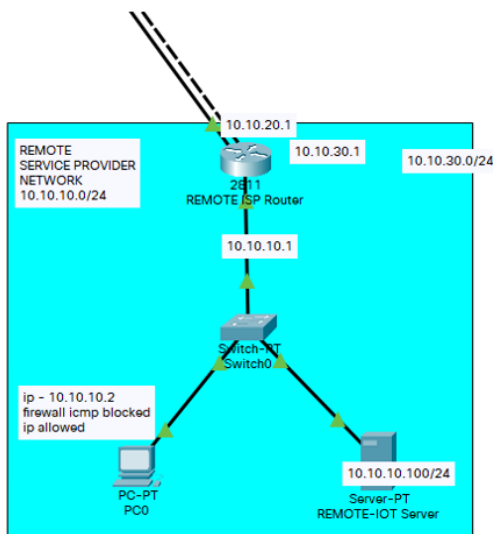


Fig. 1. Secure network topology.

C. Configuration of IoT Systems and Devices

The smart hospital is separated into three clusters for the different areas of the hospital. This allows to easily see the split up for clear analysis. The ambulance uses a smoke detector, window, LCD screen, alarm, fire detector, and water sprinkler as shown in Figure 2. These are all very important sensors and actuators that are essential in a hospital. The emergency entrance part of the smart hospital has a garage door, alarm, siren, smoke detector, trip wire, carbon monoxide monitor, and a water sprinkler. This can be seen in Figure 3 with all essential sensors and actuators an emergency entrance would include. The building portion of the smart hospital is shown in Figure 4 and it includes several monitors, a fan, alarm, trip wire, temperature sensor and light. These all correlate to each other, so when the trip wire is activated the light and fan activate. The microcontroller (MCU), which reads and transmits sensor data and outputs commands to IoT devices. All MCUs connect wirelessly to a central hospital gateway, which serves as a router to the PCs in the network, allowing users to view and control devices from the IoT monitor.

Three separate IoT systems were developed for the smart hospital design:

- Smart Ambulance: Located in the 'Vehicle Area' section in the smart hospital. The MCU takes the inputs from the different parts shown in Figure 2 and transmits them to the Gateway to be viewed on a connected PC in the hospital. These sensors read emergencies to activate the actuators and notify the IoT Monitor.
- Smart Emergency Entrance: This system warns users when the trip wire is activated to open the emergency door. This is also equipped with safety devices like fire sprinkler, monoxide monitor, and smoke sensor. This can be seen in Figure 3 and includes most essential sensors and actuators a garage would include.
- Smart Hospital: Includes a trip wire as input and light and fan as output. When motion is detected, the light and fan

turn on. This also includes an alarm, temperature sensor, and temperature monitor for viewing.

Figure 4 shows how the sensors and actuators are hard wired to the MCU, but some of these are connected wirelessly straight to the home gateway.

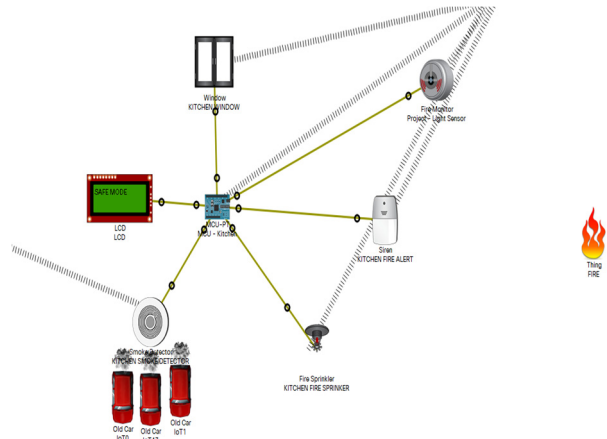


Fig. 2. WSN configuration of the ambulance in the smart hospital

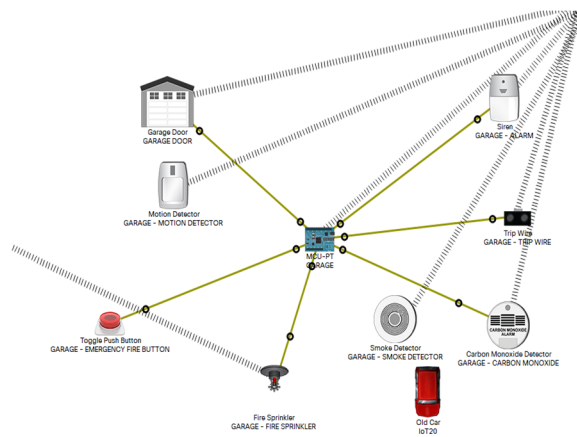


Fig. 3. WSN configuration of the emergency entrance of the smart hospital

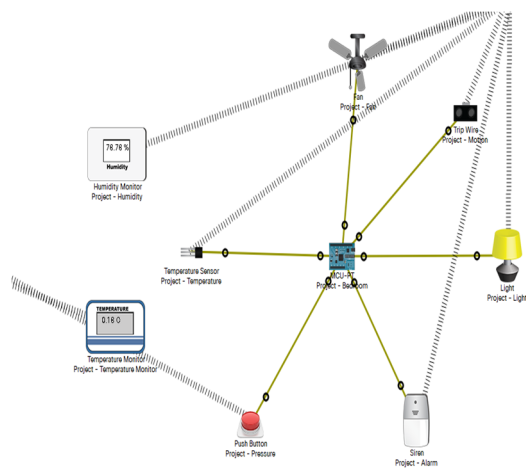


Fig. 4. WSN configuration of the building portion of the smart hospital

D. Packet Tracer Simulation and Testing

Cisco Packet Tracer is a simulation tool used for developing, testing, and troubleshooting networks [2, 3]. Packet Tracer allows users to program virtual IoT devices, sensors, and actuators, making the program a valuable tool for building and analyzing realistic IoT systems for smart cities. The network was first built and analyzed in Packet Tracer to ensure proper functioning [1].

E. Hardware Implementation and Testing

A hardware implementation of the smart hospital is being developed currently in the lab and will be tested. In addition to the functions design for the Packet Tracer simulation, this implementation includes the ability to send push notifications to user’s phone or send an email if sensor values pass user-determined thresholds.

III. PACKET TRACER SIMULATION AND ANALYSIS

A. Secure Network Communication

The ICMP Firewall is a protocol that network devices (e.g., routers) use to generate error messages when network issues are preventing IP packets from getting through. The Internet Control Message Protocol is one of the fundamental systems that make the internet work [3]. ICMP gives a little bit of feedback on communications when things go wrong. Figure 5 shows the command prompt pinging the IP address to show all connectivity is successful. The firewalls that are built are the ICMP and IP firewall. These allow the user to only browse, and not connect to other devices with the same IP address. This is a low security feature that allows users to continue safe browsing without disrupting other devices [1,3]. Figure 6 and Figure 7 show the Firewall Configuration and Firewall IP Configuration, respectively.

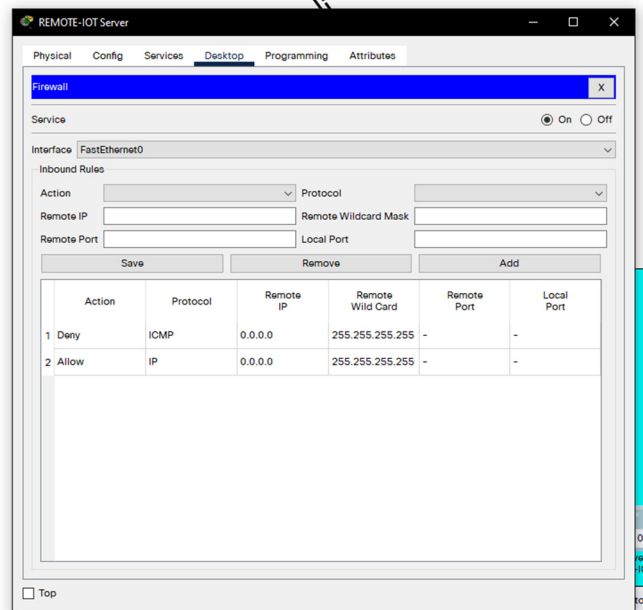


Fig. 6. Firewall Configuration

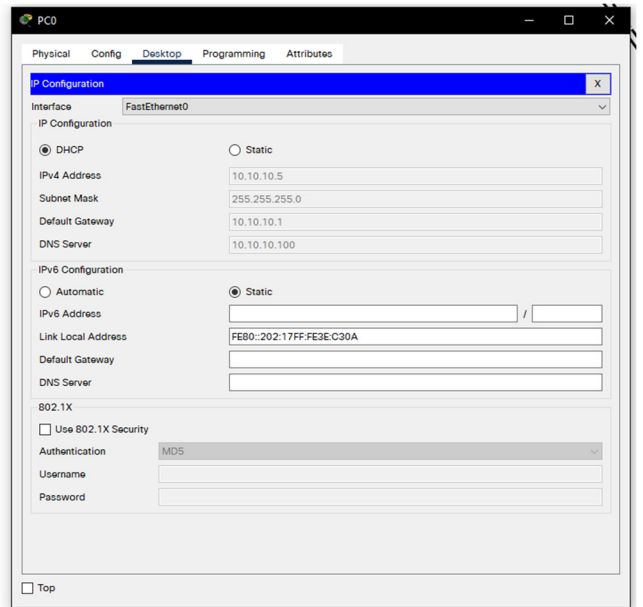


Fig. 7. Firewall IP Configuration

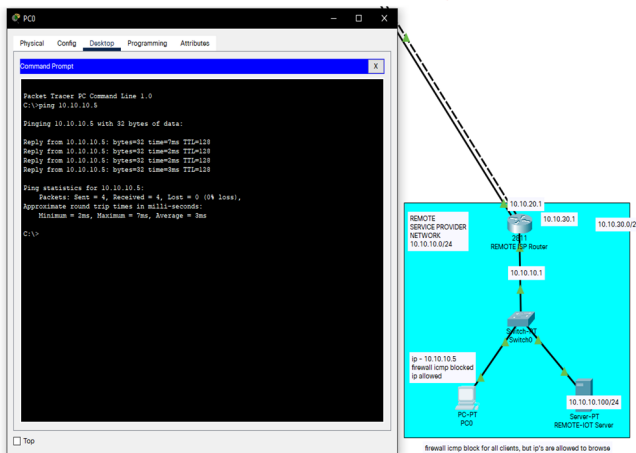


Fig. 5. Command Prompt pinging the IP address selected

B. IoT System Functions

- Smart Ambulance: Located in the ‘Vehicle Area’ section in smart hospital, this system consists of a LCD screen, window, fire monitor, alarm, fire sprinkler, and a smoke sensor. These sensors read emergencies to activate the actuators. This programming can be seen in Figure 8.

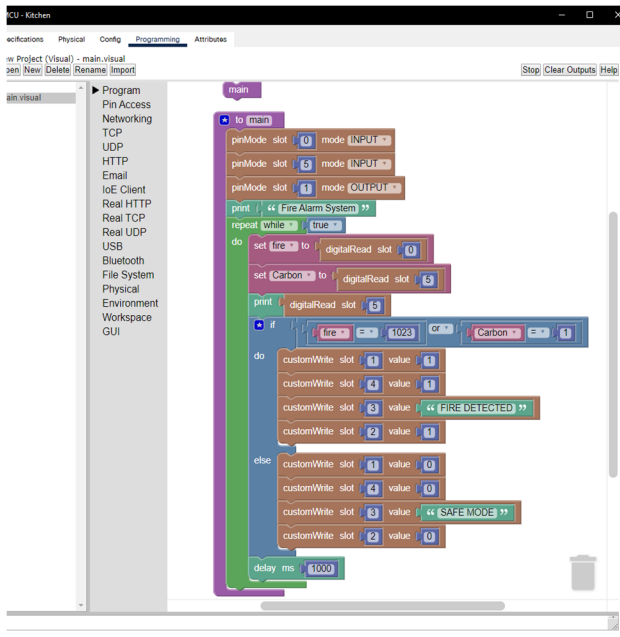


Fig. 8. Smart Ambulance – Blockly Programming

- **Smart Emergency Entrance:** This system warns users when the trip wire is activated to open the emergency door. This is also equipped with safety devices like fire sprinkler, monoxide monitor, and smoke sensor. The Blockly program for the sensors and actuators can be seen in Figure 9.

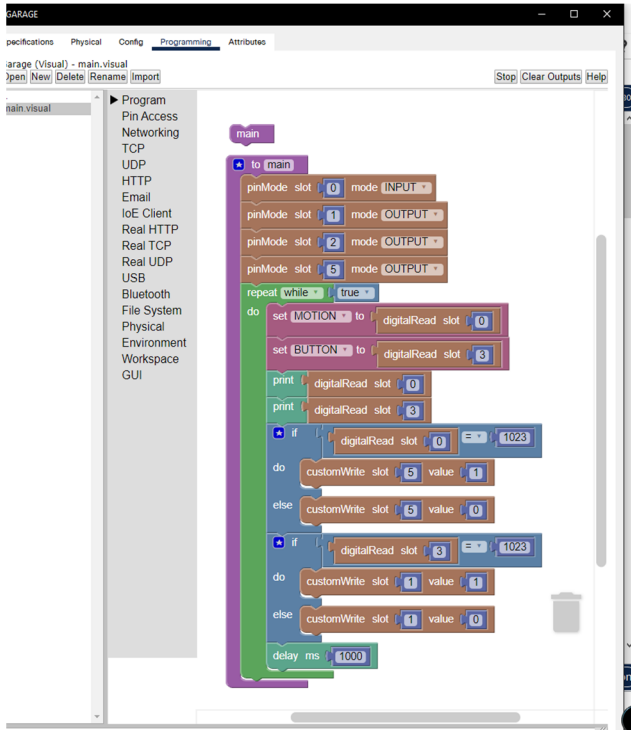


Fig. 9. Smart Emergency Entrance – Blockly Programmed

- **Smart Hospital:** Includes a trip wire as input and light and fan as output. When motion is detected, the light and fan turn on. This also includes an alarm, temperature

sensor, and temperature monitor for viewing. Figure 10 shows how the sensors and actuators are programmed with Blockly.

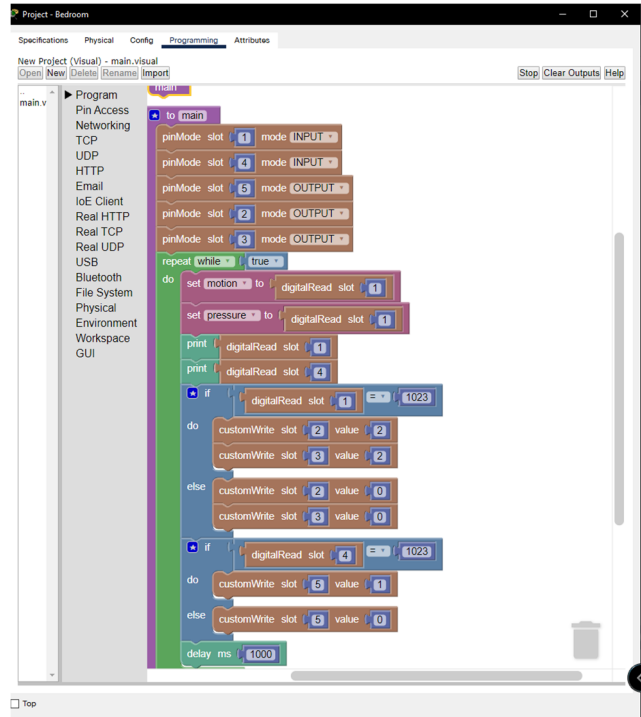


Fig. 10. Smart Bedroom – Blockly Programming

- **Remote-Access IoT Monitor**

Figure 11 shows all IoT devices that may be viewed or controlled remotely from a user PC’s IoT Monitor on Cisco Packet Tracer.

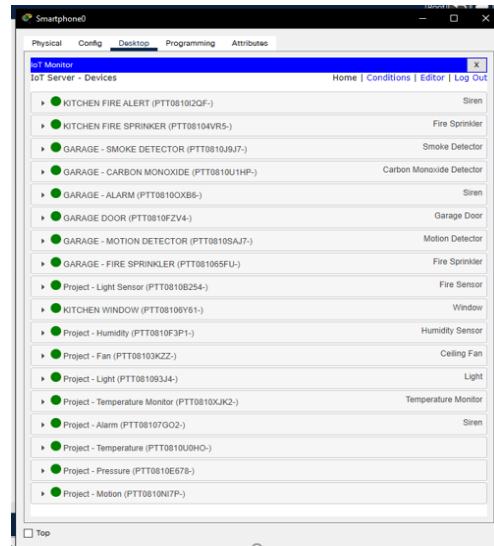


Fig. 11. All IoT devices accessible on the IoT Monitor.

IV. HARDWARE IMPLEMENTATION AND RESULTS

A. MCUs and Sensors

The smart hospital was created into three sections. This includes the smart ambulance, smart hospital, and smart

emergency entrance. These all connect with sensors and actuators and can be remotely controlled from the IoT Monitor.



Fig. 12. MKR1000 microcontroller with built-in WiFi Connectivity

The smart hospital system was implemented using an Arduino MKR1000 board and SparkFun OBD II. Sensors and actuators from the Grove Starter Kit for Seeed Studio BeagleBone® Green were used to monitor the various functions. These will be used to monitor the data remotely from the smart hospital.

B. *Arduino IoT Cloud*

The Arduino IoT cloud is a platform for creating Internet of Things (IoT) devices with Arduino and other microcontrollers. The Arduino IoT Cloud allows you to develop your own IoT devices using your existing Arduino skills. To make advantage of the cloud, you'll need to first learn some of its vocabulary. The "Thing" is the foundation of all Arduino IoT Cloud projects. The term "Thing" refers to a cloud-based virtual item. It is used to store variables and information about linked devices and networks in a safe manner. This needs a network connection details to the Thing. This will be the WiFi credentials, SSID, and password for most devices. Microcontrollers that link to the cloud are known as devices. It's likely that they are Arduino microcontrollers [4].

V. CONCLUSION

This paper provides a detailed explanation of an IoT-based smart hospital simulation. In this IoT system, an Arduino MKR1000 is utilized, with sensors connected to the OBD II port. In order to save as many lives as possible, today's hospitals are equipped with cutting-edge technology. One element is lacking from the hospital: IoT-based

communication between vehicles like ambulances and nearby hospitals. The internet will be accessed via WiFi or cellphone data. The smart hospital will employ these sensors to provide real-time patient data to other hospitals at the same time. EMTs might assess the situation and reduce the time it takes doctors to begin treating patients once they arrive. The internet will be accessed through WiFi and these sensors will be used by the smart hospital. The sensors are directly connected to the MKR1000. The MKR1000 has been calibrated to operate with all sensor data. The MKR1000 uses WiFi to send the data to the internet. This sends the data to the Arduino IoT Cloud through the internet. The data is subsequently transferred to the cloud, where IoT Analytics processes it. Before the data is uploaded to the Arduino IoT Remote, IoT Analytics examines it. This is the App component of the IoT system. Because it focuses just on the "Thing's" readings rather than all the Serial.prints and other units, the Arduino Dashboard is included in the IoT Analytics. This is a way of processing data so that graphs and gauges may be shown to monitor the patient remotely prior to arriving at the hospital.

REFERENCES

1. What is connected vehicle technology and what are the use cases? Digi International. (n.d.). <https://www.digi.com/blog/post/what-is-connected-vehicle-technology-and-use-cases>
2. Shouran, Zaied & Ashari, Ahmad & Priyambodo, Tri. (2019). Internet of Things (IoT) of Smart Home: Privacy and Security. *International Journal of Computer Applications*. 182. 3-8. 10.5120/ijca2019918450.
3. M. W., & says:, S. C. (2021, May 6). *What is ICMP?* Comparitech. <https://www.comparitech.com/net-admin/what-is-icmp/>
4. Kumar, R. (2022, January 18). *What is Arduino IOT Cloud*. iWheels. <https://iwheels.co/blog/arduino-iot-cloud/>
5. Wikimedia Foundation. (2022, March 30). *Packet tracer*. https://en.wikipedia.org/wiki/Packet_Tracer

Simulators and Testbeds for IIoT Development and Validation

Nicholas Jeffrey
Department of Computer Science
University of Oviedo
 Oviedo, Spain
 0000-0001-6384-5746

Qing Tan
Faculty of Science and Technology
Athabasca University
 Athabasca, Canada
 0000-0002-6447-2133

José R. Villar
Department of Computer Science
University of Oviedo
 Oviedo, Spain
 0000-0001-6024-9527

Abstract— The Internet of Things (IoT) and Industrial Internet of Things (IIoT) are integrated systems that combine software and physical components. These integrated systems have experienced rapid growth over the past decade, from fields as disparate as telemedicine, smart manufacturing, autonomous vehicles, industrial control systems, smart power grids, remote laboratory environments, and many more. As IIoT becomes increasingly ubiquitous throughout supply chains, malicious attacks by hostile actors have grown exponentially in recent years. Attacks on critical national infrastructure (CNI) such as oil pipelines or electrical power grids have become commonplace, as increased connectivity to the public internet increases the attack surface of IIoT. This paper presents a review of the current academic literature describing the state of the art of the use of simulated environments and testbeds in the system development life cycle for IIoT environments, with a focus on the use of simulators for rapid iteration of security validation tests during the development process. As a new contribution, this paper also identifies outstanding challenges in the field, and maps selected challenges to potential solutions and/or opportunities for further research.

Keywords— *IoT Simulator, IIoT Testbed, Cyber-Physical Systems Design, Industry 4.0, IoT Cyber Range*

I. INTRODUCTION

The Industrial Internet of Things (IIoT) is comprised of integrated systems that combine software and physical components. IIoT systems have experienced exponential growth over the past decade, from fields as disparate as telemedicine, smart manufacturing, autonomous vehicles, industrial control systems, smart power grids, remote laboratory environments, and many more.

This research has been funded by the SUDOE Interreg Program -grant INUNDATIO-, by the Spanish Ministry of Economics and Industry, grant PID2020-112726RB-I00, by the Spanish Research Agency (AEI, Spain) under grant agreement RED2018-102312-T (IA-Biomed), and by the Ministry of Science and Innovation under CERVERA Excellence Network project CER-20211003 (IBERUS) and Missions Science and Innovation project MIG-20211008 (INMERBOT). Also, by Principado de Asturias, grant SV-PA-21-AYUD/2021/50994.

Academia tends to use the term Cyber-Physical System (CPS), while industry tends to use Internet of Things (IoT) for consumer-grade devices, and IIoT for industrial control systems (manufacturing, process control, etc.).

The consumer-focused IoT industry was born in an age when ubiquitous connectivity to an increasingly hostile Internet was assumed, which helped drive adoption of standardized communication protocols around TCP/IP, with integrated authentication and encryption functionality designed for the lightweight messaging protocols of devices assumed to have constrained processing power, battery life, and unreliable network connectivity.

However, the IIoT systems that are driving Industry 4.0 have product lifecycles measured in years or decades, and the historical design assumptions of operating in a fully trusted and air-gapped isolated environment, the traditional Industrial Control Systems (ICS) are much slower to adopt new technologies than their more agile counterparts in consumer-focused IoT devices that have product lifecycles measured in months to a few years.

This disparity in product lifecycles, as well as design considerations for life safety issues in IIoT design, has resulted in the consumer-focused IoT industry diverging from their IIoT counterparts. This paper focuses on the issues related to the design lifecycle of IIoT systems, with a focus on the use of simulators and testbed environments as a tool for non-destructive testing and validation of life safety and information security issues. In a typical corporate computer network, it is common to have a production environment that runs the organization, plus a parallel dev/test environment for testing changes and upgrades. However, dev/test environments are less common for IIoT systems, due to the expense and complexity of maintaining a parallel version of all the physical components in the IIoT system [1].

This fundamental difference in Information Technology (IT) networks and Operational Technology (OT) networks was largely ignored for decades, as OT networks were typically air-gapped from IT networks for reliability and availability reasons. However, Industry 4.0 has driven the merging of IT and OT networks, also commonly known as CPS or IIoT, exposing fundamental shortcomings in safely interconnecting IT and OT networks.

To address this issue, a number of simulation platforms [2-10] that have been developed specifically for researchers and practitioners to be able to test different facets of the IIoT system in an isolated non-production environment, which alleviates significant expense and life safety issues that could result from testing on a live system.

A common use case for a simulated IIoT systems environment is vulnerability assessments, which can involve potentially disruptive activities such as input fuzzing, man-in-the-middle (MitM) attacks, DDoS, false data injection, etc. Such activities performed on a live CPS may have significant financial and/or life safety consequences, which makes simulated environments attractive for testing [11]. Another common use case is for non-disruptive user training in a simulated environment, either on a parallel virtualized IIoT, or through the use of AR/VR (Augmented Reality / Virtual Reality) simulations.

Simulated or test environments are commonplace in software-only environments, due to widespread availability of server virtualization technologies than can inexpensively simulate large networks and applications. However, since IIoT environments are often customized and unique combinations of cyber and physical devices, a universal testbed or simulator has proved elusive, due to the myriad of hardware sensors and actuators in the marketplace.

II. LITERATURE ANALYSIS / RELATED WORK

Two of the commonly recurring themes in the available literature are the extreme heterogeneity of IIoT systems, and the lack of a formalized design methodology, leading to “reinventing the wheel”, which adds time and expense to the system development process. The use of simulators or testbeds during the IIoT development process can help minimize these costs, leading to shorter time to market and more robust designs.

Poudel et al [12] developed a simulator for smart power grids that focuses on testing overall grid reliability in the face of physical attacks on power lines and electrical transformers in a smart grid. Smart power grids are designed to be fault tolerant to environmental hazards such as lightning strikes, trees falling on power lines, vehicles striking transformers, etc. However, real-world testing on a power grid has significant economic and life safety consequences, making a simulated environment very attractive. There are various hardware vendors that serve the power grid industry with their own unique and proprietary sensors and actuators, so this simulator is valid only for testing very specific hardware environments.

The RITICS (Research Institute in Trustworthy Inter-connected Cyber-physical Systems) organization [13] is perhaps the leader in academic research in this area, a consortium of universities and industry partners that collaboratively develop simulators and testbeds for a broad range of CPS, with a focus on Industrial Control Networks (ICS), and their interactions with external corporate networks and the public internet.

RITICS has a particular focus on developing simulated environments with a high level of fidelity to the real-world system being emulated, as well as repeatability and accuracy of

test runs. This high degree of fidelity to the real-world counterpart leads to higher levels of confidence in the simulator as being reflective of the real world, which is often lacking in other implementations.

Wlazo et al [14] discuss the challenges of detecting man-in-the-middle (MitM) attacks in IIoT smart grid environments, which can allow malicious actors to create false data injection (FDI) and false command injection (FCI) to compromise the stability of the power grid. This research focus primarily on network traffic, which allows a large-scale emulated smart grid to run on virtualized hardware at low cost with high observability. While the use of an easily observable emulated environment did allow the researchers to use an off-the-shelf Intrusion Detection System (IDS) to detect MitM attacks, it suffered from poor accuracy in real-world environments, so has met with limited interest from industry.

Ani et al [15] developed a mapping framework for design considerations for building credible testbeds for IIoT / ICS environments, with specific focus on how effective design considerations for testbeds make possible the modeling and simulation of cybersecurity testing and physically destructive testing that is infeasible for financial or life safety issues on real-world IIoT systems. A key issue in testbed development is the credibility or confidence level that the testbed environment is an accurate simulation of the real environment, which is affected by trade-offs between fidelity of the testbed design and implementation costs, as well as the observability of the testbed environment during simulation tests.

Negi et al [16] performed case studies on several high-profile cybersecurity incidents that compromised IIoT environments running CNI, and how the use of testbeds to perform non-disruptive vulnerability assessments could have been used to mitigate the risks that lead to system compromises. Operators of CNI or IIoT systems are hesitant to perform Vulnerability Assessment and/or Penetration Testing (VAPT) on their own systems due to fear of disruption performance impact. Continuous availability has long been the primary goal of the operators of IIoT, which leads to a general avoidance of VAPT on live IIoT systems. Negi et al developed a testbed for a smart power grid comprised of a wide variety of heterogeneous physical equipment, with minimal use of virtualization or simulation to increase real-world fidelity, although on a smaller scale than a typical smart power grid. Through extensive VAPT, Negi et al were able to discover multiple communication protocol vulnerabilities that could be used to perform false data injection attacks on the power grid. These risks were then able to be mitigated on the live IIoT systems without disruptive testing.

Green et al [17] propose a high level model that defines a baseline for IIoT / ICS testbed development, utilizing the Purdue Enterprise Reference Architecture [18] to segment the IT/OT networks into four zones. The Safety Zone is for managing safety functions, the Control Zone for operational processes, the DeMilitarised Zone is a buffer between the Control Zone and Enterprise Zone, and the Enterprise Zone is the corporate IT network. This model encourages formal design specifications that forbid inputs from a lower-security zone being used in a higher-security zone, allowing a formal proof of the design, and eases real-world administration through the use of zone-based

firewalls to control access. This model has proven popular with academics due to its simplicity in generating formal proofs, but has received limited adoption in industry due to lack of fidelity with real-world environments.

III. SELECTED CHALLENGES

The field of IIoT is still quite young, but grew out of older industries such as cybernetics, industrial process control, and control logic. The rapid adoption of Industry 4.0 practices has grown in leaps and bounds over the past few decades, largely mirroring advancements in microprocessor technology, and the increased availability of high-speed wired and wireless networks. Due to rapid rate of change, IIoT has a number of outstanding challenges, a selection of which are described below.

Rapid obsolescence: The more modern consumer-focused IoT industry has been quicker to adopt a zero-trust model of information security, accepting the reality that they operate in a potentially hostile network environment, and embedding strong authentication and encryption protocols by default. Unfortunately, the rapid advancement of IoT means that product lifecycles are very short [19], making devices become obsolete quickly, leaving many “orphaned” devices without ongoing vendor support or upgrades to counter new security threats. While some vendors have included functionality for receiving trusted over-the-air updates to counter newly discovered threats, there are many IoT devices that entirely lack any sort of update functionality, leaving them permanently vulnerable to emerging threats.

Low quality research datasets: Another significant challenge is the availability of accurate and meaningful datasets from real-world environments. Conti et al [20] discuss the ongoing research challenge of obtaining useful datasets from IIoT systems for the testing of research questions. Due to privacy concerns, private industry and governmental organizations are reluctant to provide datasets from their own IIoT environments due to fears of inadvertent vulnerability disclosure and/or loss of competitive advantage in the marketplace. Most of the datasets currently being used for testbed validation have been generated by researchers in simulated environments, so have varying degrees of fidelity to real-world environments, which leads to less than optimal research outcomes. The use of honeynets [21-24] is a specialized use case for testbeds that are particularly useful for generating real-world data for research datasets for intrusion detection and AI modeling of threat detection.

Roblez-Durazno et al [25] recognize the lack of availability of accurate research datasets, and propose a hybrid testbed architecture that combines fully simulated testbeds with small-scale physical components to achieve greater fidelity to real-world environments as a means to obtain higher quality datasets for research. A fully virtualized or simulated testbed is attractive to academic researchers due to the low cost and ease of acquisition, but lacks real-world fidelity due to timing and performance inconsistencies introduced by the virtualization layer. Conversely, fully physical testbeds have greater fidelity to real-world environments, but are prohibitively expensive to

build at scale. By combining virtual and scaled-down physical testbeds into a hybrid model, more accurate testing of physical conditions such as vibration and temperature extremes can be more accurately tested, while the communication layer can take advantage of virtualization to minimize costs. A hybrid testbed will always be a trade-off between cost and real-world fidelity, which makes the development of an accurate model an ongoing challenge for researchers. Although wide variation exists across different testbeds, a common hybrid model will virtualize the sensor and actuator components to allow for flexible manipulation during experimentation, while using physical components for items such as Programmable Logic Controllers (PLC) and Human-Machine Interfaces (HMI) to achieve higher fidelity with real-world systems. Virtualization of the sensors and actuators in a testbed is particularly attractive to researchers, as it simplifies cybersecurity testing and validation early in the IIoT development life cycle, by providing visibility into network-based attacks such as ARP spoofing, MiTM, input fuzzing, password spraying, denial of service, etc. As OT networks are increasingly connected to IT networks, these network-based attacks have become more common, making cybersecurity testing increasingly important in the early phases of the IIoT development cycle.

Duplication of effort: Due to the extreme heterogeneity of IIoT systems, the case studies in this area are particularly interesting due to their extreme differences. The case studies [2-10] cover systems ranging from water treatment plants, oil pipelines, autonomous vehicles, power grids, and more. While this extreme heterogeneity makes the field ripe with different research opportunities, the lack of a common design framework or standardized communication protocols makes it difficult to leverage prior research and industry expertise to drive forward the current state of the art. In other words, the extreme heterogeneity tends to result in duplication of effort across both academia and industry, which slows advancement in the field.

Immature design models: Eckhart et al [26] recognizes that the software engineering discipline has a well-defined methodology for including security requirements in the earliest stages of the system development lifecycle, and that due to historical assumptions about operating on an isolated and trusted network, security was often ignored in ICS / SCADA networks. Eckhart et al propose a novel method of adopting similar techniques for security designs for CPS, referred to as Security Development Lifecycle for Cyber-Physical Production Systems (SDL-CPPS).

Rehman and Gruhn [27] build on the concepts of the SDL-CPPS, by developing a security requirements engineering framework for CPS that is designed to overcome the challenges of the highly heterogeneous nature of CPS with a unified general model, recognizing that fixing security issues in later stages of the system development lifecycle can increase costs exponentially.

Lack of consensus on best practices: Popisil et al [28] build on the work of Roblez-Durazno by developing a framework of best practices for the development of physical, simulated, virtual, and hybrid testbeds. Recognizing the extreme heterogeneity of IIoT environments, Popisil et al recognize the need for unique and custom testbeds to ensure fidelity to the

real-world system being tested, and therefore recommends against a generalized or universal model for accuracy reasons. This makes the datasets generated highly accurate for researchers to perform functional tests, intrusion detection, and training of AI models on that specific environment, but those datasets are of limited use to the greater research community due to their specialized nature. Further distinction is made by defining four categories of testbed, each of which has different intended audiences and objectives. The first testbed category is for cybersecurity analysis, which emphasizes network virtualization to simplify intrusion detection testing and forensic analysis of network-based attacks. The second category is testbeds designed for educational purposes, with emphasis on training students or operators in the day-to-day operations of the real-world system. The third category is testbeds designed for functional and performance testing, so will necessarily have more physical components than the other categories to ensure real-world fidelity, although at a scaled-down size. The final category is testbeds designed for standards development, with an emphasis on research and development of industry standards. Each category is sufficiently differentiated to provide value to its intended use case, making the added complexity of maintaining multiple custom testbed environments a worthwhile endeavour for both academia and industry.

Lack of standard reference architecture: Craggs et al [18] take the opposite approach of Popisil, aiming to develop a generalized reference architecture for IIoT testbeds, focusing on a highly virtualized environment used for cybersecurity testing. Little attention is given to the testbed fidelity with real-world systems, with the emphasis on development of a standardized approach for performing vulnerability analysis, generation of datasets for AI modeling, and big data analytics. While this approach provides value specifically for improving intrusion detection capabilities in IIoT, the lack of focus on real-world fidelity will limit industry adoption outside of the research community.

IV. CONCLUSIONS

As a relatively young (since 2006) field of study, the system development life cycle for IIoT environments is still maturing, with limited consensus for industry best practice or formal development methodologies. Due to historical design requirements inherited from legacy industrial control systems, IIoT environments typically place high priority on reliability and availability, but will often intentionally exclude even the most basic cybersecurity protections, due to historical assumptions of the IIoT environment being located on a fully isolated and trusted network. As IT and OT networks are joined together, these historical assumptions no longer hold true, resulting in cybersecurity measures being added to IIoT environments too late in the system development process, resulting in higher costs and lower effectiveness, resulting in the increasingly common compromise of IIoT environments from malicious actors.

Improved simulators: The use of simulated environments or testbeds for IIoT systems shows much promise for deeper integration of cybersecurity validation early in the system development life cycle, but is still hampered by the relative immaturity of IIoT testbeds, particularly relating to the fidelity

of a testbed environment when compared to the real-world system. Improvements in virtualization technology have helped simulated environments gain closer fidelity to their real-world counterparts, but the extreme heterogeneity of IIoT environments had made accurate simulation an elusive target.

Regulatory collaboration: The increasing frequency of state-sponsored attacks on CNI have created an urgency to developing more robust IIoT environments, leading to multiple governments to enact legislative requirements for minimal acceptable cybersecurity standards for CNI. The most widely recognized is the Cybersecurity in the EU Common Security and Defence Policy (CSDP) [29]. All EU member states agreed in principle to this common cybersecurity standard in 2016, but as of 2022, the process of integration into each member state sovereign legislation was still underway. This has resulted in inconsistent application of cybersecurity postures for CNI across national borders due to bureaucratic friction. Outside of EU member states, the UK and NATO have also agreed in principle to the CSDP, but have not enshrined those policy standards into law. While this cooperation is admirable in the face of shared threats from state-sponsored adversaries, there is considerable opportunity for accelerating the timeline for implementation.

Academia / industry collaboration: Opportunities for further development include increased collaboration between academia and industry, towards the development of a more formalized methodology for the end-to-end system development life cycle of IIoT, with a focus on increased use of standardized communication protocols and observability metrics to encourage a common testbed environment that can provide high fidelity with a real-world IIoT, allowing for more reliable testing and validation earlier in the system development life cycle.

Higher education: Recognizing a skills gap in cybersecurity skills for CNI / IIoT environments, Kamsanrong et al [30] recommends development of curricula for higher education to prepare students for the cybersecurity needs of industry controlling critical infrastructure. With stakeholders throughout the EU, Erasmus+ is an academic consortium performing fundamental research into solutions to the cyber skills gap for CNI, in collaboration with government and industry. Testbeds are a significant part of driving education to address this skills gap, and Erasmus+ develops both physical and virtual testbeds with a design objective of pedagogical instruction and curriculum development. This is one of the few organizations focused directly on raising the minimum acceptable standard of CNI security through educational development, and shows great promise for future development.

REFERENCES

- [1] K. Tam, K. Moara-Nkwe, K.D. Jones (2021). "The use of cyber ranges in the maritime context: Assessing maritime-cyber risks, raising awareness, and providing training", *Maritime Technology and Research*, volume 3 issue 1, pp 16-30, <https://doi.org/10.33175/mtr.2021.241410>
- [2] R. Czekster, C. Morisset, J. Clark, S. Soudjani, C. Patsios, P. Davison (2021). "Systematic review of features for co-simulating security incidents in Cyber-Physical Systems", *Security and Privacy*, 2021; 4:e150. <https://doi.org/10.1002/spy2.150>
- [3] S. Athalye, C. M. Ahmed, J. Zhou (2020). "A Tale of Two Testbeds: A Comparative Study of Attack Detection Techniques in CPS Critical

- Information Infrastructures Security", 15th International Conference, CRITIS 2020, pp 17-30 https://doi.org/10.1007/978-3-030-58295-1_2
- [4] S. Lee, S. Lee, H. Yoo et al (2018). "Design and implementation of cybersecurity testbed for industrial IoT systems", *The Journal of Supercomputing* 74, 4506-4520 (2018) <https://doi.org/10.1007/s11227-017-2219-z>
- [5] G. Ravikumar, B. Hyder and M. Govindarasu (2020). "Next-Generation CPS Testbed-based Grid Exercise - Synthetic Grid, Attack, and Defense Modeling", 2020 Resilience Week (RWS), 2020, pp. 92-98, <https://doi.org/10.1109/RWS50334.2020.9241284>
- [6] M. Wu, J. Song, L. Wang, L. Lin, N. Aurelle, Y. Liu, B. Ding, Z. Song, Y.B. Moon (2018). "Establishment of intrusion detection testbed for CyberManufacturing systems", *Procedia Manufacturing*, Volume 26, 2018, Pages 1053-1064, ISSN 2351-9789, <https://doi.org/10.1016/j.promfg.2018.07.142>
- [7] B. Potteiger, W. Emfinger, H. Neema, X. Koutosukos, C. Tang and K. Stoffer (2017). "Evaluating the effects of cyber-attacks on cyber physical systems using a hardware-in-the-loop simulation testbed", 2017 Resilience Week (RWS), 2017, pp. 177-183, <https://doi.org/10.1109/RWEEK.2017.8088669>
- [8] T. Yardley, D.M. Nicol (2017). "Cyber-Physical Experimentation Environment for RADICS (CEER)", Information Trust Institute, <https://iti.illinois.edu/research/energy-systems/cyber-physical-experimentation-environment-radics-ceer>, https://eprints.lancs.ac.uk/id/eprint/139028/1/Open_Testbeds_deliverable_final.pdf
- [9] G. Ravikumar, A. Singh, J. R. Babu, A. Moataz A and M. Govindarasu (2020). "D-IDS for Cyber-Physical DER Modbus System - Architecture, Modeling, Testbed-based Evaluation", 2020 Resilience Week (RWS), 2020, pp. 153-159 <https://doi.org/10.1109/RWS50334.2020.9241259>
- [10] J. Gardiner, B. Craggs, B. Green, A. Rashid (2019). "Oops I Did it Again: Further Adventures in the Land of ICS Security Testbeds", CPS-SPC'19: Proceedings of the ACM Workshop on Cyber-Physical Systems Security & Privacy, November 2019, Pages 75-86 <https://doi.org/10.1145/3338499.3357355>
- [11] D.T. Sullivan, E.J.M. Colbert, B.E. Hoffman, A. Kott (2018). "Best Practices for Designing and Conducting Cyber-Physical-System War Games", *Journal of Information Warfare*, vol. 17, no. 3, Peregrine Technical Solutions, 2018, pp. 92-105, <https://www.jstor.org/stable/26633168>
- [12] S. Poudel, Z. Ni, N. Malla (2017). "Real-time cyber physical system testbed for power system security and control", *Electrical Power and Energy Systems*, Volume 90, pp 123-133, 2017, <https://doi.org/10.1016/j.ijepes.2017.01.016>
- [13] C. Hankin, D. Chana, B. Green, R. Khan, P. Popov, A. Rashid, S. Sezer (2018). "Open Testbeds for CNI", Imperial College London,
- [14] P. Wlazlo, A. Sahu, Z. Mao, H. Huang, A. Elisa, P. Goulart, K.R. Davis, S. Zonouz (2021). "Man-in-The-Middle Attacks and Defense in a Power System Cyber-Physical Testbed", *IET Cyber-Physical Systems: Theory and Applications* 2021, <https://doi.org/10.1049/cps2.12014>
- [15] U. Ani, J. Watson, B. Green, B. Craggs, J. Nurse (2019). "Design Considerations for Building Credible Security Testbeds: A Systematic Study of Industrial Control System Use Cases", *ArXiv*, <https://arxiv.org/pdf/1911.01471.pdf>
- [16] R. Negi, P. Kumar, S. Ghosh, S. K. Shukla and A. Gahlot (2019). "Vulnerability Assessment and Mitigation for Industrial Critical Infrastructures with Cyber Physical Test Bed", 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019, pp. 145-152, <https://doi.org/10.1109/ICPHYS.2019.8780291>
- [17] B. Green et al (2020). "ICS Testbed Tetris: Practical Building Blocks Towards a Cyber Security Resource", CSET @ USENIX Security Symposium (2020), <https://www.usenix.org/conference/cset20/presentation/green>
- [18] B. Craggs, A. Rashid, C. Hankin, R. Antrobus, O. Serban, N. Thapen (2019). A Reference Architecture for IIoT and Industrial Control Systems Testbeds. *Living in the Internet of Things (IoT 2019)*, <https://doi.org/10.1049/cp.2019.0169>
- [19] B. Jeannotte, A. Tekeoglu (2019). "Artrias: IoT Security Testing Framework", 2019 26th International Conference on Telecommunications (ICT), <https://doi.org/10.1109/ICT.2019.8798846>
- [20] M. Conti, D. Donadel and F. Turrin (2021). "A Survey on Industrial Control System Testbeds and Datasets for Security Research", *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2248-2294, Fourthquarter 2021, <https://doi.org/10.1109/COMST.2021.3094360>
- [21] M. Haney (2019). "Leveraging Cyber-Physical System Honey Pots to Enhance Threat Intelligence", *Critical Infrastructure Protection XIII. ICCIP 2019. IFIP Advances in Information and Communication Technology*, vol 570. Springer, Cham, https://doi.org/10.1007/978-3-030-34647-8_11
- [22] M. A. Hakim, H. Aksu, A. S. Uluagac and K. Akkaya (2018). "U-PoT: A Honey Pot Framework for UPnP-Based IoT Devices", 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC), 2018, pp. 1-8 <https://doi.org/10.1109/IPCCC.2018.8711321>
- [23] S.M. Wade (2011). "SCADA Honeynets: The attractiveness of honeypots as critical infrastructure security tools for the detection and analysis of advanced threats", *Proceedings of 13th IFIP WG 11.10 International Conference, ICCP 2019* <https://dr.lib.iastate.edu/server/api/core/bitstreams/a9069931-a3db-4726-824e-493e6255520f/content>
- [24] J. Franco, A. Aris, B. Canberk and A. S. Uluagac (2021). "A Survey of Honey Pots and Honeynets for Internet of Things, Industrial Internet of Things, and Cyber-Physical Systems", *IEEE Communications Surveys & Tutorials* <https://doi.org/10.1109/COMST.2021.3106669>
- [25] A. Robles-Durazno, N. Moradpoor, J. McWhinnie, G. Russell, J. Porcel-Bustamante (2021). "Implementation and Evaluation of Physical, Hybrid, and Virtual Testbeds for Cybersecurity Analysis of Industrial Control Systems", *Symmetry* 2021, 13, 519. <https://doi.org/10.3390/sym13030519>
- [26] M. Eckhart, A. Ekelhart, A. Lüder, S. Biffl, E. Weippl (2019). "Security Development Lifecycle for Cyber-Physical Production Systems", *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, 2019, pp. 3004-3011, <https://doi.org/10.1109/IECON.2019.8927590>
- [27] S. Rehman, V. Gruhn (2018). "An Effective Security Requirements Engineering Framework for Cyber-Physical Systems", *Technologies* 2018, 6, 65, <https://doi.org/10.3390/technologies6030065>
- [28] O. Pospisil, P. Blazek, K. Kuchar, R. Fujdiak, J. Misurec (2021). "Application Perspective on Cybersecurity Testbed for Industrial Control Systems", *Sensors* 2021, 21, 8119. <https://doi.org/10.3390/s21238119>
- [29] P. Trimintzios, G. Chatzichristos, S. Portesi, P. Drogkaris, L. Palkmets, D. Liveri, Andrea Dufkova.(2017). "Cybersecurity in the EU Common Security and Defence Policy(CSDP)", *European Parliament*, [https://www.europarl.europa.eu/RegData/etudes/STUD/2017/603175/EP_RS_STU\(2017\)603175_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2017/603175/EP_RS_STU(2017)603175_EN.pdf)
- [30] J. Kamsamrong, B. Siemers, S. Attarha, S. Lehnhoff, M. Valliou, A. Romanovs, J. Bikovska et al. (2022). "State of the Art, Trends and Skill-gaps in Cybersecurity in Smart Grids.", *Erasmus+ Strategic Partnership Project*, https://www.uwasa.fi/sites/default/files/2022-04/State%20of%20the%20Art%2C%20Trends%20and%20Skill-gaps%20in%20Cybersecurity%20in%20Smart%20Grids%20CCRSRG%20Project_0.pdf

Image Captioning- Bangladesh's Heritage Perspective Using Deep Learning

Sarowar Alam, Khalidul Islam, Nishat Sharmila, Ziaur Rahman Sovon,
Rashedur M. Rahman

*Department of Electrical and Computer Engineering
North South University*

Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh

{sarowar.alam, khalidul.islam, nishat.sharmila, ziaur.sovon, rashedur.rahman}@northsouth.edu

Abstract—Image captioning aims to make a textual short explanation of a given image. Despite the fact that it looks to be a straightforward task for human being, it is difficult for computers since it involves the ability to analyze the image and provide a human-like description. Encoder-decoder architectures have recently reached advanced outcomes in the form of picture captioning. With some existing datasets, e.g., Flickr_data, Flickr8k_token.txt, and heritage dataset, we build our model that can create captions from the images related to Bangladeshi culture, tradition and historical places. Bangladesh is enriched with great culture; many heritage places and cultural programs that attract travelers to visit our country. We try to relate our culture, place, and food, together with machine learning techniques by appropriate captioning and spread over our cultural strengths through proper captioning. Our image captioning tool can be very helpful for travel lovers who want to know more about Bangladesh.

Keywords—Encoder-decoder, datasets, LSTM, CNN, RNNs, ResNet-50, tensorflow & keras, RNN.

I. INTRODUCTION

Natural beauty is the key for rich lands in Bangladesh. There are many other natural features like rivers, beaches, hills, waterfalls etc. Some of our world heritage sites are Sundarban, Bagerhat's Historic Mosque, Ruins of the Buddhist Vihara at Paharpur. Many local and international tourists visit the mentioned locations to admire the natural beauty. However, due to some obstacles Bangladesh has been unable to establish itself as a tourism destination. By keeping that in mind, this paper empirically focuses on unleashing the natural beauty of Bangladesh among travelers using image captioning.

A. Background and Motivation

In our daily lives, images are pretty important. They are in people's homes, on the streets, on their jobs, and even in the internet. Stop signs in the street, artwork at the Louver museum hung behind protective glass, or

modified photos for enjoyment are accessible online. Regardless of their intended use, all photographs serve the same purpose: to communicate [1]. We want people worldwide to know about our country and its tradition, culture, and daily life, which worked as our motivation.

B. Study objective & Aspiration

Image captioning is known as the task of producing written explanations for photos. The earliest and most fundamental methodologies in picture captioning are pattern recognition based image labeling and framework-based captioning. There is a database of information in the system which has photographs and their accompanying captions or descriptions using this way. Before tying those concepts to a specified sentence template and proceeding, the system examines an image for visual ideas. This method still has a rigorous selection of words for photos with similar visual notions and minimal variability even after producing a more appropriate description method. It has proven that deep neural network-based approaches receive extraordinary outcomes on image captioning problems when large datasets are open. These methods depend laboriously on recurrent neural nets (stands for RNNs) and are powered often by a Long-Short-Term-Memory (stands for LSTM) segment [2,3].

This study experiments with the help of image captioning using TensorFlow, resnet, and flask based on the discussion above. We provided a complete assessment of the tourism industry's perspective using image captioning algorithms. In this paper, we looked at various existing raw image captioning techniques to see how they create new captions for unseen photos. We also compared and provided the results of our implementation of these models.

C. Cognitive Challenges

Extra layers are sometimes added in CNNs (for convolutional neural networks) to take care of challenges found in computer vision. These extra coatings support

the faster explanation of complex issues since the separate layers can be qualified for additional jobs to create accurate outcomes.

A deeper network can indicate the degradation issue, while the number of stacked layers might improve the model's qualities. If we put it in another way, as the number of layers in a neural network increases, the precision levels may become soaked and slowly decrease. Consequently, the model's performance spoils both training and testing data. Overfitting may not be the reason for this degeneration. It may be due to the network's initialization, optimization algorithm, or, more crucially, the problem of vanishing or ballooning slopes.

II. RELATED WORKS

Recently, encoder-decoder models for image captioning have been intensively investigated [4-12]. In its most raw version, the input image gets into the vector representation with the help of a CNN converter, and after that, the starting input gets into an RRN. The following word in the caption is speculated by RNN rooting on the preceding word without needing the secular vulnerability to be constricted to a pre-decided order, as n-gram methods do. In different ways the CNN image representation can be sent into RNN. A number of researchers [6, 10] compute the RNN's initial state, whereas others [5, 8] use that in every RNN repetition.

Attention-based image captioning was first suggested by Xu et al. [11]. The visual display of an image region incorporated by RNN position improvement. With RNN's starting state the picture region gets attended to. A rounded fusion of several area descriptors was presented, as well as a "hard" variant that uses only one region, which is used by a "soft" modification. Although one runs quite well, it is much harder to train and that to anti sampling procedure inside the condition update. The loci of attention in their method are locations in a convolutional CNN layer's activated grid. Each site is represented by the activation column that corresponds to it throughout the layer's channels.

Recently, there has been a lot of research into automatic image captioning. The research is divided into major categories: template-based, retrieval-based, and ways for creating original photo captions.

The goal of the template-based approach is to produce captions from preset templates with various blank areas. After recognizing the various objects, features, and actions, the empty spaces are filled. For example, in [13] a triplet of scene elements was used to fill the template slots in constructing image descriptions [13]. Li et al. [14] achieve this by extracting phrases relevant to specified objects. Kulkarni et al. [15] use a conditional random field (CRF) approach to infer the objects, characteristics, and prepositions. Template-based solutions can produce grammatically correct captions.

On the other hand, templates are preset, and description length cannot be modified.

The retrieval-based strategy seeks to generate a picture description by selecting the most conceptually comparable phrases from a phrase pool or duplicating terms from those other visibly similar pictures. For instance, Gong et al. [16] built visual descriptions from millions of badly labeled photos using a stacked auxiliary embedding method. Ordonez et al. [17] used the Flickr database to locate similar images in the millions of photographs and their associated descriptions, then returned the explanations of these extracted features to the query. Sun and his colleagues [18] grouped similar phrases and photos together first, then used captions from similar images in the same group to generate captions for image features. Hodosh and his associates [19] created a ranking-based framework for using sentence-based image descriptions as an information source. The authors in [19] developed a system that treats sentence-based picture description as a ranking of a set of captions for each test image. These techniques generate captions that are both general and syntactically correct. Creating image-specific and semantically appropriate captions, on the other hand, is difficult for them.

In contrast to the other two categories, novel caption creation methodologies rely heavily on deep learning and machine understanding to develop unique captions. This method is often implemented by analyzing the picture's visible information and utilizing a speech example to create picture captions from the optical material. Vinyals et al. [20], used CNN to generate the description sentence as an encoder for picture variety and LSTM as a decoder. Karpathy et al. investigated creating a natural language image description [21]. Their strategy looked for an intermodal link between the words in the description and the visual data in picture datasets with natural language descriptions. The first method aligns sentence fragments to visible places before combining them into a single description using multimodal embedding. A second recurrent neural network model is trained to construct captions using this description as training data. Xu and his associates used a convolutional neural network to extract feature maps, and an LSTM was used to characterize the input image with already extracted feature maps [22].

So, following the footsteps of these works, we implemented our image captioning project and tried to uphold the Bangladeshi culture in front of the world through images and appropriate captions. We try to use all the techniques of related works and set our dataset according to Bangladeshi heritage pictures, and try to generate the nearest possible captions in terms of many methods which we mentioned earlier [23-25].

III. RESEARCH PLAN

We develop our image captioning model by blending structures. It is also known as the RNN - CNN model. The CNN algorithm pulls features and all from a photo. We will use CNN information to help create an image caption by using the pre-trained model. Figure 1 describes it.

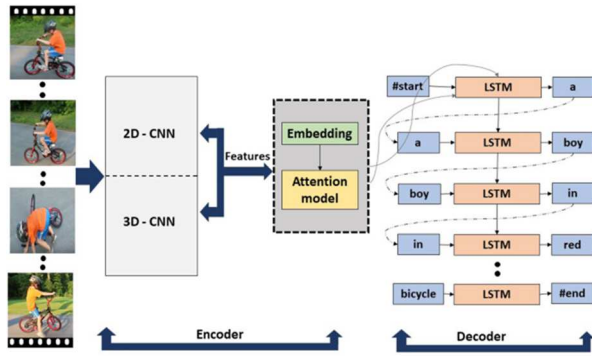


Figure 1: RNN-CNN Model [30]

We use the ResNet50 convolutional neural network (CNN), a profound network with 50 layers, to represent the image. For neural networks, web depth is critical, yet more in-depth networks are more challenging to train. ResNet50's topology facilitates network training and allows considerably more profound training, resulting in improved performance in various stirs. ResNet50 is not only more complicated than its "simple" contenders, but it also has a lesser number of parameters (weights). Consider the table below, which shows how many parameter comparisons VGG16 and ResNet-50 have. New evidence shows network depth is quite essential (Table 1).

CNN	Number of parameters
VGG16	138,357,544
ResNet50	23,587,712

Table 1: The entire number of VGG16 and ResNet50 parameters are compared.

The accuracy of networks proliferates as network deepness increases, unsurprising, and then swiftly diminishes (saturated). Overfitting does not cause this decline, and adding even more coatings increases the error of learning. ResNet50 may quickly obtain a deeper model that is comparable to the less deep network. It works by employing categories rather than layers to bypass the incoming signal without modifying it. ResNet50's higher levels must forecast the difference

between preceding layers' output and the objective function. They might simply skip the transmission by setting the scales to zero. Deep residual learning teaches the network to forecast divergences from past layers as a result.

IV. STUDY AREA

Our project target is studying the concepts of a CNN and LSTM model and then implementing the concepts of CNN and LSTM to develop a workable prototype of an image caption generator.

We will use CNN (Convolutional Neural Networks) and LSTM to construct the caption generator in this Python project. The image attributes are from Xception, a model with convolutional neural networks prepared on the flickr image dataset which is of course an imagenet database and then provided into the model which is LSTM and that's how the image illustrations will be generated [26].

Representing an image in 2D matrix is easy and CNN can do that as it is a neural web which is a type of in-depth web. Fig.2 describes it.

Mainly convolutional neural networks are used to identify the representation of an image, let us say it identifies if the image is a sky, water, living creatures etc. [26].

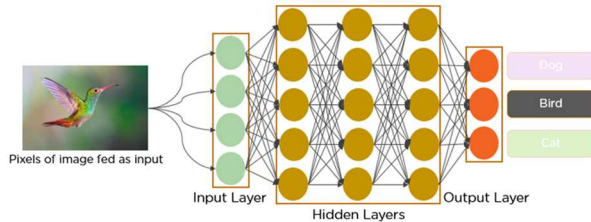


Figure 2: CNN Architecture [31]

For arranging/managing prediction challenges LSTM is being used. We can assume what the following word will be based on the prior section. It has transcended standard RNNs to overcome the constraints of RNNs with short-term memory. The LSTM may fetch out suitable data throughout the processing of intakes, and it can scrap non-related data using a forget gate [26].

This is how an LSTM cell appears [Fig.3]

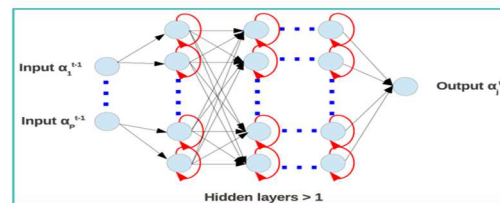


Figure 3: LSTM [32]

TensorFlow is a line-to-line, open-source machine learning platform; for differentiable programming it's a framework coating. It combines four crucial skills: Low-level tensor operations on the CPU, GPU, or TPU are efficiently performed, calculating the growth of differentiable words at every given time, extending estimate to many machines, such as clumps of hundreds of GPUs, and shipping programs ("graphs") to servers and other external scripts, portable, browsers and inserted gadgets.

Keras: It is a high level API of tensorflow. It is an obliging, greatly-rich ally for decoding ML (machine learning) states, focusing on modern deep learning. It furnishes fundamental conceptions and formation unions to create and send machine-learning solutions with increased iteration momentum.

Keras designates architects and experimenters to benefit from TensorFlow's scalability and cross-platform abilities fully; Both on TPU and large groups of GPUs keras can be executed, and our keras patterns can be exported in order to carry out in both browsers and on a movable gadget[27].

ResNet-50 [28] is a 50-layer deep CNN. It is a pre-trained edition of the network which was trained on more than a million images that can be imported from the ImageNet database. It can classify objects into 1000 different categories. ResNet's variants operate on the identical vision but have distinct digits of coatings. 50 neural network layers can be identified by ResNet50 . [28].

Keras is a popular deep learning API because of how simple it is to create prototypes using it. Keras comes with a number of pre-trained prototypes, such as Resnet50, that can be used in experiments by anyone.

Creating a remaining network in Keras for computer image applications makes image categorization astonishingly simple. Essential steps to follow are described below:

#1: We must preferably run a code to define the essential blocks for modifying the convolutional neural networks into a remaining network and evolve the convolution coalition.

#2: By mixing both blocks we can build a 50 layer resnet model.

#3: Eventually, we must design the task's model. Keras assembles it simply to make a comprehensive summary of the web architecture we have created. For future use, this can be saved or printed [28].

V. DATA COLLECTION & PREPARATION

Datasets are groups of interconnected, separate parts of data that can be observed singly or together or handled as an individual unit.

Datasets are organized in a way where one can read the data easily and understand it. For example, for a car database there will be several rows and columns which will carry different entities, such as, car name, car model, engine model, horsepower etc. [29].

In our dataset, we picked the Flickr 8k dataset, where there are approximately 8k+ pictures and 40k+ captions. We have five different captions for each picture. So, there are two folders in our Flickr 8k dataset. Mainly the name is Flickr_Data. If we go inside that folder, we will see two more folders, and their names are Flickr_TextData & Images. In the folder Flickr_TextData, we are only using the file named Flickr_8k.token.txt as our captions for the images. If we open that file, we will see all the pictures with unique codes, and each picture has five captions. So basically, the images are not named randomly. These pictures are in perspective of Bangladesh Heritage. They are named according to a code name. They maintain a serial, and we can notice that if we look at the images in our dataset. We have used 1500 images to train our model due to memory issues.

VI. METHODOLOGY : LANGUAGE BASED MODEL

Image captioning is split into two components: an image-based sample that pulls the features and subtlety from our image and a linguistic/language model that transforms the elements and entities extracted by our image-based model into a meaningful phrase.

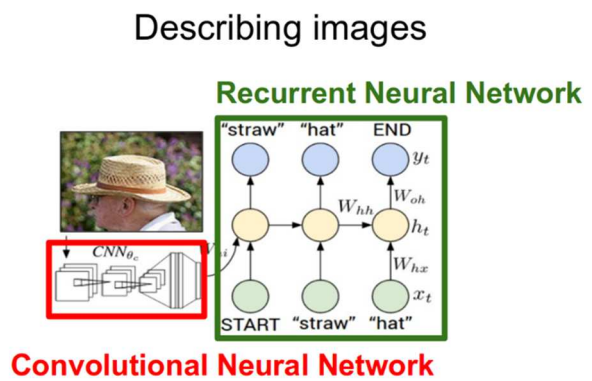


Figure 4: Image Description from [32]

In most cases, pre-trained CNN features are derived from our input data. The RNN/LSTM network's input dimension is continuously transformed to a particular dimension as the feature vector. We establish our label and target text before we train our LSTM model. This is done to ensure that our model recognizes the beginning

and conclusion of our labeled sequence [30]. By following the technique, we implement our Bangladeshi cultural and heritage dataset and train them to get the maximum value.

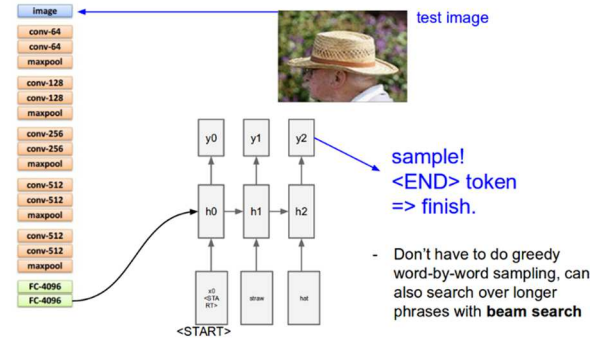


Figure 5: Machine Understanding Label Sequence [33]

VII. EXPERIMENT SETUP

A. Software Requirements

- VS code (Visual Studio)
- Jupyter Notebook

B. Model Train

First, we must import numpy, pandas, cv2, os, and glob, which are all straightforward imports. They will only be used to read and alter pictures. Then, using glob, we read the photos from our specified directory, only reading the .jpg image format. After that, we used matplotlib to display some image samples. Then we specified include_top=True and imported the resnet-50 model from keras. The input size is 224×224 , and the output size is 1000 classes, according to our model summary. This means it must anticipate 1000 different classes. The dense layer was then eliminated, allowing us to use avg pool as our final layer. After deleting the last layer, we built a new model called 'modele.' After that, we can observe that the size of our last output layer has increased to 2048. After that, we preprocessed our photos. We used the resnet model for all of our photos. We constructed an 'images feature' dictionary in which we preprocessed our images by iterating image names, values, and changing colors so that our machine could read them simply. Then, while preprocessing, we made sure that our image size was reduced. We used 1500 photos to help us improve our memory. We reshaped our image to make it easier for resnet to edit it, and we shrunk it down to 224 by 224 pixels. There are three color channels in total. The image is then altered one more time before being output as a 2048-pixel output.

The subtitles for the photographs were then pre-processed. To do so, we had to first include the caption path, and then build code to make the captions readable in our IDE. We chopped our caption dictionary so that we did not have to use the entire path name. After that,

we added the 'start of seq' and 'end of seq' strings. After that, vocabulary was added to the captions so that they could be converted into integer numbers. This indicates that the strings have been transformed to integers.

After that, we made three variables. One for storing picture attributes. Another is for holding the previous word, and the last is for holding the expected next word. We constructed a vocab size in which those three factors were used to improve our model training. Then we changed them all to numpy arrays to speed up the processing and ensure that everything works properly.

After that, we built our model and entered that data into it. We used libraries and packages that we imported. As evidenced by its code, it was a good-sized model. Then, before training, we plotted our model. Then we put our model through 150 epochs of training. The training took us nearly a day and a half to finish. Then we used the model to make predictions in which value was a key and key was a value. After that, we titled our weight files 'mine model weights.h5' and 'model.h5.' 'mine vocab.npy' was the name of the vocabulary file we used for our caption.

Verification

Unlike other machine learning assurance pipelines, in this case we will not guarantee our model against a particular matrix. Instead, we verified our model based on whether it produces the right captions, and, most importantly, whether it paid attention to the relevant features when producing those captions. We achieved this by covering the weight of the matrix of attention produced when we produced captions for a particular image in the image itself. This produced an image with specific areas that indicate what the network was paying attention to while generating captions.

Next Steps

We are trying out models on a different dataset, like the Flickr 8k dataset. We have built a new dataset based on Bangladesh's Cultural Heritage. Though we could not get far due to the lack of a high configuration PC, we did well enough to get a good caption.

C. Implementation to Website

We are creating a webpage using Python and Flask.

Flask is used to develop web applications faster. It has an easy to expand essence.

```
//importing flask
app = Flask(__name__)
def example_code():
```

```
return 'This is a test!'

if __name__ == '__main__':

    app.run()
```

Python, Programming language which allows us to work quickly and blend systems more virtually.

The presented model can be used using an illustrated a user interface that follows association:

- Authorizes the user to operate his or her trained model, choose a picture from the device.
- Generate and show a caption for a current image in the procedure defined in the design.

Numpy, cv2, os, glob, and other imported libraries were installed first. The flask was then fitted. After that, we developed the code for delivering our model's flask. We started with an elementary page for our website and then added other libraries to connect the flask to our model. We constructed captions by applying weight files and the vocab file to the exact pages.

VIII. RESULT & ANALYSIS

We built our model and entered that data into it. We have taken pictures in perspective of Bangladesh Heritage. Our system is described in Fig.6-Fig.9. We used libraries and packages that we imported. As evidenced by its code, it was a good-sized model. Then, before training, we plotted our model. Then we put our model through 150 epochs of training. The training took us nearly a day and a half to finish. Then we used the model to make predictions in which value was a key and key was a value. After that, we titled our weight files'mine model weights.h5' and'model.h5.' 'mine vocab.npy' was the name of the vocabulary file we used for our caption. The model accuracy was almost 89%.



Figure 6: Our website (Image uploading page)

In the figure 6 as we can see a website which is our BD Heritage Image captioning website and that is mainly our homepage where we can upload any picture for captioning.



Figure 7: After Prediction Page

After uploading the image, if we click on the button 'predict caption', it will take a little time and then the caption will be predicted just like in figure 7.

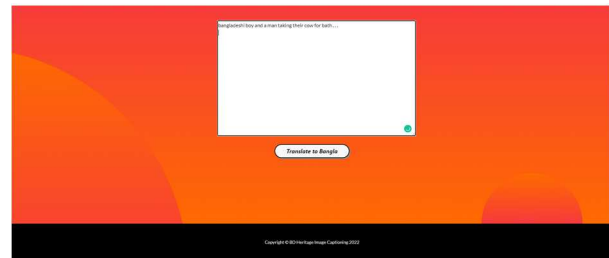


Figure 8: Translation box underneath the prediction

Then we can copy the predicted text and put it in the translation box as shown in figure 8.



Figure 9: Translated caption from the original caption

After pasting the predicted caption if we click on the button 'Translate to Bangla', it will translate the predicted caption to Bangla as shown in figure 9. For the translation part we used Google translator package.

IX. RESEARCH IMPACT

With this work, people worldwide will know more about our country. With that, we can attract more visitors, and they can enjoy the beauty of our county, which will strengthen our economic tourist site. We have gathered different types of Bangladesh heritage pictures from all around the sites. Captioned the pictures in 5 different scenarios.

There are so many places in Bangladesh which are very attractive and can blow anyone's mind with surprise. There are Sundarban, Sixty domed mosque, Paharpur, Shalbon Bihar, Maynamoti, Sonargone, Rangamati, Bandorbon, Cox's bazar etc. Every place has its own history, culture, and stories. If one digs into the history of these places then he can never react normally but be curious to know more about our country. Because that is how amazing Bangladesh's Heritage is!

If we look upon this generation then it's also beneficial towards the toddlers who have not yet got in touch with their motherland's heritage. They have to know and learn because we have to know about our beautiful motherland and therefore not only for the world but for the future generation in our country they have the right too to know everything and represent our country in the future when they will be grown up. So that they can say proudly "That's our country with its tradition and culture".

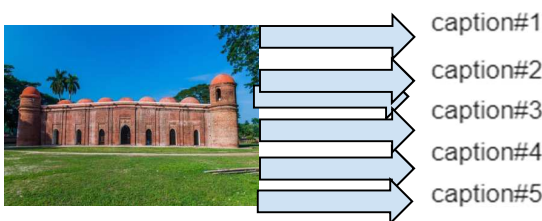


Figure 10: One of Bangladesh's Historical places

Simply we are having 5 captions for each of our images. So from those five captions we will be having one predicted by the machine and the caption will be set according to the picture as shown in the figure 10. So by this, whatever the picture related to Bangladesh's heritage a person can always get to know about Bangladesh and get curious to know more about the cultures, traditions, places, people, our daily life etc.

Our goal was to have decided to build this application in perspective to Bangladesh, to highlight that progress in machine learning, as there are many image caption sites but none in perspective to Bangladesh cultural domain. Our aim was to model train a full 8k dataset in perspective to Bangladesh culture.

X. CONCLUSION

Image captioning is the procedure of constructing natural language descriptions based on details observed in a picture to describe the situation that we visualize in

any kind of image. We have assembled all of the critical resources, including datasets. Our application had to deal with many issues: such as how we can build a model of predicting our country's heritage, finding good quality pictures related to our heritage, going places and collecting pictures, collecting datasets, editing datasets, increasing accuracy etc. We were able to overcome those obstructions. This program will be beneficial because it has a wide range of applications. We will continue to look at areas to improve our application and make the world know about our country. In the near future we plan to translate the auto generated caption into different languages so that even if someone who is not familiar with the English language can read the captions in their language.

REFERENCES

- [1] Y. Rui, T. Huang and S. Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues", *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39-62, 1999. Available: 10.1006/jvci.1999.0413.
- [2] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3128–3137, June 2015.
- [3] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. July 2004.
- [4] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In NIPS, 2015.
- [5] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, 2015.
- [6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.
- [7] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In ICML, 2014.
- [8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). ICLR, 2015.
- [9] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In ICLR, 2016.
- [10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.
- [11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.
- [12] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In NIPS, 2016.
- [13] A. Farhadi, M. Hejrati, M. A. Sadeghi et al., "Every picture tells a story: generating sentences from images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., pp. 15–29, Springer, 2010.
- [14] S. M. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. J. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 220–228, Portland, Oregon, USA, 2011.
- [15] G. Kulkarni, V. Premraj, S. Dhar et al., "Baby talk: understanding and generating image descriptions," in *CVPR means IEEE*

- Conference on Computer Vision and Pattern Recognition, pp. 2891–2903, 2011.
- [16] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, “Improving image sentence embeddings using large weakly annotated photo collections,” in European Conference on Computer Vision, pp. 529–545, Springer, 2014.
- [17] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2Text: Describing images using 1 million captioned photographs,” Advances in Neural Information Processing Systems, pp. 1143–1151, 2011.
- [18] C. Sun, C. Gan, and R. Nevatia, “Automatic concept discovery from parallel text and visual corpora,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2596–2604, Santiago, Chile, 2015.
- [19] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, 2013.
- [20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: a neural image caption generator,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164, Boston, MA, USA, 2015.
- [21] A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, Stanford University, 2017.
- [22] K. Xu, J. Ba, R. Kiros et al., “Show, attend and tell: neural image caption generation with visual attention,” in International conference on machine learning, pp. 2048–2057, Lille, France, 2015.
- [23] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” Workshop on Neural Information Processing Systems (NIPS), 2014.
- [24] X. Yu, Y. Chu, F. Jiang, Y. Guo, and D. Gong, “SVMs Classification based two-side cross domain Collaborative Filtering by inferring intrinsic user and item features,” Knowledge- Based Systems, vol. 141, pp. 80–91, 2018.
- [25] X. Yu, F. Jiang, J. Du, and D. Gong, “A cross-domain collaborative filtering algorithm with expanding user and item features via the latent factor space of auxiliary domains,” Pattern Recognition, vol. 94, pp. 96–109, 2019.
- [26] <https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn/>
- [27] https://www.tensorflow.org/tutorials/text/image_captioning
- [28] <https://viso.ai/deep-learning/resnet-residual-neural-network/>
- [29] <https://www.techtarget.com/whatis/definition/data-set#:~:text=A%20data%20set%20is%20a,some%20type%20of%20data%20structure.>
- [30] https://media.springernature.com/lw685/springer-static/image/art%3A10.1007%2Fs42979-021-00487-x/MediaObjects/42979_2021_487_Fig6_HTML.png
- [31] <https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/>
- [32] https://www.researchgate.net/figure/LSTM-deep-learning-architecture_fig3_318430079
- [33] <https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/>

Audio Band Analog Signal Measurement Instrument for Vocational School Practicum Aids

1st Nyoman Karna
School of Electrical Engineering
Telkom University
 Bandung, Indonesia
 aditya@telkomuniversity.ac.id

2nd Ridha Negara
School of Electrical Engineering
Telkom University
 Bandung, Indonesia
 ridhanegara@telkomuniversity.ac.id

3rd Bagus Aditya
School of Electrical Engineering
Telkom University
 Bandung, Indonesia
 goesaditya@telkomuniversity.ac.id

4th Adinda Fatkhah Gifary
School of Electrical Engineering
Telkom University
 Bandung, Indonesia
 adindagifary@student.telkomuniversity.ac.id

5th Dewa Rahyuni
Department of Accountancy
Faculty of Economics and Business
Universitas Padjajaran
 Bandung, Indonesia
 dewa21002@mail.unpad.ac.id

Abstract—In 2018, Indonesia has 14,064 vocational school (SMK) with 14,989 computer laboratories. All these computer laboratories are mainly used to provide students with skill about office tasks, such as word processor, spreadsheet, and presentation, with little addition like programming and design skill. As the emerging trends of electronics especially the IoT, it would be prudent to provide such skill to students to understand how signal and system works. However, many measurement instruments are quite expensive and not affordable for many vocational schools. To answer this problem, this research provides a prototype for measurement instrument to show analog signal on audio band (20Hz-20kHz) that utilize the PC in computer laboratories. To ensure students are all have the same understanding on electronics devices, this research also design a lab guide for student's lab activity. To provide an audio band analog signal measurement instrument, authors use ADC (Analog to Digital Converter) within NodeMCU to create digital oscilloscope and spectrum analyzer.

Keywords—digital oscilloscope, spectrum analyzer, microcontroller, SMK practicum kit, analog audio signal, audio band, NodeMCU

I. INTRODUCTION

Based on the 2018 SMK or Sekolah Menengah Kejuruan (vocational school) Statistics document released by the Ministry of Education and Culture, Secretariat General of the Center for Education and Culture Data and Statistics, in Indonesia there are a total of 14,064 SMKs with a composition of 25.44% with state status and 74.56% private status. In total, all of these SMKs serve 5,009,265 students, with a composition of 43.64% being in State Vocational High Schools and 56.36% in Private Vocational Schools. Of the total number of SMKs available, there are a total of 21,942 laboratories with a composition of 8,284 laboratories at State Vocational High Schools and 13,658 laboratories at private SMKs. Of the total number of 21,942 laboratories, there are 14,989 computer laboratories with a composition of 4,701 computer laboratories at State Vocational High Schools and 10,288 computer laboratories at Private Vocational High Schools [1].

This research provides the development of a computer laboratory in SMK to also be able to provide electronics and IoT labs, and at the same time change the paradigm from a computer laboratory to an IoT laboratory.

The problem raised in this study is the lack of electronic laboratories in vocational schools where students can learn, one of which is about signals and systems that can support the acceleration of technological readiness, which is one of the pillars of development based on the Global Competitiveness Index of the World Economic Forum [2]. It is hoped that from this development, the computer laboratory will not only provide computer lab work, but also electronics and IoT practicums.

One of the measuring tools needed in electronics and IoT practicum is an oscilloscope and spectrum analyzer. Oscilloscopes help students understand the shape of the signal based on the time domain, while spectrum analyzers can help students understand the shape of the signal based on the frequency domain. Unfortunately, the investment price of oscilloscopes and spectrum analyzers is very expensive because they are feature-capable of capturing signals with very high frequencies up to the order of hundreds of Mega Hertz, while for practicum in vocational schools it only requires measuring instruments for the order of tens of kilo Hertz.

With the existence of computer laboratories that are widely available in each SMK, this research provides a solution for the empowerment of these laboratories to also be able to support electronics practicum by making analog signal measuring instruments prototypes in the form of oscilloscopes and spectrum analyzers.

The method used in realizing the proposed solution is through several stages:

1. begins with a market study related to similar devices (competitors) to seek development opportunities both in terms of features and economic value;
2. then proceed with prototype development and testing to prepare it for ready-to-use products.

The objectives of this research are:

1. implementation of a prototype analog signal measuring instrument on the audio band (20Hz - 20kHz) for Electronics and IoT practicum tools in SMK, including digital oscilloscope and spectrum analyzer

- in addition to prototyping, this research also provides a curriculum and syllabus for electronics and IoT practicums that will take advantage of this measuring device.

The research was broke down into 2 (two) stages, the market study stage and prototype development, and the prototype testing phase and its readiness to become ready-to-use products. In the first stage (this research), this research provides output in the form of a prototype device that can provide functions as an oscilloscope measuring device and spectrum analyzer to describe the audio signal to a television screen or monitor. Based on the TRL (Technology Readiness Level) [3], research in this first year includes TRL 2 to TRL 4, which is what is called the Research to Prove Feasibility phase and half of the Technology Development phase.

In the second stage (future research), this research prepares prototypes developed in the first stage to become ready-to-use products. So that the output from this second stage is in the form of prototypes that have been tested for their performance along with packaging design and plans for use in practicum. Based on the TRL (Technology Readiness Level), research in this second stage covers TRL 4 to TRL 6, which is what is called the Technology Development phase and half of the Technology Demonstration phase. Fig. 1 illustrates the relationship between the research plan and TRL.

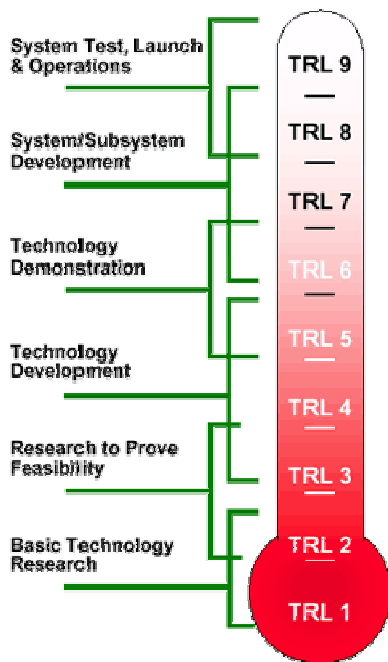


Fig. 1. Technology Readiness Level [3].

In accordance with the Indonesia government program for superior human resources, competence in electronics and IoT is one of the most needed competencies in Indonesia. This competency can be focused on vocational students who are indeed vocational schools so that they can have special skills. Skills in the field of electronics are greatly supported by a structured and applicable practicum with the help of measuring instruments as needed.

Based on the ITSM (IT Service Management) concept in Fig. 2, by definition, urgency is the time it takes for an event to have an impact on business and the environment. Therefore this research has a P2 (Important) level, namely impact =

HIGH because the existence of this measuring instrument affects the HR competence of SMK graduates throughout Indonesia, but urgency = NORMAL because this measuring tool, although it has several advantages such as low investment costs, it can be replaced with another device so that the time lag between the provision of this measuring instrument and the resulting impact is quite long.

		Impact		
		High	Medium	Low
Urgency	Urgent	P1 Critical	P2 Important	P3 Normal
	Normal	P2 Important	P3 Normal	P4 Low

Fig. 2. Relationship of Urgency and Impact (ITSM) [4].

II. RELATED RESEARCH

In his writing, Pereira [5] explained that an oscilloscope is a minimal or basic device that should be available in every electronics laboratory. There are 2 types of oscilloscopes, namely analog oscilloscopes and digital oscilloscopes. Digital oscilloscope itself is divided into 3 types, namely digital storage oscilloscope (DStO) which uses real-time sampling technique, digital sampling oscilloscope (DSaO) which uses time-equivalent sampling technique, and digital phosphor oscilloscope (DPO) which uses image and signal processing techniques. Pereira also explained that an oscilloscope has 5 main functions:

- Acquisition of electric input signal
- Signal conditioning (amplification or attenuation)
- Synchronization for a stable representation of the signal on the display media
- Visualization of the signal image on the display media
- The ability to measure and analyze signals and save and print the results of the analysis and measurements

By using the basic principles of the oscilloscope, this research also develops the prototype's ability to provide a function as a spectrum analyzer measuring device. This can be done with the help of Fourier transforms to convert from time domain to frequency domain.

Several aspects that affect the quality of the oscilloscope measurement results are the quality of the probe [6] and the sampling process [7]. The development of an oscilloscope that has economic value has also been built by utilizing a sound card available in a PC (Personal Computer) [8] and also building an external ADC (Analog to Digital Computer) device that displays the signal visualization on a PC monitor screen [9].

Most oscilloscope use built in display to show the signal to make it portable and more noise resistance. Even when the oscilloscope is using external display, they will go through a PC for processing and displaying the signal on the PC's monitor [9]. Even though all these approaches will give a low-cost oscilloscope, we need to balance between portability and features upgradability. The oscilloscope in this research will give both features, portability and displaying capability. The display capability can give the future opportunity for

knowledge discovery using online cloud-based dashboard [10].

III. SYSTEM DESIGN

According to market review, the provision of an oscilloscope measuring instrument with an economical price advantage does not yet provide the data processing capabilities required for further storage and processing, such as converting the ADC results from the time domain to the frequency domain. Fig. 3 explains the advantages of this study.

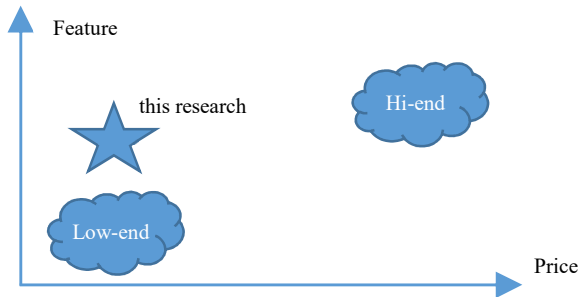


Fig. 3. Comparison of features against costs.

Hi-end oscilloscopes currently available have the ability to capture signals in the order of hundreds of Mega Hertz, so the investment price is very expensive. Meanwhile, for low-cost oscilloscopes according to literature reviews that use sound cards and also build external devices, the capabilities provided are also low, such as the limited maximum frequency that can be sampled and also require a separate computer for data processing.

Some of the well-known brands that have previously released hi-end oscilloscopes are Hewlett-Packard, Tektronix, and Hitachi. However, with a very high price, reaching tens or even hundreds of millions, it is the main obstacle for SMK schools to invest. This high price is due to the ability of the oscilloscope to be able to capture signals up to the Gigahertz order, whereas for practicum in the electronics and IoT laboratories at SMK, only an oscilloscope that is capable of capturing signals in the audio band is needed, which is the order of tens of kilohertz.

This research offers a prototype which besides being low cost, also has data processing features. This capability is also needed to add features to become a spectrum analyzer that requires Fourier transform processing.

This research uses NodeMCU for acquiring, processing, and sending sampled analog signal. There are five main reasons why authors choose NodeMCU, they are:

1. availability in Indonesia market,
2. (lower) price,
3. (bigger) memory capacity,
4. (smaller) power supply, the NodeMCU is powered through 3.3VDC, which in this prototype is supplied by 2 coin-batteries CR1025, and
5. built-in ESP8266 to transmit data via Wi-Fi.

Although, however, there are also two drawbacks on choosing NodeMCU for the system, they are:

1. (slightly) bigger size compared to, for example Arduino Nano, and
2. simpler ADC (Analog to Digital Converter) in terms of resolution (8-bit) and voltage range input.

The prototype was designed like an ordinary pen, with a sharp tip to capture the analog signal and at the other end is a cable with alligator clip to be connected to the same ground of the circuit being observed. The design is drawn in Fig. 4. The block diagram of the prototype is described in modules according to Fig. 5. Two probes, red as the sharp tip of the pen to capture the signal and black for the ground, just like regular digital oscilloscope, will determine which analog signal to be sampled. This analog signal is sampled using NodeMCU 8-bit ADC and could either stored in its RAM or sent directly to HTTP server or MQTT broker.

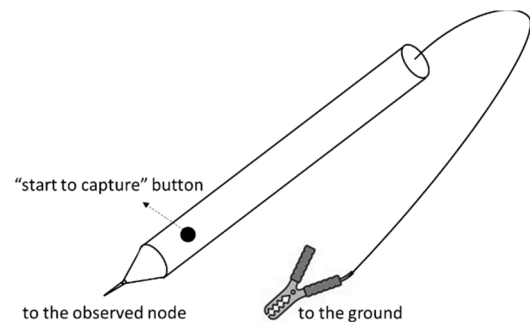


Fig. 4. Prototype design.

The prototype design in Fig. 4 shows a “start to capture” button. Student may press this button to capture the analog signal, once the sharp tip of the pen and alligator clip are already in place. The batteries in Fig. 5 were stored inside the pen. This is the reason the design is using coin batteries, so it can be easily stored inside the pen.

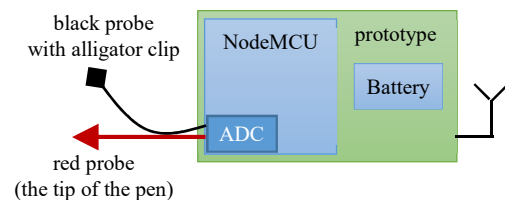


Fig. 5. Block diagram of the prototype.

Since the ADC resolution of the NodeMCU used in this research is 8 bit and the size of its usable RAM is around 50 KB then the system can store up to 50,000 samples. However, with the addition of timestamp and other additional information and for ease of addressing (data retrieval), the prototype is designed only to store 48,000 samples. If the prototype is going to sample a 20 kHz analog signal (the highest frequency in audio band), we need to sample the signal using at least 40,000 samples per second, which means the RAM is more than enough to capture a 1 second interval.

IV. TESTING AND MEASUREMENT

To ensure the oscilloscope can sample audio signal band, this research uses 2 signal generators:

1. A low-cost IC 555 timer in astable operation mode to generate a 20 kHz sawtooth periodic

signal, the maximum audible frequency, just like an audio analog signal,

2. A more sophisticated, yet rather expensive, XR2206 to simulate 20 kHz sine wave signal, the maximum audible frequency.

A. Sawtooth Signal Generator

Fig. 6 shows the IC 555 timer in astable operation, while the frequency is determined by the values of resistor R_a , resistor R_b , and ceramic capacitor C , while the duty cycle is determined by the values of resistors R_a and R_b [11].

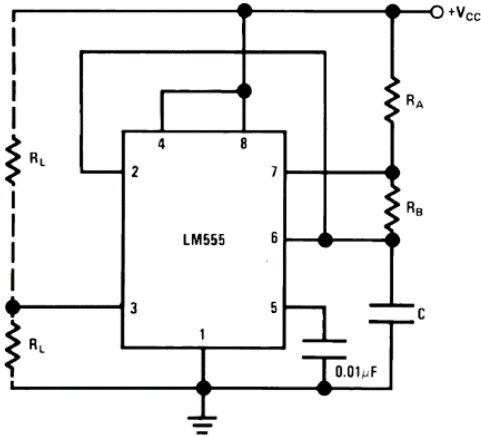


Fig. 6. Timer 555 in astable operation to generate sawtooth signal [11].

Equation (1) describes the calculation of the output frequency for sawtooth signal generator in Fig.6 (for sawtooth signal, the output is taken from pin 6 of Timer 555), just as depicted from Fig. 7. With the value of resistor $R_a = 120 \Omega$, resistor $R_b = 300 \Omega$, and ceramic capacitor $C = 100 \text{ nF}$, we get frequency = 20 kHz, while using (2) we get Duty Cycle = 42%.

$$f = 1.44 / ((R_a + 2 * R_b) * C) \tag{1}$$

$$D = R_b / (R_a + 2 * R_b) \tag{2}$$

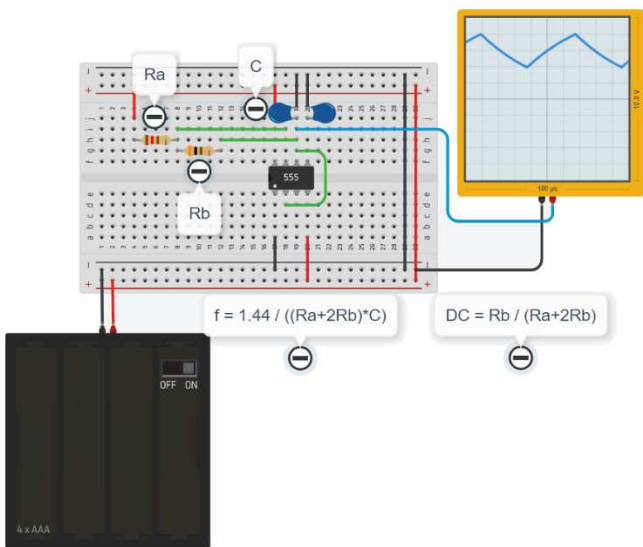


Fig. 7. Sawtooth signal for ADC input tester.

B. Sine Wave Signal Generator

For sine wave signal, we use XR 2206. Fig. 8 shows the circuit diagram for FSK (Frequency Shift Keying), which will give 2 frequencies, F_1 and F_2 , depending on the value on pin 9. Since we are going to use only sine wave with 1 frequency only, we are going to put pin 9 in open-circuited and use R_1 and C to determine the frequency using Equation (3) [12].

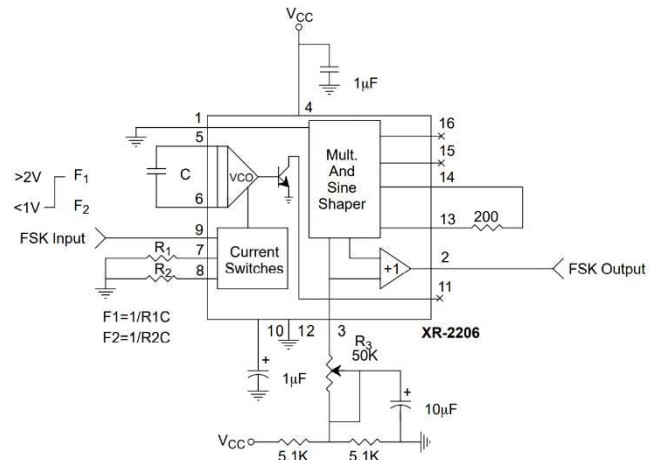


Fig. 8. XR 2206 generate sine wave signal [12].

$$f = 1 / (R_2 * C) \tag{3}$$

From Fig. 8 and Equation (3), to generate sine wave with a single 20 kHz frequency, we use $R_1 = 5 \text{ k}\Omega$ and $C = 10 \text{ nF}$.

C. Electrical Performance Analysis

Fig. 9 shows the sawtooth signal wave on a digital oscilloscope, while Fig. 10 shows the reconstructed signal from sampled signal taken previously using our prototype digital oscilloscope.

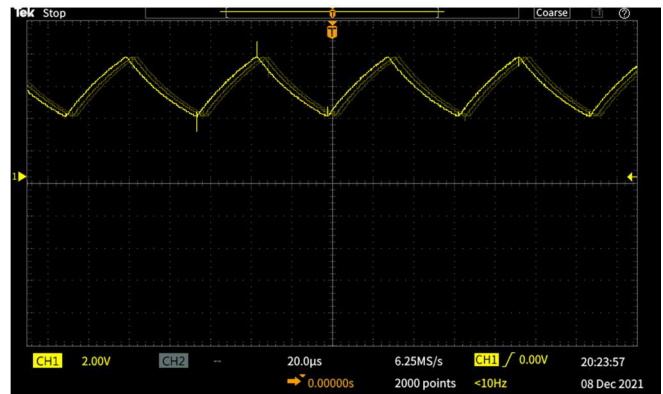


Fig. 9. Sawtooth signal from Timer 555.

Fig. 9 is taken by digital oscilloscope Tektronix TBS1052C with Timer 555 as the signal source. With a signal period of 64 µs shows the signal frequency of 15.625 kHz, lower than expected by the circuit, which supposed to provide 20 kHz. This happens due to the tolerance value of the resistor and the capacitor.

From the picture we can also see the pitch noise happened at the top peak and bottom peak of the signal, along with its shadowy signal. This happens probably due to the use of breadboard, instead of PCB (Printed Circuit Board).

Fig. 10 shows the sawtooth signal, sampled for 70 times by our prototype's ADC. The 70 sampled signal is printed out to the Serial Monitor and processed and plotted using Microsoft® Excel.

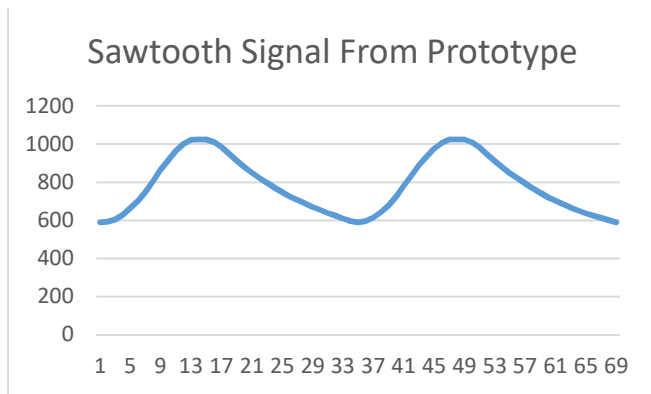


Fig. 10. Sawtooth signal from ADC of our digital oscilloscope prototype.

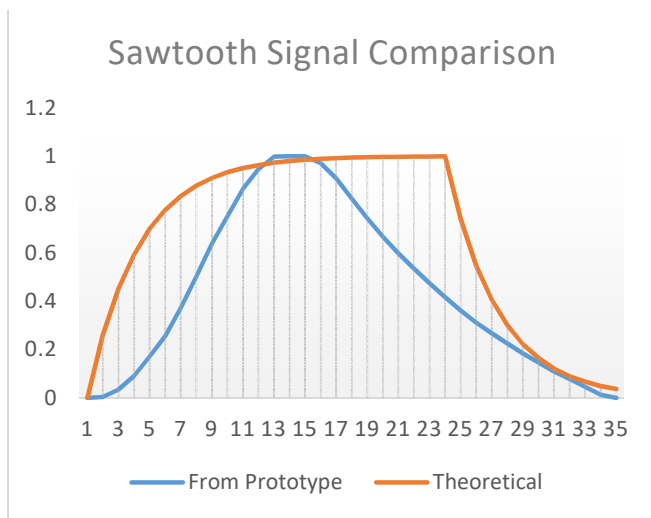


Fig. 11. Sawtooth signal comparison from ADC of our digital oscilloscope prototype and from theoretical formula.

Fig. 11 shows the comparison between the normalized sawtooth signal captured by the prototype with the theoretical approach (charging and discharging capacitor through resistor formula). Based on the comparison, we get $MSE = 0.29$.

Fig. 12 shows the sine wave signal on a digital oscilloscope, while Fig. 13 shows the reconstructed signal from sampled signal taken previously using our prototype digital oscilloscope.

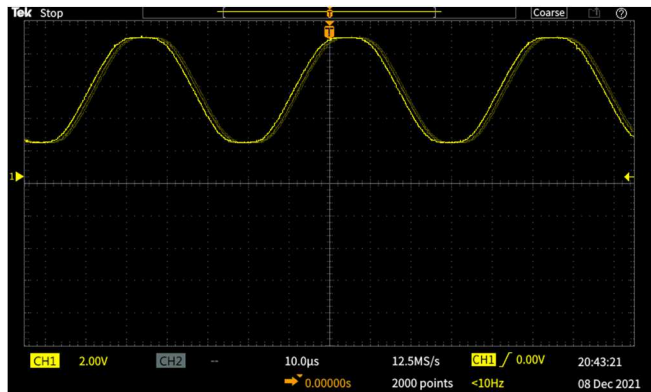


Fig. 12. Sinusoidal signal from XR2206.

Just like in Fig. 9, Fig. 12 is also taken by digital oscilloscope Tektronix TBS1052C with XR2206 as the signal source. With a signal period of $52 \mu s$ shows the signal frequency of 19.231 kHz, a little bit lower than expected by the circuit, which supposed to provide 20 kHz. This happens due to the tolerance value of the resistor and the capacitor. From Fig. 12 we can also see 2 shadow signals made by unstable sampling noise.

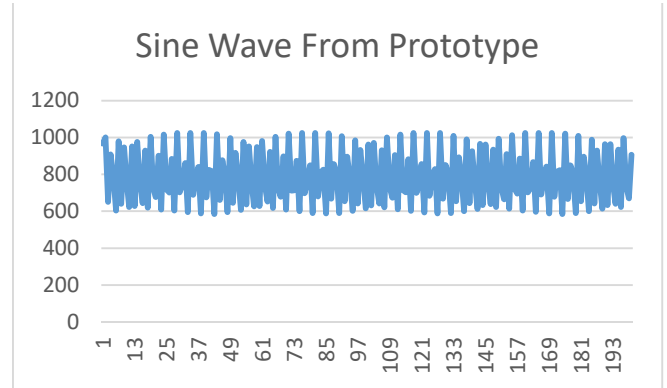


Fig. 13. Sinewave signal from ADC of our digital oscilloscope prototype.

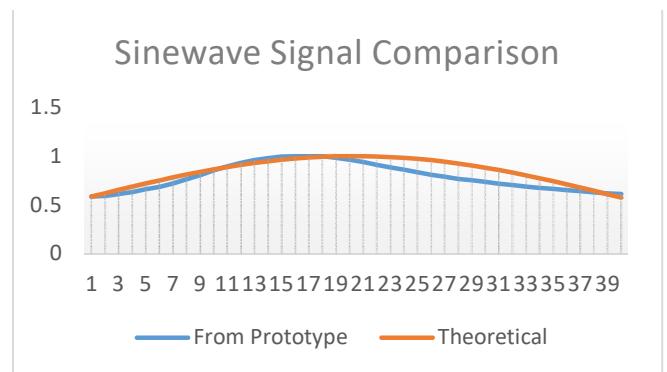


Fig. 14. Sinewave signal from ADC of our digital oscilloscope prototype.

Fig. 13 shows the comparison between the normalized sinewave signal captured by the prototype with the theoretical approach (sinusoidal formula). Based on the comparison, we get $MSE = 0.08$.

D. Price Analysis

The prototype is built using components listed in Table I. There are 4 components needed to build the prototype, 1 unit of NodeMCU, 1 unit of alligator clip, 20cm of copper cable, 1 unit of plastic casing (built by using 3D printer), and 1 unit solder tip.

TABLE I. COMPONENT LIST

No.	Component Name	Quantity	Price
1	NodeMCU Lolin	1 unit	Rp 32000
2	Alligator clip	1 unit	Rp 500
3	Copper Cable	20cm	Rp 100
4	Casing	1 unit	Rp 15000
5	Solder tip	1 unit	Rp 10000

Based on Table I, the total amount needed to build the prototype is Rp 57600 or approximately US\$ 4 each. This is affordable for many vocational schools to build the measurement instrument, in this case is digital oscilloscope, by their self.

V. CONCLUSION

The purpose of this research is to build a measurement instrument to capture signal just like a digital oscilloscope with the help of built in ADC (Analog to Digital Converter). The prototype should be small, affordable, and provide enough precision for any student in vocational school to help them understand analog signal and its application in audio range below 20 kHz. Based on our analysis, the prototype cost only Rp 57600 and gives MSE=0.29 for the sawtooth signal and MSE=0.08 for sinusoidal signal.

ACKNOWLEDGMENT

This work is supported by Basic and Applied Research program from Telkom University.

REFERENCES

- [1] Sekretariat Jenderal Pusat Data dan Statistik Pendidikan dan Kebudayaan, "Statistik Persekolahan SMK 2018/2019," Jakarta, 2018. [Online]. Available: <http://repositori.kemdikbud.go.id/13819/>.
- [2] K. Schwab and S. Zahidi, "Global Competitiveness Report Special Edition 2020: How Countries are Performing on the Road to Recovery," Switzerland, 2020. [Online]. Available: <https://www.weforum.org/reports/the-global-competitiveness-report-2020>.
- [3] Wikipedia, "Technology readiness level," *online*. https://en.wikipedia.org/wiki/Technology_readiness_level.
- [4] Octopus ITSM, "Priority Definition and Basic Service Levels," *online*. <https://wiki.octopus-itsm.com/en/articles/priority-definition-and-basic-service-levels>.
- [5] J. M. Dias Pereira, "The history and technology of oscilloscopes," *IEEE Instrum. Meas. Mag.*, vol. 9, no. 6, pp. 27–35, 2006, doi: 10.1109/MIM.2006.250640.
- [6] K. Johnson and D. Maliniak, "Oscilloscope Probes for Power Electronics: Be Sure to Choose the Right Probe for Accurate Measurements," *IEEE Power Electron. Mag.*, vol. 5, no. 1, pp. 37–44, 2018, doi: 10.1109/MPEL.2017.2782399.
- [7] D. Williams, P. Hale, and K. A. Remley, "The Sampling Oscilloscope as a Microwave Instrument," *IEEE Microw. Mag.*, vol. 8, no. 4, pp. 59–68, 2007, doi: 10.1109/MMW.2007.383954.
- [8] M. O. Hagler and D. Mehrl, "A PC with sound card as an audio waveform generator, a two-channel digital oscilloscope and a spectrum analyzer," *IEEE Trans. Educ.*, vol. 44, no. 2, pp. 15 pp.-, 2001, doi: 10.1109/13.925849.
- [9] C. Bhunia, S. Giri, S. Kar, S. Haldar, and P. Purkait, "A low-cost PC-based virtual oscilloscope," *IEEE Trans. Educ.*, vol. 47, no. 2, pp. 295–299, 2004, doi: 10.1109/TE.2004.825527.
- [10] N. Karna, "Executive dashboard as a tool for knowledge discovery," in *Proceedings - 2017 International Conference on Soft Computing, Intelligent System and Information Technology: Building Intelligence Through IOT and Big Data, ICSIT 2017*, 2017, vol. 2018-Janua, doi: 10.1109/ICSIT.2017.10.
- [11] Texas Instruments, "LM555 Timer LM555," no. February. pp. 1–21, 2000, [Online]. Available: <https://www.ti.com/lit/ds/symlink/lm555.pdf>.
- [12] EXAR Corporation, "Datasheet XR-2206," no. 510. pp. 1–16, 2008.

Enhancing SARS-CoV-2 variants Research with Blockchain Architecture

Oluwaseyi Ajayi
 Engineering Department
 Vaughn College of Aeronautics Technology
 New York, USA
 oluwaseyi.ajayi@vaughn.edu

Tarek Saadawi
 Department of Electrical Engineering,
 City University of New York, City College
 New York, USA
 saadawi@ccny.cuny.edu

Abstract— The emergence of the novel SARS-CoV-2 (Covid-19) virus in 2019 has led to continuous monitoring of the outbreak attempting to generate accurate reports of people's health information to understand the pandemic's impact. It is likely that more variants will emerge since not all countries and populations have been vaccinated. Thus, with SARS-CoV-2's constant mutation, researchers need to collect individuals' health data to study these variants and vaccine efficacy, especially those who show symptoms. However, researchers have difficulties building comprehensive datasets because people are unwilling to release their health information or have no way to report their health statuses (i.e., at-home testing). This problem stems from a lack of complete control over who assesses their health data. Hence, they cannot guarantee the security, privacy, and integrity of the disclosed health information. As the problem of building secure databases persists, researchers find it challenging to accurately report any evolving variants within a short period. This problem has resulted in several new outbreaks of the pandemic. In this work, we propose a blockchain architecture that can guarantee patients' health data integrity, privacy, and security, encouraging individuals to disclose their health information freely. This solution gives patients complete control over who assesses their health information. The framework proposed access management to patients' health data for researchers and contact tracers. This solution classifies patient health information to different sensitivity levels and manages access based on this sensitivity. In case of unauthorized access, the proposed solution detects and prevents such access, thereby ensuring the patient's health information's security, integrity, and privacy. Based on the classification, contact tracers can quickly assess the information needed from patients while the patients will be confident that no sensitive information is disclosed, reducing the burden on contact tracers.

Keywords — Blockchain, Integrity, privacy, SARS-CoV-2 virus, Security, Privacy, Confidentiality, Public Health Information, Access Management.

I. INTRODUCTION

The outbreak of the SARS-CoV-2 (Covid-19) virus since 2019 has called for an urgent need for effective research and reports to curtail further spreading and deaths from the pandemic. Based on the studies, it is highly contagious and constantly spread around the globe [1]. The Covid-19 virus most often causes respiratory symptoms similar to cold, flu, or pneumonia. Since its discovery, more than 506 million cases have been reported worldwide and have accounted for 6.2 million deaths[2]. 80.7 million cases and more than 989 thousand deaths have been reported in

the USA[2]. Currently, about nine virus variants are being monitored, while three of the evolved variants (Delta, Omicron, and Ihu) are of great concern to researchers [3]. Due to its rapid mutation rate, researchers collect patients' health information to study and report new variants. Researchers failed to accurately study the variants despite their effort to collect updated health information. This is due to a lack of willingness from people that have been exposed to disclose health information. The problem stems from an inability of the patients to control who assesses their health data. Hence, they cannot guarantee the security, privacy, and integrity of the disclosed health data. As the problem persists, researchers find it cumbersome to give an updated report of new variants and vaccine efficacy. If no solution is provided, the problem might lead to an outbreak of another variant.

Academia and industrial researchers have put forward diverse blockchain applications to preserve the privacy and confidentiality of the health information collected by contact tracers and researchers. For instance, [3] proposed BeepTrace to bridge the user/patient and authorized solvers to desensitize the user ID and location information. A blockchain architecture that preserves the privacy of contact tracing Covid-19 patients was proposed in [4], while [5] evaluates the benefits of using Big data and blockchain technology to deal with the pandemic and some data quality issues that still present challenges to decision making. The authors in [6] proposed a blockchain architecture that enforces accountability and transparency to bridge the gap between the on-chain and off-chain user information collected by contact tracers.

Despite the numerous researches on privacy-preservation of individual health information, none of the available solutions focus on access management of the health information, hence the motivation to this work.

A. SARS-CoV-2 Virus researchers

These people are dedicated to monitoring the new variants symptoms to make an informed decision and recommendation to the government. These people study the symptoms to i). reports about the new variants, and ii) report the vaccine efficacies. To obtain accurate information based on the new or existing symptoms, they rely on information supplied by contact tracers. They can also request covid related health information from Covid isolation centers and treatment sites or healthcare providers and hospitals.

B. Contact Tracing

This is a process of identifying, assessing, and managing people who might have contact with Covid-19 infected person and subsequent collection of further information about these contacts[7]. The purpose of contact tracing is to minimize the spread of the virus. Contact tracing has been used and proven effective in minimizing the spread of infectious diseases in general. The contact tracing process relied heavily on recalling a list of people they have been in contact with over the previous weeks or locations the confirmed person has been. Letters, phone calls, or emails are being used to inform people who might be exposed. Thus, completeness and accuracy of the list and timeliness and efficiency of the tracing are limited by such a labor-intensive, traditional contact tracing approach. Recently, different digitized contracting through smartphone apps has been deployed to solve the labor-intensive problem. For instance, Singapore TraceTogether[8], GAEN[9], National Health Service (NHS) COVID-19 App [10], and China Health code system [11]. Researchers rely on contact tracing information to study the virus accurately.

C. Contribution

The proposed solution aims to develop a blockchain framework that can guarantee people's health information privacy, security, confidentiality, and integrity so that individuals will be willing to disclose necessary information to study and report new cases of Covid-19 variants. In addition, this framework promotes health equity by often engaging underserved communities. The specific aims of the solution are summarized below:

- **Guarantee PHI privacy, integrity, security, and confidentiality using access management:** One of the significant problems experienced by researchers in obtaining accurate information about the SARS-CoV-2 virus is that there is no way to guarantee privacy or integrity and confidentiality of the health information requested. As a result, individuals find it challenging to trust non-medical practitioners(e.g., Researchers, contact tracers, etc.) requesting their health information. Even if researchers turn to healthcare providers to obtain health information, they are restricted by government policies (e.g., HIPAA) about sharing health records with third parties. Our proposed solution combat this challenge by ensuring that patients have complete control over who accesses health records and that no sensitive health information is disclosed during this process.
- **Classifies individual health information based on its sensitivity.** People find it challenging to share their health information because there is no access control strategy to deter requesters from accessing more than requested information. This means there is no way patients can control the level of access once the health records have been shared. Our approach approached this challenge by classifying the health information based on

sensitivity and defining different access levels to enforce restrictions. Apart from creating these access levels, the architecture manages access to health information via verification, authentication, and permission steps.

- **Fosters SARS-CoV-2 virus research:** One of the reasons it is required to study and report the SARS-CoV-2 virus continuously is its constant mutation. This constant mutation has led to a strong need for a prompt and constant collection of health information to generate a dataset for compelling studies and reports. by implementing the proposed solution, people will be willing to release helpful information about their covid status without privacy and integrity breach concerns
- **Reduces the burden on Contact Tracing:** The bulk of the dataset building for effective studying of the variants relies on the information provided by contact tracing. It is often challenging to get accurate health information as people might be unwilling to talk or misrepresent their situation. Our architecture allows the individual's health providers (e.g., testing center) to upload accurate information to the blockchain network. Contact tracers can access this information from the blockchain platform by requesting a permit from the health data owner. In this case, both the contact tracer and owner are anonymous, and the contact tracer only accesses the permitted/required information.

This paper's remainder is organized as follows: Section II discusses the background and related works on blockchain applications for healthcare data access. Section III describes the proposed architecture's method. Section IV presents the preliminary results, and finally, section V presents the conclusions of this paper and possible future works.

II. BACKGROUND AND RELATED WORKS

Researchers have applied blockchain in diverse areas, including data security in healthcare, in the past years. The blockchain capabilities make the blockchain technology more secure than other distributed database platforms. Such features are attractive for their secure data collection and management, prompting other researchers to propose blockchain applications in sharing health information. For example, the authors in [1] proposed Gem Health Network, allowing different healthcare specialists to access the same health information from a shared infrastructure. The authors [2] proposed a user-centric health data sharing solution that utilizes a decentralized and permissioned blockchain. In their work, a mobile application is deployed to collect health data from personal wearable devices, manual input, and medical devices, and synchronize data

to the cloud for data sharing with healthcare providers and health insurance companies.

The authors in [3] proposed a conceptual design for sharing personal continuous dynamic health data using blockchain technology supplemented by cloud storage to share the health-related information in a secure and transparent manner. The primary goal of their proposed system is to enable users to own, control and share their personal health data securely. The solution also provides an efficient way for researchers and commercial data consumers to collect high quality personal health data for research and commercial purposes. A blockchain-supported architectural framework for secure control of personal data in a health information exchange by pairing user-generated acceptable use policies with smart contracts was proposed in [4]. In [5], the authors proposed blockchain-enabled IoMT to address the security and privacy concerns of IoMT systems to combat COVID-19. The authors in [6] presented a blockchain-enabled privacy-preserving contact tracing scheme, BeepTrace. Their work proposed adopting blockchain bridging the user/patient and the authorized solvers to desensitize the user ID and location information.

Our proposed application is unique from previous proposals in that it ensures data security with additional steps. Although the available solutions utilized the same blockchain platform to share personal health information, our approach introduced two specific, novel steps: i. our proposed solution classifies health information based on sensitivity for easy access management. ii. our application mounted an extra verification step on the existing blockchain framework to ensure that no unauthorized access is permitted. This mode of operations makes our approach different from existing blockchain applications in healthcare. The proposed architecture classifies and stores people's health information (PHI) then manages access based on the sensitivity and class of the requester, thereby guaranteeing the privacy, integrity, and security of an individual's health information. The requesters are classified (e.g., Researcher, Contact tracer, Physician, Nurse), and access to sensitive information is defined based on each class. Any entity that submits a request to access health data goes through verification steps. If the verification fails, the access is denied, and no health information is returned to the sender. Although the proposed framework can manage access from different requester classes, this work focuses on SARS-CoV-2 virus studies and reports.

III. METHODOLOGY

This approach aims to build a distributed database that guarantees health information security, privacy, and integrity by providing access management. The novel access management strategy encourages individuals to disclose their health information, enhancing covid-19

studies and vaccine efficacies testing. Figure 1 describes the high-level mode of operation. The key actors involved in the proposed approach are:

- **PHI generators:** These are health professionals that generate patient health information. An example is a Covid testing center, a healthcare provider etc. These entities send the PHI into the blockchain network[1] via a classification module(described below).
- **Owner:** The individual that owns the health information uploaded to the blockchain network. The owner approves whenever there is a request to access the stored health records. The owner grants or denies the access based on the request and sender's information. The request pushed to the owner is a successfully verified request.
- **Requesters:** These entities need health records for different purposes, e.g., research. They request health information relating to the Covid-19 virus to conduct better studies, make accurate reports, adjust the vaccines, evaluate the performance, advise government agencies like CDC, task-force on the pandemic, etc. Examples are SARS-CoV-2 virus researchers and contact tracers.

Apart from the above descriptions, the architecture can enable the inclusion of underrepresented populations, especially those whose health issues have been neglected in research

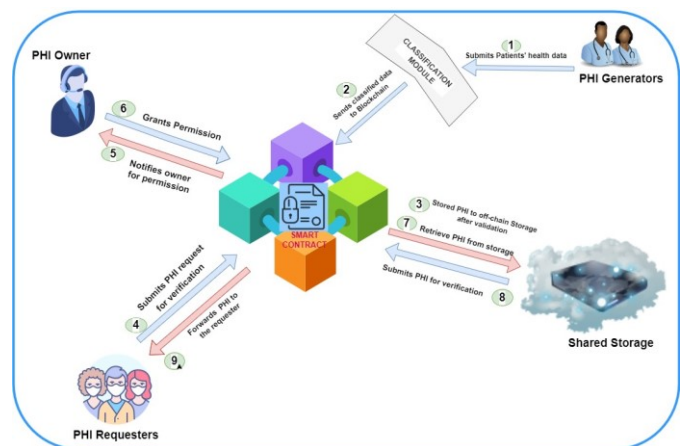


Fig. 1. The Proposed architecture

The architecture comprises a blockchain network that features a programmable smart contract and an off-chain storage system, e.g., IPFS. The main building blocks of the proposed architecture are shown in Fig. 2. The building blocks comprise a classification system, access management module, and storage system.

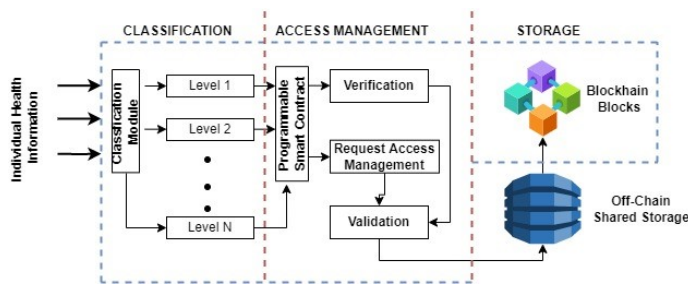


Fig. 2. Building blocks of the proposed architecture

1. Classification module

The classification module is an interactive script developed in a python programming language that identifies sensitive health information and groups them according to the sensitivity level. This module reformats the health record according to sensitivity features written in the smart contract. For the script to effectively classify ingress health records, it is required that the health record be sent in a particular format. Based on this format, the script interacts, extracts, and separates the information according to sensitivity. For proof-of-concept evaluated in this work, we defined three sensitivity levels:

- *Basic information:* The general or basic information about the individual is extracted here. An example of the information extracted includes name, age, date of birth, gender, weight, height, race, origin, etc. The basic information is accompanied by a tag that defines the sensitivity and access level.
- *Covid-19 information:* The health records here are Covid-19 and vaccines-related information. Individual Covid-19 related information is obtained and registered here. An example of the information extracted includes underlining health problems, age, gender, race, origin, vaccine status, Covid-19 symptoms, etc. The Covid-19 information is accompanied by a tag that defines who can access the information and the level of sensitivity
- *Other Health information.* The health information here is other health records not covid-related. Examples include health history from PCP, list of drugs, previous tests, etc. This information is not to be accessed by Covid-19 researchers or contact tracers. The health information is accompanied by a tag that defines the sensitivity and access level

Algorithm 1 below illustrates how the classification module script reads the incoming individual health information (IHI), classifies it, and sets the sensitivity level. To achieve this classification, the module matches the field of the received IHI to what it runs then retrieves the features.

```

Algorithm 1: Classification Module
1 Require:
2   Mapping(Incoming IHI ->struct) private Sensitivity;
3
4 Procedure: ReadIncomingIHI (Individual Health Information)
5   For ihi in IHI:
6     If ihi ∈ basic information
7       basic information ← ihi
8       Sensitivity => 1
9     else
10      If ihi ∈ covid-19 related information
11        covid-19 information ← ihi
12        Sensitivity => 2
13      else
14        other health information ← ihi
15        Sensitivity => 3
16      end If
17    end For loop
18  end procedure
    
```

Based on these classifications, the module prepares separate transactions. These transactions are digitally signed and submitted to the blockchain network for verification and validation. In addition to the information, the sending blockchain node submits its information for sender authentication.

2. Access Management Module

This module describes the verification, validation, and request access management of health information. The transaction sensitivity level, public keys, and authorized nodes' information are stored in the smart contract. The transaction owner is authenticated, and the submitted transactions are verified. Verification and request access control are managed on the server-side of the architecture, i.e., smart contract, while transaction validation is handled by blockchain consensus protocol. Thus, smart contract handles the following functions:

i. Node Authentication

The smart contract handles nodes authentication. The purpose of authenticating a node is to ensure that all transactions are sent from authorized nodes only. The smart contract retrieves the sender's information that accompanies the transaction and invokes a code that compares it with stored information. The information confirmed includes the transaction account, MAC and IP addresses, and the digital signature. We recognized that hackers could easily spoof the MAC and IP addresses in a highly unsecured environment; hence, we added a multifactor authentication process that involves verifying the sender's transaction account and digital signature in addition to the verified MAC and IP addresses. The pseudocode in Algorithm 1 describes the snippet of the smart contract that handles the node's authentication. The smart contract invokes the code that reads the sender's information. It verifies the digital signature using the owner's public key (o.pk) and confirms the transaction account, MAC,

and IP addresses. The algorithm invokes the transaction verification code if the node authentication is successful. For a node to be successfully authenticated, the sender must be an authorized node i.e., the digital signature and other information must be verified. If any of these fail, the node's transaction is dropped.

Algorithm 2: Node Authentication

```

1 Require:
2 Mapping (pubkey=>bytes32) public Keys;
3 Mapping (account => bytes32) public Account;
4 Mapping (ip_addr=>bytes32) public IP;
5 Mapping (mac_addr => bytes32) public MAC;
6
7 procedure: ReadOwnerData (Transaction tag)
8     Transaction.owner ←msg.sender
9     Transaction.account ← account
10    Transaction.pubkey ← o.pk
11    Transaction.ip ← ip_addr
12    Transaction.mac ←mac_addr
13    Return TRUE
14 end procedure
15
16 procedure: NodeAuthentication (OwnerData)
17     require (ip_addr ∈ IP)
18     require (mac_addr ∈ MAC)
19     require (pubkey ∈ Keys)
20     require (account ∈ Account)
21     require (o.pk verifies digital signature)
22     return successful authentication with timestamp
23 end procedure

```

ii. *Transaction verification*

One of the novelties of the proposed solution is the implementation of the verification stage. The verification stage ensures that any submitted transaction conforms with the agreed-upon format and contains the correct information. The verification step also ensures no leakage of sensitive information due to incorrect classification—the smart contract reviews and matches tags accompanying each transaction to its sensitivity level. A snippet from the smart contract runs through the submitted transaction to ensure no features are missing. Every authorized node sign its transaction with the private key, and a smart contract code verifies the digital signature using the known public keys whenever a transaction is submitted to the blockchain. For digital signature implementation, we generate key pairs with Digital Signature Algorithm (DSA) with 1024 bit-length using the command "*ssh-keygen -t dsa -b 1024*" and store the public key in the smart contract. The smart contract also stores the tag information accompanying each transaction to manage access. A snippet from the smart contract shows algorithm 3 pseudocode describing the transaction verification process. For transaction verification to be successful, the transaction must agree with the defined format, and the sensitivity level must match the information. If the sensitivity level does not match the

information, the sender is notified, and the transaction is dropped.

Algorithm 3: Transaction Verification

```

1 Require:
2 Mapping (transaction=>struct) public Sensitivity;
3 Mapping (transaction => struct) public Format;
4
5 procedure: ReadTransaction (format, tag)
6     Transaction.field ←format
7     Transaction.tag ← level
8     Return TRUE
9 end procedure
10
11 procedure: NodeAuthentication (OwnerData)
12     require (format != NULL)
13     require (level == tag[level])
14     return successful verification with timestamp
15 end procedure

```

iii. *Request Access Management*

Another novelty of this proposed work is its ability to enforce access control on the stored health information. The purpose of managing the requester's access is to ensure that a requester is assessing only the required information. We implement the access control function by defining a transaction retrieval policy that uses the transaction's sensitivity. We defined each class of requesters' information and wrote it into the smart contract. Every requester must submit the required information according to the access level requested. If a requester fails to provide the required information, access is denied. After the smart contract has successfully verified the requester, an access request is pushed to the owner for approval. If approved, the requester can access the information; else, access is denied. Algorithm 4 shows that for a request to be granted, a smart contract must verify the requester's information, the owner must approve the request, access level must match the requester level (e.g., a researcher requesting covid-19 related health information), and a sender cannot request multiple information (i.e., no cross-level request).

Algorithm 4: Access Management

```

1 Require:
2 Mapping(request => struct) public Request;
3 Mapping (requester_sender => struct) public Information
4
5 procedure: ReadRequest(requester)
6 Requester_info ← info
7 Requester_request ← health_info
8 Requester_level ← access_level
9 Return True
10 end procedure
11
12 procedure: RetrieveTransaction (Transaction)
13 require (access_level ∈ Tag[level])
14 require (info must be correct)
15 require(health.owner must approve)
16 require (request_info == 1)
17 Allow Access to Specified Information
18 end procedure
19 end procedure

```

3. Storage Module

i. Off-chain shared Storage

Owing to the massive amount of data accompanying patient health records, we introduce a decentralized off-chain shared storage medium called interplanetary File System(IPFS). One of the features of IPFS is that each stored file can be allocated a unique hash according to its content to identify it uniquely. So, if somebody uploads two files with the same hash value, IPFS pins them to the server and creates only one entry in the content addressed storage block. Due to this feature, IPFS has no access control, and files over IPFS can be directly accessed by their respective hash value [14]. Unlike [14], access to the IPFS is limited to the blockchain network. Blockchain network offloads health records to the IPFS after successful verification and authentication. A successfully verified transaction is pushed to the off-chain Storage, and reference to the location of this information is sent to the blockchain network for validation and Storage. The reference location of the newly added information undergoes a validation process and is added to the blockchain network blocks. Our previous works have extensively described the validation process [15-17].

ii. Blockchain Blocks

After successfully validating the location reference, the newly added block reflects on the ledger of all blockchain nodes, and the transaction is ready to be retrieved. All blockchain nodes receive the newly added block notification but do not access the block's content. The smart contract retrieves the transaction address after the requester has been successfully verified and the owner approves the request.

IV. RESULT

The on-going work is implemented on an Ethereum blockchain platform. We use *Solidity v 0.8.11*

implementation for smart contracts and *geth v 1.10.13* for Ethereum. For initial testing of the proof-of-concept, the private blockchain network is set up in the laboratory with five computers serving as blockchain nodes (Fig. 1) to evaluate the performance. We evaluated access management on the stored information, and the preliminary result was as described below. We also present the further work to be carried out.

A. Access Control/Management

We evaluate the access control function by implementing a node attempt to retrieve an unauthorized transaction. We submit a transaction using one of the blockchains. The submitted information was successfully verified and pushed to the off-chain Storage. The reference to the location was successfully verified and added to blocks. Another random node was used to query this stored information. The smart contract handles each transaction's access control by comparing the information of the requesting node to the authorized requester. The node queries the blockchain for the content of a newly added block that it is not privileged to download. We investigate further by querying the blockchain using a node that can retrieve the data. The node successfully downloads the block's content from the blockchain. No information was returned to the requester because the node is not privileged to retrieve the data.

V. CONCLUSION

We proposed a blockchain framework that can guarantee the privacy and integrity of shared patient's health information, thereby encouraging people to disclose more helpful health information to enhance the SARS-CoV-2 (Covid-19) virus study and monitor vaccine efficacies. Implementing the proposed architecture will foster the study and reporting of new SARS-CoV-2 variants. Due to its distributed database implementation, the proposed solution can enhance the inclusion of underrepresented populations, especially those whose health issues have been neglected in research.

Further works

As part of the further work, we plan to implement the following

- Add Off-chain shared storage medium to offload the Storage of health information from the blockchain
- Evaluate the effectiveness of the classification with real data
- Evaluate and present the results of the architecture's effectiveness towards guaranteeing privacy, integrity, confidentiality, and security of Individuals' public health information.
- Describes how the architecture achieves the user training module and inclusion of underserved populations

REFERENCES

- [1] <https://www.yalemedicine.org/news> accessed 4/20/2022
- [2] <https://coronavirus.jhu.edu/map.html> accessed 4/20/2022
- [3] H. Xu, L. Zhang, O. Onireti, Y. Fang, W. J. Buchanan and M. A. Imran, "BeepTrace: Blockchain-Enabled Privacy-Preserving Contact Tracing for COVID-19 Pandemic and Beyond," in *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3915-3929, 1 March 2021, doi: 10.1109/JIOT.2020.3025953.
- [4] S. Tahir, H. Tahir, A. Sajjad, M. Rajarajan and F. Khan, "Privacy-preserving COVID-19 contact tracing using blockchain," in *Journal of Communications and Networks*, vol. 23, no. 5, pp. 360-373, Oct. 2021, doi: 10.23919/JCN.2021.000031.
- [5] I. Ezzine and L. Benhlima, "Technology against COVID-19 A Blockchain-based framework for Data Quality," 2020 6th IEEE Congress on Information Science and Technology (CiSt), 2020, pp. 84-89, doi: 10.1109/CiSt49399.2021.9357200.
- [6] H. R. Hasan, K. Salah, R. Jayaraman, I. Yaqoob, M. Omar and S. Ellahham, "COVID-19 Contact Tracing Using Blockchain," in *IEEE Access*, vol. 9, pp. 62956-62971, 2021, doi: 10.1109/ACCESS.2021.3074753.
- [7] Ferretti et al., "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, 2020, Art. no. eabb6936. [Online]. Available: <https://science.sciencemag.org/content/368/6491/eabb6936>
- [8] J. Bay et al., "BlueTrace: A privacy-preserving protocol for community driven contact tracing across borders," Singapore's Government Technol. Agency, Singapore, White Paper, p. 9, 2020.
- [9] Exposure Notification, Apple Inc., Cupertino, CA, USA and Google LLC., Mountain View, CA, USA, May 2020.
- [10] I. Levy. (2020). The Security Behind the NHS Contact Tracing App. Accessed: May 8, 2020. [Online]. Available: <https://www.ncsc.gov.uk/blog-post/security-behind-nhs-contact-tracing-app>
- [11] P. Mozur, R. Zhong, and A. Krolik, In *Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags*, New York, NY, USA, 2020. [Online]. Available: <https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html>
- [12] O. Ajayi, M. Abouali and T. Saadawi, "Secure Architecture for Inter-Healthcare Electronic Health Records Exchange," 2020 IEEE International IoT, Electronics, and Mechatronics Conference (IEMTRONICS), 2020, pp. 1-6, doi: 10.1109/IEMTRONICS51293.2020.9216336.
- [13] O. Ajayi and T. Saadawi, "Detecting Insider Attacks in Blockchain Networks," 2021 International Symposium on Networks, Computers, and Communications (ISNCC), 2021, pp. 1-7, doi: 10.1109/ISNCC52172.2021.9615799.
- [14] Meryem Abouali, Kartikeya Sharma, Oluwaseyi Ajayi, Tarek Saadawi, "Blockchain Framework for Secured On-Demand Patient Health Records Sharing" 2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON 2021) December 1-4, 2021, New York, USA.
- [15] O. Ajayi, M. Cherian and T. Saadawi, "Secured Cyber-Attack Signatures Distribution using Blockchain Technology." 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), New York, NY, USA, 2019, pp. 482-488.
- [16] O. Ajayi and T. Saadawi, "Blockchain-Based Architecture for Secured Cyber-Attack Features Exchange," 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 2020, pp. 100-107, doi: 10.1109/CSCloud-EdgeCom49738.2020.00025.
- [17] O. Ajayi, O. Igbe and T. Saadawi, 2019. Consortium Blockchain-Based Architecture for Cyber-attack Signatures and Features Distribution, 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 0541-0549, doi: 10.1109/UEMCON47517.2019.8993036.

Convolutional Neural Network Structure to Detect and Localize CTC Using Image Processing

Shorouq Al-Eidi*, Omar Darwish†, Ghaith Husari‡, Yuanzhu Chen§, and Mahmoud Elkhodr¶

*Computer Science Department, Memorial University of Newfoundland, Canada

†Information Security and Applied Computing, Eastern Michigan University, USA

‡Department of Computing, East Tennessee State University, USA

§School of Computing, Queen's University, Canada

¶School of Engineering and Technology, Central Queensland University, Australia

Abstract—Many cybersecurity attacks utilize Covert Timing Channels as a method to secretly transmit (steal) sensitive information from target networks such as untrusted Internet of Things (IoT) and 5G/6G networks. Such attacks aim to violate the confidentiality and privacy of the data that resides in the targeted networks by transmitting the stolen information in a stealth manner over a prolonged period of time to avoid detection by cyber defenses and anti-exfiltration tools.

In this work, we proposed a novel approach that utilize novel Artificial Intelligence (AI) algorithms, in particular, deep learning to detect and localize covert channels over cyber networks. Taking advantage of the rapidly improving deep learning algorithms in image processing, we convert the malicious and normal network traffic (or packets) inter-arrival times to colored images. Then, we implement an AI-based approach using the popular deep learning algorithm Convolutional Neural Network (CNN) to process images and detect the ones that contain malicious CTC activities. Finally, we design and conduct a set of experiments to evaluate the ability of our proposed system to detect and localize CTC-based privacy attacks. The conducted experiments show that our approach yielded a high accuracy of 96.75% in detecting stealth covert channels.

Keywords—Big data analytics, artificial intelligence, privacy-aware defense, covert timing channel detection, image processing, deep learning.

I. INTRODUCTION

Fifth-generation (5G) wireless communication networks have been deployed in many countries to date. 5G technology boosts mobile capacity and data rate. However, the advancements promised by 5G technologies go far beyond increasing the data rate for mobile users to the provision of massive inter-connectivity at a scale never witnessed before. The current volume of data exchanged over 5G networks shows explosive growth. Moreover, 5G technology transforms many services, including location-based services, given its location positioning accuracy. Coupling 5G accurate positioning and low latency capabilities with the Internet of Things (IoT) allows for a significant overhaul in inter-connectivity and automation, paving the way to new models of seamless and ubiquitous interactions [1]. Using IoT sensors, many companies, businesses, or even individuals would have the possibility of collecting and analyzing the massive amounts of data collected or provided by various IoT applications either directly or via other low power

technologies. The cohesion of IoT, 5G, and 6G, will transform many industries and move humanity many steps ahead on the road of realizing the vision of intelligent or smart cities. Nevertheless, given that global 5G implementation requires substantial investments in communications infrastructure, 6G is poised to augment the revolution in ubiquitous and cross-borders communications.

Industrial collaborative robots [2], holographic meetings and health [3], and metaverses' cross and external connectivity with real-world applications are some examples of the innovative applications and technologies be driven and enabled by 6G technology. The massive increase in data rate currently witnessed with 5G is poised to be at least ten folds more in 6G. This advancement in data rate will also extend to provide advances in service availability, increase reliability, and removing many of the current network and geographical boundaries currently experienced with 5G.

However, in addition to inheriting the security vulnerabilities of the IoT and the previous generations, 6G has its own set of challenges [4]. New radio technologies will introduce many unknown threat vectors; the physical location of ultra-massive MIMO systems will also be a concern and many other challenges such as quantum-safe communications and cyberattacks against pervasive intelligence. While network steganography is relatively not fully explored, the integration of billions of embedded devices that communicate to each other in 6G networks will open the door to several new covert timing channels (CTCs) attack. Given the CTCs' ability to evade traditional detection and prevention methods and the massive data rate capabilities of 6G networks, transmitting hidden data bits over 6G networks poses a serious security threat [5].

Traditional CTCs detection methods presented in [6]–[10] use anomaly-based detection techniques. Using these techniques, the statistical features of both the covert and normal traffic are extracted and analyzed to detect covert communication. However, the volume and variability of data provisioned in 6G networks hinder the effectiveness of CTCs statistical-based detection methods.

Recently, rather than focusing on non-visible features for CTC detection, our previous work [11] proposed a CTC

visualization and classification model to improve the accuracy of CTCs' detection. The visualization approach introduced a technique for converting traffic inter-arrival times to colored images and transforming CTC detection into an image classification problem. The CTC visualization method can handle the attack detection and localization problems, but it suffers from the high time cost needed for image feature extraction. Moreover, these feature extraction methods also demonstrate low efficiency when exposed to large datasets. Additionally, the quality of extracted features significantly affects the overall performance of classifiers, which satisfies specific criteria or assumptions. Therefore, the challenge for building CTC detection models is to find a means for extracting features effectively and automatically.

To this end, this work employs a convolutional neural network (CNN), one of the most common and powerful deep learning algorithms in computer vision. CNN can automatically extract image features and avoid the drawbacks of manually extracting inappropriate features for images classification. It is critical to have a quick and light technique for detecting CTC attacks with large data to make decisions appropriately. Moreover, discovering the covert part of the traffic containing covert messages is another essential objective. It provides the ability to drop only the malicious part of traffic flows while allowing the rest of the traffic flow to pass through. This precise detection allows alerting the existence of CTCs to start the mitigation defend mechanisms, which significantly improves the quality of service that gets impaired upon dropping the entire traffic flow.

The main contributions of this work are summarized as follows:

- Utilize the convolutional neural network to improve CTC detection accuracy.
- Determine the optimum model architecture that provides high-performance outcomes in CTC detection by evaluating model hyper-parameter effects and determining optimal parameter values.
- Use a variety of image-based datasets to evaluate the proposed model's effectiveness and quantitative performance.
- Propose a mechanism that uses the convolutional neural network to accurately pinpoint the covert part that contains covert messages within traffic flows.

The rest of this paper is organized as follows. Section II, illustrates the detection model. Section III, describes the experimental setup. Section IV shows the results and analysis of this research. Finally, Section V, concludes this work.

II. CTC DETECTION USING CNN APPROACH

This section provides a detailed description of our proposed detection approach and the methods used to accomplish the detection objective of covert traffic.

A. Collecting and Processing Network Data

Figure 1 overviews the major components of our proposed approach. First, we collect a dataset of network traffic induced by two processes 1) a normal (non-malicious) client-server packet exchange and 2) a malicious process that implements covert channels to steal information from the network. Then, we calculate and extract the inter-arrival times (IATs), the time duration between two packets, from our traffic flows. Once these IATs are extracted, we convert these numbers into colored images to utilize the advanced image processing techniques in the domain of deep learning. To convert IATs to colored images, we implemented a module that creates a 2D matrix of size $n \times n$ where n ranged from 16 to 128. More details about these sizes are provided in Section III. This module then systematically places each IAT in this matrix in a row by row manner. Intuitively, each stream of IAT numbers are placed in one matrix. Then, the module normalizes the numbers in each matrix into a range between 0 and 255 to represent a colored image. We used the matplotlib library [12] to create colored images from these 2D matrices.

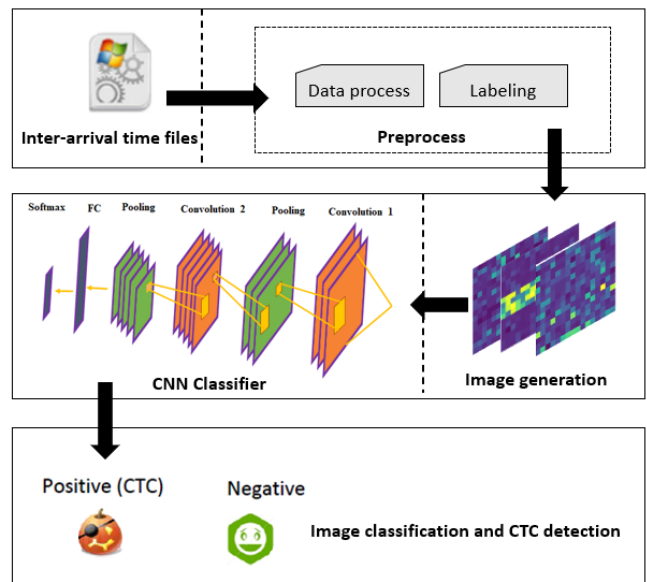


Figure 1: CNN framework.

B. Shallow CNN classifier

CNN uses weight sharing and pooling methods to reduce the number of parameters and decrease computation time and memory. This reduction allows the training depth model to be implemented. Deep CNNs have shown stellar performance in the field of image processing as demonstrated by AlexNet [13], and VGGNet [14]. However, CNN remains a time and memory-demanding algorithm due to its complicated learning network and network topologies that contain many parameters such as kernel size and layers. For instance, VGG16 contains 16 layers with more than a million parameters making it

a memory-demanding convolutional neural network model. Shallow CNN is underutilized, and research on such networks has lately gained interest.

To utilize deep learning in network security resources with limited memory and computation resources, we chose shallow CNN. The shallow CNN network has less layers and smaller convolution kernels, making it a lighter-weight model that places less demand on memory and computation.

C. CNN feature extraction and classification of images

The previous step yielded colored images such that some of these images resulted from a normal (non-malicious) traffic exchange, and the remainder of these images resulted from a malicious traffic exchange that used covert time channels to steal data from the victim network.

To construct a deep learning model that detects covert channels from these images, the first step is to extract effective features from these images. To this end, we used the feature extraction algorithm in CNN.

A colored image is essentially a matrix (or simply a set) of pixels. Once an image is fed into the CNN feature extraction module, the convolution layers and filters of a CNN produce feature maps. First, as shown by Algorithm 14, the model extracts 32 features to feed into the first hidden layer. Then, the second hidden layer utilizes 64 features for its learning. To reduce the computation requirements, the max-pool layers, which have the size of 2×2 , come after the second layer.

Then, the feature maps are normalized using CNN batch normalization [15] to adapt to the changing parameters during the training process. This adaptation increases the stability of the neural network and, in turn, improves classification accuracy. After the normalization process, the features are converted into one dimension for the fully connected layer. To reduce overfitting, dropout is used to improve the generalization of the model. Finally, the softmax classifier is used to identify the images into two categories, covert and overt.

III. EXPERIMENTAL SETUP

A. Experimental dataset

This work focuses on detecting one of the most sophisticated CTC attacks known as Cautious Covert Timing Channel (CCTC) [11]. CCTC is a cyberattack that attempts to mimic normal network traffic behavior to some extent, making it more difficult to detect by security detection methods such as firewalls and intrusion detection systems.

For the detection process, four CTC datasets have been built with 4000 streams (50% covert and 50% overt) in each dataset. The main difference between these datasets is the stream length (256, 1024, 4096, and 16384). Covert message size in each covert stream is 64 bits regardless of the length of the stream. The delay time used to generate CTCs in all datasets is 0.5μ where μ is the mean inter-arrival times of normal traffic, which equals 0.664 seconds. However, for each stream-based dataset

Algorithm 1 CNN Based CTC Detection Algorithm

Input: Training data and parameters

Output: Class label

Create classification model:

- 1: Generate the first convolutional layer, followed by batch normalization
- 2: Add max pooling layer
- 3: Generate the second convolutional layer, followed by batch normalization
- 4: Add max pooling layer
- 5: Add a dense layer

Model training:

LOOP Process

- 6: **for** I_{th} epoch **do**
- 7: **for** J_{th} mini-batch **do**
- 8: Extract feature representation
- 9: Minimize the cost function
- 10: Update model parameters using Adam optimizer
- 11: **end for**
- 12: **end for**
- 13: Output the result according softmax classifier

Model testing:

- 14: Test the model using the test dataset
-

(256, 1024, 4096, and 16384), an image-based dataset version has been generated with different image dimensions ranging from small images with the dimensions of 16×16 pixels to large images with the dimensions of 128×128 pixels. Table II details the different image dimension that were used in our evaluation. In fact, for the detection process, the covert message is located in the middle of the stream, while in the localization process, it has been either in the beginning, middle, or end of the stream. Accordingly, a new dataset has been generated for the localization process of 6000 covert streams (2000 in each location) with 1024 bits length, which then it converted to an image of 32×32 pixels.

B. Experimental environment

To run the training and testing experiments of CTC detection and localization, we used a ThinkStation-P920 server running Linux version, with 24 cores and 62.42GB RAM. Python, the programming language with various libraries such as TensorFlow and Keras is used to implement the models and experiments.

C. Measures of CTC detection

In this paper, most common performance metrics have been used to evaluate approaches performance. Our evaluation aims of providing a comprehensive analysis measures to provide more details about CTC detection and localization process. The following list presents each one of these performance measures.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$F_1score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

where True Positive (TP) is the number of image samples that are classified correctly as CTCs. True Negative (TN) is the number of image samples that are correctly identify as non-CTC. False Positive (FP) is the number of image samples that are incorrectly classified as CTCs. False Negative (FN) is the number of images that are incorrectly classified as non-CTC while, in fact, they are CTCs.

D. Choice of model hyperparameters

The CNN model has various hyper-parameters that determine the network structure (e.g., number of filters) and how the network model is trained (e.g., type of optimizer). The model performance can vary considerably according to the selected set of hyper-parameters. For example, the number of layers was ascertained after repeated experiments for designing the CNN model. The convolution layers were determined to be optimal with small kernels, in this work the kernel size of the convolution layers was determined at 3×3 .

We evaluated several model hyper-parameters to determine the best classification model. We adjusted and updated the model hyper-parameters in the search space and found the best value in the grid search. For example, we evaluate the model performance using the epochs range between 10-30. Based on the model results, we found that when the number of model epochs reaches 15, the model results did not continually improve, and the training time increased; based on that, the best number of epochs for the CNN proposed model is set as 15. The learning rate was another important hyper-parameter determining whether the objective function converges to a local minimum in the appropriate time and speeding up training in the early stages. In this study the learning rate was sit to 0.001 and the decay of the learning rate to 0.02. Moreover, the Adam optimizer and Cross-Entropy loss function are used with a 0.5 dropout probability for all model experiments.

IV. EXPERIMENTAL RESULTS EVALUATION

Our evaluation seeks to measure the performance of the proposed CNN approach in the following terms: (a) the effectiveness of the proposed approach to detect CTCs under different batch normalization configurations; (b) the efficiency of the proposed approach to detect CTCs using different hyper-parameters; (c) the impacts of various traffic image sizes and

shapes on the proposed approaches performance; d) compare the results of the proposed approach with different traditional machine learning and deep CNN classifiers based on their accuracy and interpret-ability in detecting CTC, and e) the ability of the proposed approach to pinpoint the covert part of the traffic flows.

The training and testing experiments were repeated ten times, and then the average of the findings was used as the final assessment value. 10-fold cross-validation is used to evaluate each classifier model in this work. We also used well-known metrics in the model's evaluation, including accuracy, precision, recall, and F_1 score.

A. Effects of using batch normalization on the CNN model performance

Our first set of evaluation results shows our approach's performance in detecting CTCs under different batch normalization (BN) configurations to demonstrate how BN may speed network training and enhance model performance. This section provides two CNN variation instances with two assumptions for utilizing BN, as shown in cases 1 and 2. The results of these two cases are compared with the CNN model that utilizes BN after each convolution layer.

case 1: Remove the BN technique after the first convolutional layer only and leaving all other layers and model parameters without any changing.

case 2: Remove BN techniques after the first and second convolutional layers, while all remaining layer parameters have no changes.

According to the experiment accuracy results, the CNN approach with BN after each convolutional layer achieved the higher accuracy of 96.75%, which is 0.3 % better than the proposed CNN approach in case 1 that obtained 96.45% of accuracy and 0.55 % higher than the CNN approach in case 2 with 96.20% of accuracy.

CNN with batch normalization technique have significantly quicker convergence than the results obtained with the standard parameterization without BN. The BN technique helps to normalize the parameters of each layer, making updating the model parameter process is more stable. This indicates that the CNN with BN can learn data characteristics more effectively and quickly. Consequently, we can infer that our classification accuracy obtains the best classification result when combining CNN with BN, proving the validity and efficiency of the proposed approach.

B. Effect of hyperparameters on CNN approach performance

Our second set of evaluation results shows our approach's performance using different model hyper-parameters, including the type and the number of layers with other activation functions. Table I shows the proposed CNN approach performance based on using three convolutional layers and one fully-connected layer with sigmoid, Tanh, and ReLU activation functions.

Table results show that the diverse activation functions affect the proposed approach’s performance and training time. The results demonstrate that the ReLU function is more time-saving than other functions, which is overall more effective in terms of approach performance with convolutional and fully connected layers. As shown by the table, the proposed CNN approach with sitting two convolutional layers and one fully-connected layer obtained the highest detection CTC accuracy of 96.75%, and the lowest training time, which is 5.03 seconds, where ReLU function is set for both types of layers.

Table I: Accuracy results of proposed model based on various hyperparameters.

Layer	Function	Accuracy	Training time(s)
Convolution 1	Sigmoid	95.10	5.50
	ReLU	95.51	4.00
Convolution 2	Sigmoid	96.60	7.22
	ReLU	96.75	5.03
Convolution 3	Sigmoid	96.20	9.50
	ReLU	96.55	6.20
Fully connected	Sigmoid	96.60	7.22
	Tanh	96.65	6.50
	ReLU	96.75	5.03

C. Image size effect on CNN model performance

Our third set of evaluation results shows the performance of our approach for detecting CTC using various image-based datasets. In the CNN technique, image size considers one of the crucial parameters affecting the approach classification results. The small size of the image could not help extracting enough information, which causes losing the attack properties. At the same time, using a large size of image would increase the computational time without any improvement in the accuracy of classification.

In this paper, we conducted a variety of experiments based on using different image sizes, including 16×16 , 32×32 , 64×64 , and 128×128 pixels as discussed in Section III-A. Fixed image size is used as required for CNN approach input. Figure 2 show the accuracy results of two classes classification using the four image sizes.

Also, Table II shows various performance metrics of using these same four image sizes. The results show the approach performance is increased when the image size is increased. However, a large image needs more training time. We noted that when the image size is getting more than 32×32 pixels, the approach performance does not significantly increase, and the training time is increased rapidly. For these reasons, we choose 32×32 pixels as the best size of input image-based data that can be used to achieve the best CTC classification and detection.

D. Image reshaping effect on CNN approach performance

This work also studies the effect of image shapes on the proposed approach performance. We used the traffic inter-arrival times to generate a non-square image with different dimensions. We choose the image size 16×16 , and then we

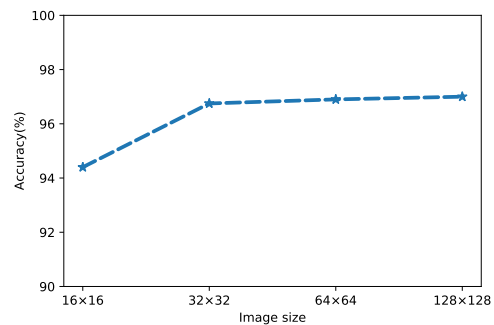


Figure 2: Accuracy results of using different image sizes.

Table II: Performance measures of CNN model using four image sizes.

Size of image	Precision	Recall	F_1 score	Time(s)
16×16	94.46	92.00	92.56	4.00
32×32	95.95	95.75	95.85	5.03
64×64	95.93	95.22	95.84	12.00
128×128	95.44	95.00	95.22	27.00

reshape it to three non-square image shapes, including 4×64 , 8×32 , and 2×128 pixels. The experiment results shown in Table III present the image shape of 64×4 obtained 94.451% of accuracy, 32×8 achieved of 94.455% accuracy, 128×2 achieved also 94.438% of accuracy, comparing to what is the model achieved 94.46% of accuracy for the 16×16 pixels.

Table III: Accuracy results using different image shapes.

Size of image	Accuracy
64×4	94.451
32×8	94.455
128×2	94.438
16×16	94.46

Based on the performance results, we noticed that changing the image shapes did not affect the classification approach’s performance, which means the width and height of the image did not affect the pattern of covert traffic in the image. Thus, a shallow convolution network with a small filter size and one stride would be worked well; the filter can correctly detect the split pattern in the covert images.

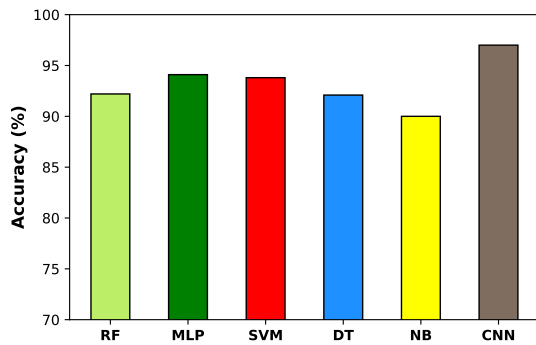
E. A comparison of the proposed approach with machine learning and deep CNN approaches

To validate the performance and efficiency of the proposed approach, we present our experimental results and compare them with various popular approaches. These detection approaches include traditional machine learning and deep CNN approaches, same as shown in Figure 3 and Table IV.

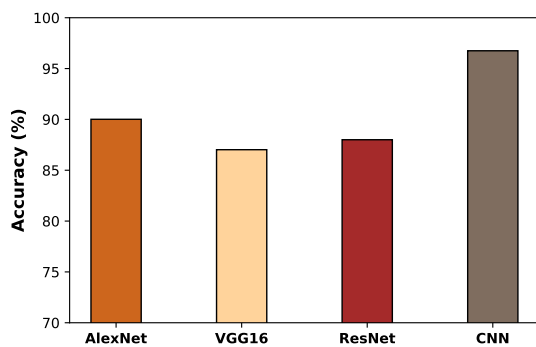
In this part, firstly, we compare our approach’s effectiveness in detecting covert channels with five traditional machine approaches, including Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB). These approaches utilized

extracted image features to identify and detect CTCs. For this comparison, 32×32 image size was used. The results in Figure 3(a) and Table IV show the highest CTC detection accuracy is achieved by the CNN proposed approach, which achieved the accuracy of 96.75%. The second highest CTC detection was achieved by the MLP, which reached 94.1%. The SVM approach achieved third place with an accuracy of 93.7%, whereas the NB approach achieved the lowest accuracy of 90%.

Then, to verify the benefit of the proposed shallow CNN approach over deep CNN approaches, we compare the proposed shallow CNN approach with the most common other deep CNN approaches, including VGG16, AlexNet, and ResNet. These approaches were designed with a high number of convolution, and fully connected layers, such as the ResNet approach consists of 419 filters with 5x5 convolutions in the first layer only. Figure 3(b) shows the shallow proposed CNN approach with 32 and 64 filters in the first and second convolutional layers, outperforms the AlexNet approach by 6.74%, achieving an accuracy of 90.01%, and 9.73% and 8.75% more accurate than the VGG16 and ResNet approaches, with accuracies of 87.02% and 88%, respectively.



(a) Accuracy results with machine learning approaches



(b) Accuracy results with deep CNN approaches.

Figure 3: Accuracy results of proposed approach with machine machine learning (a) and deep CNN approaches (b).

Table IV also shows the shallow CNN proposed approach

is achieved the highest score of correct predictions with less training time compared with other deep CNN approaches. For instance, the AlexNet approach achieved the second level of precision, which can predict 89.02% of the CTC samples correctly. In contrast, the ResNet classifier is reached 85.50%, which is the lowest precision level after the VGG16 approach. In the final column of Table IV the convolutional layers are denoted by Conv in the last column of the table, while fully connected layers are denoted by Fc (including output layer).

Table IV: Classification results of CNN approach with shallow machine learning and deep CNN approaches.

Classifier	Precision	Recall	Training time(s)	Con./Fc
RF	92.11	90.00	63.30	-
MLP	93.01	93.00	36.02	-
SVM	92.00	89.00	73.96	-
DT	91.80	88.99	66.25	-
NB	90.00	90.88	25.10	-
AlexNet	89.02	86.00	25	5/3
VGG16	86.00	85.50	42	13/3
ResNet	85.50	87.00	35	7/1
CNN	95.95	95.75	5.03	2/2

F. Localization of CTC in traffic flows

As mentioned earlier, detecting a CTC attack is essential to prevent malicious applications that steal sensitive data over the network. The traffic will be dropped to stop the exfiltration data process by successfully detecting these covert data. Dropping traffic flows containing covert data is an effective defense against CTC, but it is also very disruptive to non-malicious applications' quality of Service of the non-malicious applications.

Finding the parts of traffic flows containing the covert data is an important goal, which provides the ability to drop only the malicious part while allowing the normal traffic to pass through. This accurate identification substantially reduces the disruptions of overt traffic caused by non-malicious applications.

This part of our study investigates and presents the proposed CNN approach's performance in pinpointing the location of covert data within traffic flows. To do this, we designed experiments in which the covert messages are injected into one of three locations: *beginning*, *middle*, and *end*. Then, we investigated our approach's performance to find the correct traffic flow segment that contained these covert data.

These location labels are generated automatically using the malicious agent. For instance, if the malicious agent injects the covert data into the beginning of traffic, it will also provide the label beginning to that traffic flow then generate the image with the same label of the inter-arrival times of that traffic flow. Further, it does the same with the other two labels, the middle and the end. This labeled image-based dataset of covert traffic flow is then used as input for training and testing CNN proposed approach.

As demonstrated in Table V, our proposed approach successfully identified the segments that contained the covert data

with an accuracy of 94.01%, and less false positive (FP) and false-negative (FN). We also ran machine learning and deep CNN approaches, including SVM, AlexNet, and Decision Tree (DT), to provide a baseline for comparison. Based on the table results, AlexNet achieved second place with the detection accuracy of 92.01%. The SVM achieved third place with an accuracy of 91.05%. DT came in last place with an accuracy of 90.54% and high FN.

Table V: Results of pinpointing covert messages within traffic flows using different approaches.

Classifier model	Accuracy	FP	FN
SVM	91.05	5.53	12.98
DT	90.54	5.13	14.39
AlexNet	92.01	4.29	12.29
CNN	94.01	3.17	6.72

V. CONCLUSION

Cyber attacks targeting the Internet of Things (IoT) and 5G/6G networks with the objective of stealing sensitive information have been rapidly increasing. Covert Timing Channels provides a maliciously effective method to steal data from the targeted networks in a defense evasive manner.

In this work, we presented a novel deep learning-based approach to detect and localize CTC-based cyber attacks at the exfiltration phase. First, we converted the traffic data (packets) to colored images, then we constructed a Convolutional Neural Network (CNN) model to classify those images to malicious or normal. Finally, we evaluated our CNN model using a set of comprehensive experiments and the model achieved an accuracy of 96.75% , a relatively high accuracy in this domain.

In addition, we evaluated the capability of the CNN model to pinpoint the location of covert data with traffic flows. The model achieved a high accuracy of 94.01% in pinpointing the stolen covert data in traffic flows.

In the future, we plan to extend this work by test the capability of our CNN model in detecting more advanced stealthy spyware utilizing port knocking. In this attack (e.g., Covert knocks), the malware sends signal packets to a closed network port to enable it and allow stolen information to flow through it. This attack is more difficult to detect as it can utilize multiple ports (even closed ones) to create covert channels. Detecting such distributed malicious activity will require a novel method that can investigate network traffic on each port as well as the aggregated traffic activity of multiple ports to detect advanced stealthy spyware.

REFERENCES

[1] M. Elkhodr, S. Shahrestani, and H. Cheung, "Managing the internet of things," in *2015 IEEE International Conference on Data Science and Data Intensive Systems*, 2015, pp. 579–585.

[2] Y. Tashkoush, I. Haj-Mahmoud, O. Darwish, M. Maabreh, B. Alsinglawi, M. Elkhodr, and N. Alsaedi, "Enhancing robots navigation in internet of things indoor systems," *Computers*, vol. 10, no. 11, p. 153, 2021.

[3] M. Rath, "Real time analysis based on intelligent applications of big data and iot in smart health care systems," *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, vol. 3, no. 2, pp. 45–61, 2018.

[4] M. W. Akhtar, S. A. Hassan, R. Ghaffar, H. Jung, S. Garg, and M. S. Hossain, "The shift to 6g communications: vision and requirements," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–27, 2020.

[5] S. Al-Eidi, O. Darwish, and Y. Chen, "Covert timing channel analysis either as cyber attacks or confidential applications," *Sensors*, vol. 20, no. 8, p. 2417, 2020.

[6] K. Borders and A. Prakash, "Web tap: detecting covert web traffic," in *Proceedings of the 11th ACM conference on Computer and communications security*, 2004, pp. 110–120.

[7] S. Cabuk, "Network covert channels: Design, analysis, detection, and elimination," Ph.D. dissertation, Purdue University, 2006.

[8] S. Gianvecchio and H. Wang, "An entropy-based approach to detecting covert timing channels," *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 6, pp. 785–797, 2010.

[9] R. Archibald and D. Ghosal, "A comparative analysis of detection metrics for covert timing channels," *Computers & security*, vol. 45, pp. 284–292, 2014.

[10] P. L. Shrestha, M. Hempel, F. Rezaei, and H. Sharif, "Leveraging statistical feature points for generalized detection of covert timing channels," in *2014 IEEE Military Communications Conference*. IEEE, 2014, pp. 7–11.

[11] S. Al-Eidi, O. Darwish, Y. Chen, and G. Husari, "Snapcatch: Automatic detection of covert timing channels using image processing and machine learning," *IEEE Access*, 2020.

[12] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

FPGA Implementation of Phase Recovery Technique for Complex Transforms

Poorvi Bhaskar
 Department of ECE
 SRM Institute of Science and
 Technology
 Tamil Nadu, India
pb9202@srmist.edu.in

Yuvaraj S
 Department of ECE
 SRM Institute of Science and
 Technology
 Tamil Nadu, India
yuvarajs@srmist.edu.in

Palanisamy P
 Department of ECE
 NIT, Trichy
 Tamil Nadu, India
palan@nitt.edu

Thilagavathy R
 Department of ECE
 NIT, Trichy
 Tamil Nadu, India
thilagavathy@nitt.edu

Abstract—The ECG signals are one of the most important signals to check the human heart's condition. On monitoring the heart continuously, a large amount of ECG signal data will be produced. So, there is a need for efficient compression techniques. Discrete Anamorphic Stretch Transform (DAST) is one of the most efficient techniques. It is a one-dimensional complex transform that includes the phase recovery technique for recovering the phase from the magnitudes. This paper deals with implementing the phase recovery block in Field Programmable Gate Array (FPGA), which will recover the phase by using magnitudes. Phase recovery block plays a key role in reconstructing the phases from the magnitudes. First, the required signal is passed through the linear filter or phase recovery filter. Then the phase value is estimated using a non-iterative algorithm depending on the linearity and causality conditions. The new approach for the phase recovery block is also used for any complex signal transmission. The input ECG signal is taken from the MIT-BIH Arrhythmia database and implementation is carried out in Artix-7 NEXYS 4 DDR FPGA Board. The performance of the phase recovery block is quantified in terms of hardware and computational complexity.

Keywords—*Electrocardiogram; Discrete Anamorphic Stretch Transform; Coordinate Rotation Digital Compute; Phase Recovery; Field Programmable Gate Array.*

I. INTRODUCTION

The electrocardiogram (ECG) signals are generally used for detecting various heart diseases. They are recorded, processed, stored, and transmitted continuously through wired and wireless communication networks. Cardiac monitoring continuously over the long term is required in many cases such as monitoring patients in critical conditions,

detecting pathological events occurring sporadically over a long time, and continuous remote monitoring of the aged population. The enormous volume of ECG data is generated in these cases, resulting in a “BIG DATA” that needs to be transmitted and stored with less bandwidth and storage space. It requires a higher ECG data compression factor while maintaining a high degree of accuracy in the reconstructed signal. The millions of ECGs are taken yearly, thus putting a heavy burden on the physicians and cardiologists, who have to diagnose these ECGs manually. Big Data leads to challenges in acquisition, analysis, storage, and transmission [4]. Several wireless technologies such as Bluetooth, Zig bee, Wi-Fi, and GSM are used to transmit the ECG signals. The maximum signal rate limits all these technologies. So, to send at a lower rate than this, bandwidth should be reduced. Hence, a novel approach to data capture, compression, and transmission is required to deal with such data loads. A two-dimensional DAST is proposed for Image data compression in [4]. Application of the DAST for ECG data compression is proposed in [1-3]. The phase recovery for complex transform is simulated and verified in MATLAB in the literature [1] and [3]. Field Programmable Gate Arrays (FPGA) contain a large number of configurable logic blocks (CLBs) connected through programmable interconnects. Due to their programmable nature, FPGAs are best suited for various applications. Xilinx offers complete solutions comprising FPGA devices, advanced software, and configurable, ready-to-use IP cores [13]. The FPGA implementation of the phase recovery block is carried out first time in this paper. The phase recovery block contains many mathematical and trigonometric functions. The CORDIC algorithm is used to implement the trigonometric functions in FPGA [6-17].

II. PHASE RECOVERY TECHNIQUE SS

A. Discrete Anamorphic Stretch Transform

A complex one-dimensional (1D) DAST is presented in [1] as a pre-processor for ECG data compression. This preprocessing is achieved through a mathematical reshaping of the signal and non-iterative. But the sampling process is modified in Compressive sensing and needs feature detection and as well as iterative in nature. The complex 1D DAST of a 1D ECG signal $B[m]$ is denoted as $\tilde{e}[n]$ and is expressed by

$$\tilde{e}[n] = \sum_{m=-M}^M K[m].B[n-m] \quad (1)$$

where $K[m]$ is the 1D two-sided DAST kernel of length $2M+1$. $K[n]$ is defined as

$$K[n] = e^{i\Phi[n]} \quad (2)$$

$K[n] = 0$ for $n > M$ where $2M+1$ is the size of the kernel the DAST kernel phase profiles ($\Phi[n]$) can be obtained as

$$\text{Sublinear } \Phi[n] = a \left((n \cdot \text{atan}(b \cdot n)) - \frac{\ln((b^2 \cdot n^2) + 1)}{2 \cdot b} \right) \quad (3)$$

where \ln indicates the natural logarithm, \cos is the cosine function. After the suitable selection of the variable ‘a’, the resulting signal size is associated with the warping strength b . DAST is well suited for the pre-compression of the ECG signal for real-time transmission using channels with limited bandwidth [1].

B. Compression using Complex Transform

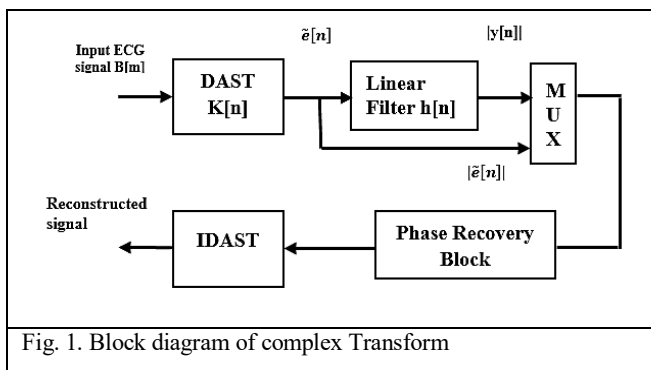


Fig. 1. Block diagram of complex Transform

The block diagram of ECG compression using complex transform is shown in Fig. 1. The input ECG signal $B[m]$ is convolved with DAST sublinear kernel $K[n]$ with kernel size 5.

To enable the phase recovery, the DAST performed complex ECG signal $\tilde{e}[n]$ is fed to a linear frequency filter (phase recovery filter) with impulse response $h[n]$. A two-tap all-pass filter with constant magnitude and the sum of the phases $\Phi_h[0]$ and $\Phi_h[1]$ equal to 180 degrees is proposed in [1] and is used here. Only the magnitude of the unfiltered signal $\tilde{e}[n]$ ($|\tilde{e}[n]|$) and filtered signal $y[n]$ ($|y[n]|$) are transmitted. At the receiver side, the phase of the signal is recovered using these two magnitudes in the phase recovery block. The ECG signal is properly reconstructed after performing the inverse DAST.

C. Phase Recovery using a linear filter

The procedure proposed in [1] and [5] for phase recovery is adopted in this work. $\Phi_e[n]$, the phase of $\tilde{e}[n]$ is given by

$$\Phi_e[n] = \cos^{-1} \left(\frac{|y[n]|^2 - |\tilde{e}[n]|^2 |h[0]|^2 - |A_i|^2}{2|\tilde{e}[n]||h[0]||A_i|} - \Phi_h[0] + \Phi_{A_i} \right) \quad (4)$$

where

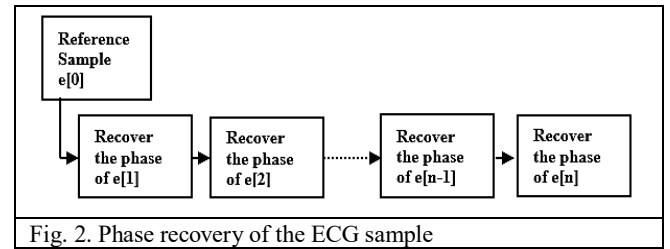
$$y[n] = \sum_{m=0}^n \tilde{e}[m] h[n-m] \quad (5)$$

$$= \tilde{e}[n]h[0] + A_i \quad (6)$$

where

$$A_i = |A_i| e^{j\Phi_{A_i}} = \sum_{n=0}^{i-1} \tilde{e}[n] h[i-n] \quad (7)$$

From Fig. 2 it is noted that the phase of $e[n]$ is calculated using the previous sample $e[n-1]$. The $\Phi_e[n]$ is implemented in FPGA and the required DAST complex samples are generated using MATLAB.



III. FPGA IMPLEMENTATION OF PHASE RECOVERY

The MIT-BIH Arrhythmia database (mitdb) ECG signals are considered as the input [14]. The record 112 (V1) will have the voltage (mV) values with respect to time. A total of 300 samples (equivalent to 1 beat) are considered as the input data. The ECG signal is sampled at 360Hz and has an 11-bit resolution. Hardware results of the design are analyzed using Chip scope in the Xilinx platform.

The FPGA Specifications are FPGA Board: Artix-7 NEXYS 4 DDR, Clock frequency: 100MHz, Xilinx part number: XC7A100T-1CSG324C. It has 15,850 logic slices, 240 DSP slices, and 4,860 Kbits of fast block RAM.

The concept of phase recovery block is based on the propagation of the signal under test through a linear filter or phase recovery filter. The phase value is then calculated using a non-iterative algorithm based upon linearity and causality conditions. Equation (4) is implemented in FPGA. To design the phase recovery block in Xilinx, trigonometric functions like sine, cosine, inverse tangent, and inverse cosine are needed.

These advanced mathematical operations cannot be implemented directly. To implement these functions CORDIC algorithm is very much useful. CORDIC algorithm is used to implement these operations by using IP core in Xilinx ISE 14.4 tool.

A. Computation of Magnitude

The computation of magnitude is performed using the Non-Restoring Square Root algorithm. It is one of the effective ways to implement at a very fast clock rate. It mainly emphasizes the non-restoring on the “partial remainder”, with each iteration. It needs one normal adder/subtractor in each iteration and produces the correct value as well as in the last bit position. Next, using the last bit result, the accurate remainder is achieved immediately without any correction or addition operation. There are two advantages: i) fully pipelined high-performance operations. It may accept a new square-root instruction for each clock cycle with each pipeline stage. It requires a minimum number of gate counts. ii) low-cost operations that require simply a single adder/subtractor for iterative operation.

B. Computation of taninv using CORDIC

The CORDIC algorithm has two modes, vectoring (rectangular to polar) or rotation (polar to rectangular) modes. The vectoring is achieved by selecting α_i , such that θ' converges towards zero. It produces the scaled output Vector $Z_i * (X', Y')$. The rotation mode rotates the vector (X, Y) around the circle till the Y reaches zero. For the input vector (X, Y), magnitude X, and phase θ , are the outputs. It is achieved by choosing α_i in such a way that Y converges to zero. The outcomes of the CORDIC algorithm are corresponding to a vector rotation or vector translation that is scaled by a constant Z_i , where Z_i is the CORDIC scale factor. The concept adopted in [6-7] is applied here to get the required trigonometric functions. Vector rotation is expressed as a sequence of ‘n’ micro-rotations.

$$X' = \prod_{i=1}^n \cos(\text{atan}(2^{-i}))(X_i - \alpha_i Y_i 2^{-i}) \quad (8)$$

$$Y' = \prod_{i=1}^n \cos(\text{atan}(2^{-i}))(Y_i + \alpha_i X_i 2^{-i}) \quad (9)$$

$$\theta' = \prod_{i=1}^n \theta - (\alpha_i \cdot \text{atan}(2^{-i})) \quad (10)$$

$$\alpha_i = (+ \text{ or } -) 1.$$

Where α_i is the direction of rotation.

C. Computation of cosinv using taninv

The computation of the cos inverse function is done using the trigonometric identity

$$\cos^{-1}(x) = \tan^{-1}\left(\frac{\sqrt{1-x^2}}{x}\right) \quad (11)$$

D. Recovery of real and imaginary components

The real and imaginary components of the complex transform are recovered using the phase and magnitude of the samples. Consider the phase θ and the magnitude r , then the trigonometric identity is

$$r * e^{i\theta} = r * (\cos \theta + i * \sin \theta) \quad (12)$$

Hence the real part will be $r * (\cos \theta)$ and the imaginary part will be $r * (\sin \theta)$. Here $r = |\tilde{e}[n]|$ which one is transmitted and $\theta = \Phi_e[n]$ is calculated using a phase recovery block at the receiver side. Using $|\tilde{e}[n]|$ and $\Phi_e[n]$, the DAST ECG samples $\tilde{e}[n]$ is recovered.

E. Modular Hierarchy

The unfiltered signal magnitude and filtered signal magnitude applied to the phase recovery block is shown in Fig. 3. The corresponding real and imaginary portions of the complex numbers are generated. The Verilog code for the phase recovery block is written as in the above hierarchy. The inputs and outputs of the phase recovery block are 16 bits. Initially, input ECG signal samples are scaled by 1000, since the samples will be in mV (which are less than 1). To work in Verilog easily the samples are scaled with 1000 and descaled at the output samples. The input angle i in degree is scaled and descaled to work with the trigonometric functions by 31994. Since the angle ranges from 0 to 360 degrees is very small for working with 16-bit resolution.

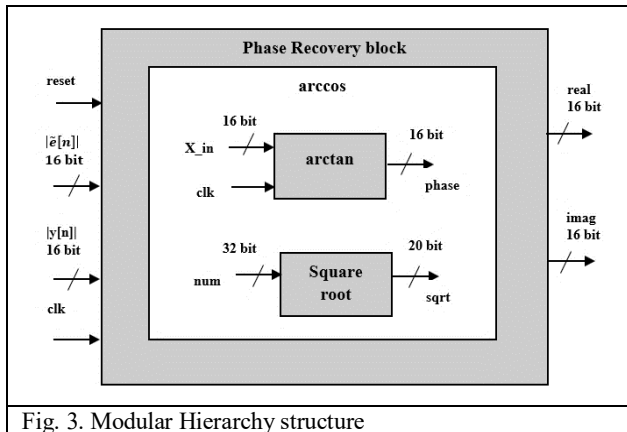


Fig. 3. Modular Hierarchy structure

F. Hardware blocks

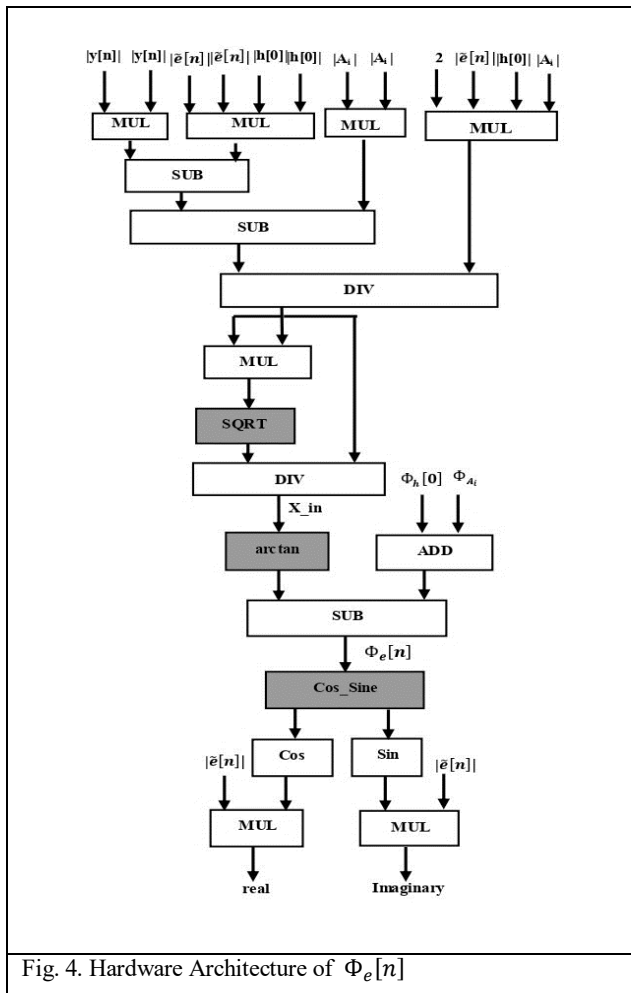


Fig. 4. Hardware Architecture of $\Phi_e[n]$

The different hardware blocks required for the phase recovery implementation are shown in Fig. 4. ADD and SUB indicate the 16-bit adder and subtractors respectively. The 16-bit multiplication blocks (MUL) and 16-bit division blocks (DIV) are also used. The block SQRT i.e. the square root function is implemented using the Non-Restoring Square Root algorithm. Arctan and cos_sine blocks are the trigonometric functions that are implemented using the CORDIC algorithm.

IV. SIMULATION AND RESULTS

All the modules are coded in Xilinx ISE 14.4 Design suite in Verilog HDL, simulated and synthesized using Xilinx Artix-7 FPGA device. Hardware results of the design are analyzed using Chipscope in the Xilinx platform. The filtered and unfiltered DAST ECG coefficient magnitudes are calculated using the Non-Restoring Square Root algorithm. It is implemented and verified with a simple example in Fig. 5. i.e $\sqrt{3^2 + 4^2} = \sqrt{25} = 5$. Each value is scaled by 1000. Cosine & Sine implementation is done with the CORDIC rotational mode in Verilog shown in Fig. 6. The input angle is specified in terms of $(2^{32} \cdot i) / 360$, where i indicates the input phase angle in degree. The output is scaled with a gain of 32000. Arccos is implemented using arctan by CORDIC vectoring mode in Verilog is shown in Fig. 7. $\arccos(x) = \arctan(\sqrt{1-x^2} / x)$. Input angle is specified in terms of $x \cdot 1000$, where x signifies input in radian. This gain is compensated in $(\sqrt{1-x^2} / x)$. The phase output is scaled with a gain of 342068275. The reconstructed complex DAST ECG sample is $esr + i esi$ recovered using the phase recovery block shown in Fig. 8. i.e $8 + i 619$ is recovered from two magnitudes. The hardware result for phases recovery block using chipscope is shown in Fig.9. The required number of adders, multipliers, and clock cycles are reported in Table I. For computing phase recovery for N samples, N + D clock cycles are required. Where D represents initial delay while computing trigonometric functions (sine, cosine, and arccos) using the CORDIC algorithm (Pipeline).

TABLE I. COMPUTATIONAL AND HARDWARE COMPLEXITY

Required Hardware Blocks	Number of blocks	Latency (cycles)
Adders / Subtractors	732	N + D
Multipliers	16	
Shifts	88	
Table lookup	44	

The hardware utilization details for the phase recovery block are reported in Table II.

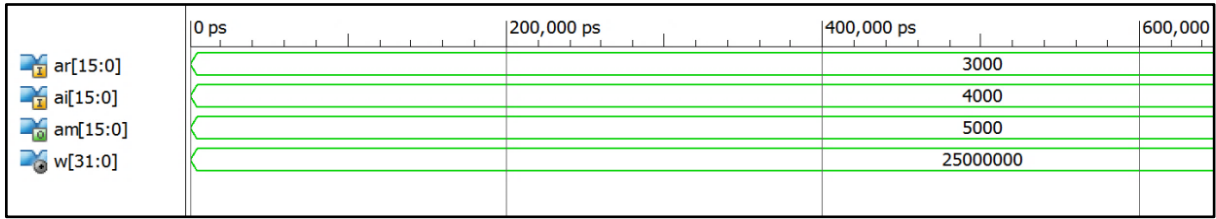


Fig. 5. Xilinx Simulation results for square root

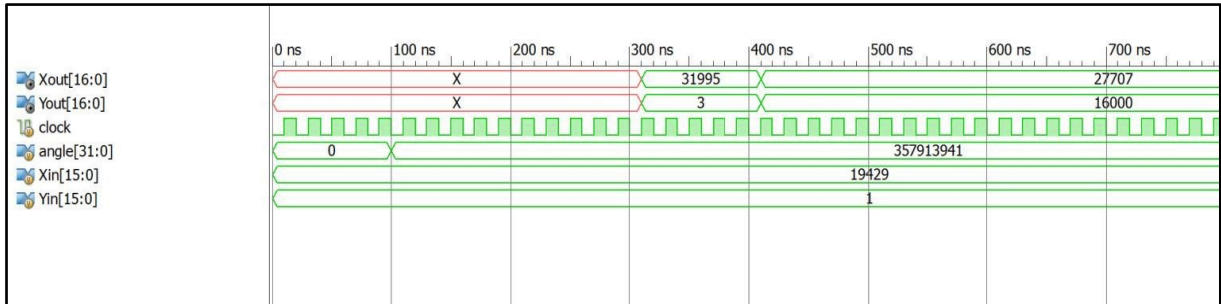


Fig. 6. Xilinx Simulation results for sine and cosine

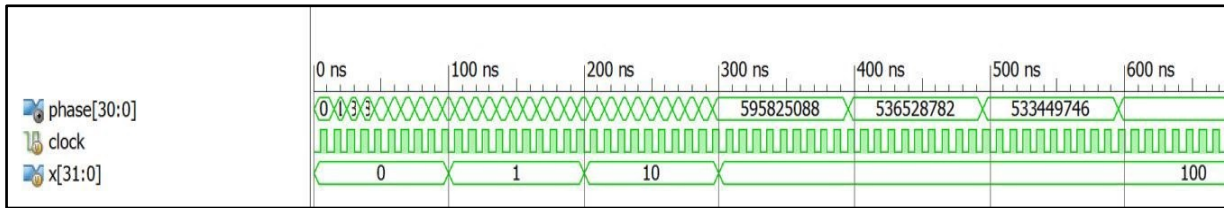


Fig. 7. Xilinx simulation results for cosinv

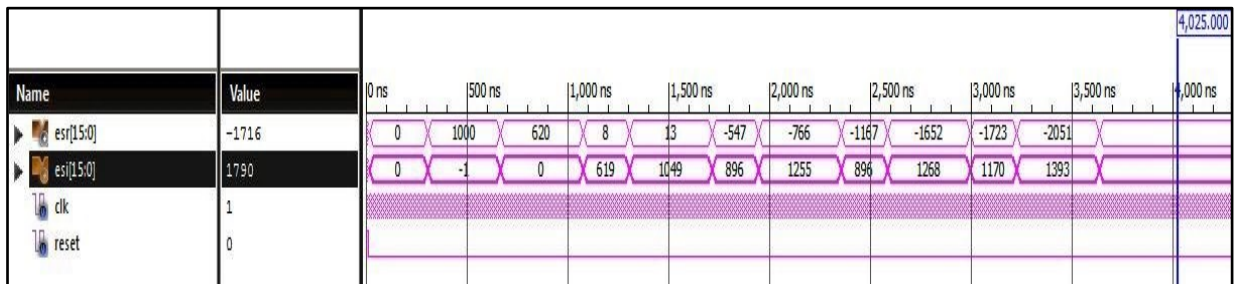


Fig. 8. Xilinx simulation result for phase recovery block

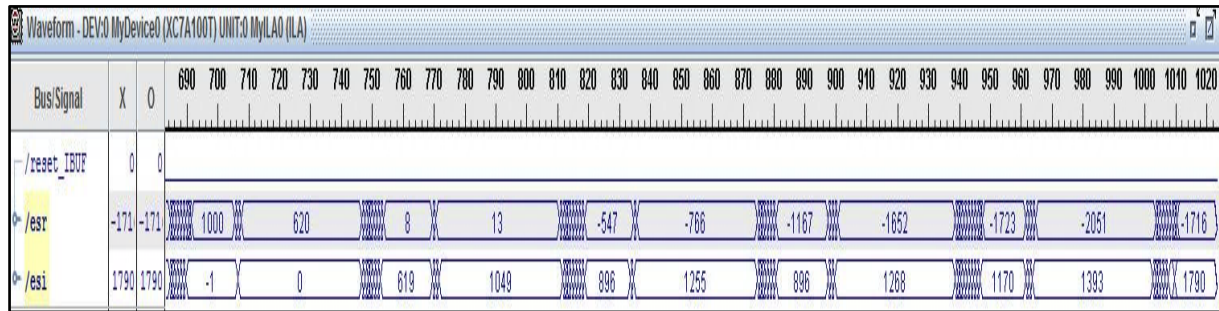


Fig. 9. Hardware result for phase recovery block using Chipscope

TABLE II. RESOURCES UTILIZATION SUMMARY IN ARTIX-7 FPGA DEVICE

Artix-7 resources	Total Available	Used	Utilization factor in %
No. of slice registers	126800	2105	1.66
No. of slice LUTs	63400	15391	24.27
No. of full used LUT-FF pairs	17009	143	0.84
No. of bonded IOBs	210	46	21.9
No. of Block RAM/FIFO	135	1	0.74
No. of BUFG/BUFGCTRLS	32	2	6.25
No. of DSP48E1s	240	51	21.25

V. CONCLUSION

The complex transforms are reported in the literature for improving the compression factor and reducing the reconstruction errors. The one-dimensional DAST is one of the complex transform used in ECG signal compression. Instead of sending real and imaginary parts separately, only the magnitudes are transmitted. At the receiver side, the necessary phase is recovered using these magnitudes. The proposed phase recovery block is effectively implemented in the ARTIX-7 FPGA board and the results are compared with MATLAB simulation results which is not reported in the literature. Advanced mathematical functions like square root and trigonometric functions are implemented using the CORDIC algorithm. Since it uses shift registers, adders, and look-up tables the Hardware requirement and the cost of a CORDIC processor are less. Working with the smaller values (less than one) in Verilog is a complex task. The problem is solved by multiplying the input samples by the scaling factor leads to error. The floating-point implementation may be considered in the future. FPGA implementation of real-Time ECG Signal processing can be carried out instead of the ECG signal database.

REFERENCES

[1] Thilagavathy R, Venkataramani B, "A Novel ECG Signal Compression using Wavelet and Discrete Anamorphic Stretch Transforms", *Biomedical Signal Processing and Control* (Elsevier), Volume 71, Part B, January 2022, 102773. doi.org/10.1016/j.bspc.2021.102773.

[2] Thilagavathy R, Venkataramani, "Optimization of Discrete Anamorphic Stretch Transform and Phase Recovery Techniques for ECG Signal Compression", *IETE Journal of Research* (Taylor & Francis) – doi.org/10.1080/03772063.2021.2012281

[3] R. Thilagavathy, B. Venkataramani, "ECG Signal Compression using Discrete Anamorphic Stretch Transform", 5th International Conference on Microelectronics, Circuits & Systems, organized by Applied Computer Technology, 2018. ISBN: 81-85824-46-1

[4] M. H. Asghari and B. Jalali, "Discrete Anamorphic Transform for Image Compression", 1070-9908 © 2014 IEEE.

[5] Mohammad H. Asghari and Jose Azana, "Self-reference temporal phase reconstruction based on causality arguments in linear optical filters", *CLEO Technical Digest* © OSA 2012.

[6] Vladimir Hahanov, Olga Melnikova, Dmitriy, Melnik, Philat Levchenko. "CAD Tools for CORDIC IP Cores Generation", 2006 International Conference - Modern Problems

[7] Ray Andraka, "A survey CORDIC algorithms for FPGA based computers", Andraka Consulting Group .inc 2016

[8] DS 249 Xilinx LogiCORE™ IP CORDIC core implements a generalized coordinate rotational digital computer (CORDIC) algorithm.

[9] V. Torres, J. Valls and M.J. Canet, "Optimised CORDIC-based atan2 computation for FPGA implementations", *Electronics Letters*, Volume: 53, Issue: 19, 9 14 2017)

[10] Supriya Aggarwal, Pramod K. Meher, And Kavita Khare, "Concept, Design, And Implementation of Reconfigurable CORDIC", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 24, No. 4, April 2016

[11] Rajkumar Ramadoss, Mehran Mozaffari Kermani, Reza Azarderakhsh, "Reliable Hardware Architectures of the CORDIC Algorithm With a Fixed Angle of Rotations", *IEEE Transactions On Circuits And Systems—ii: Express Briefs*, Vol. 64, No. 8, August 2017

[12] P. K. Meher, J. Valls, T.-B. Juang, K. Sridharan, and K. Maharatna. "50 years of CORDIC: Algorithms, architectures, and applications", *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 9, pp. 1893–1907, Sep. 2009.

[13] FPGA design flow, Copyright © 2008, Xilinx® Inc.

[14] The MIT-BIH Arrhythmia Database, 2005. Feb., [Online]. Available: <http://www.physionet.org/physiobank/database/mitdb>.

[15] Jay P. Lim, Matan Shachna, "Approximating trigonometric functions for posits using the CORDIC method "ACM International Conference on Computing Frontiers, pp. 19-28, 2020. DOI:10.1145/3387902.3392632

[16] Ravi Mogili, Raju Katru, Kandukuri Shobha, "Efficient Pipelined CORDIC Architecture for Generation of Sine and Cosine Function", *International Journal of Innovative Research in Technology*, Volume 4, Issue 8, pp: 447 – 452, 2018

[17] S. Bhukya and S. C. Inguva, "Design and Implementation of CORDIC algorithm using Integrated Adder and Subtractor," *International Conference for Convergence in Technology (I2CT)*, pp. 1-5, 2021. doi: 10.1109/I2CT51068.2021.9418002.

VLSI Implementation of a Real-time Modified Decision-based Algorithm for Impulse Noise Removal

1st Pradyut Kumar Sanki

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 pradyut.s@srmmap.edu.in

2nd Vasudeva Bevara

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 vasudeva_b@srmmap.edu.in

3rd Medarametla Deepthi Supriya

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 medarametla_deepthi@srmmap.edu.in

4th Devireddy Vignesh

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 devireddy_vignesh@srmmap.edu.in

5th Peram Bhanu Sai Harshath

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 peram_bhanu@srmmap.edu.in

6th Sravya Kuchina

Electronics & Communication Engineering
SRM University AP (UGC)
 Guntur, India
 sravya_kuchina@srmmap.edu.in

Abstract—In this paper, a real-time impulse noise removal (RTINR) algorithm and its hardware architecture are proposed for denoising images corrupted with fixed valued impulse noise. A decision-based algorithm is modified in the proposed RTINR algorithm where the corrupted pixel is first detected and is restored with median or previous pixel value depending on the number of corrupted pixels in the image. The proposed RTINR architecture has been designed to reduce the hardware complexity as it requires 21 comparators, 4 adders, and 2 line buffers which in turn improve the execution time. The proposed architecture results better in qualitative and quantitative performance in comparison to different denoising schemes while evaluated based on the following parameters: PSNR, IEF, MSE, EKI, SSIM, FOM, and visual quality. The proposed architecture has been simulated using the XC7VX330T-FFG1761 VIRTEX7 FPGA device and the reported maximum post place and route frequency is 360.88 MHz. The proposed RTINR architecture is capable of denoising images of size 512×512 at 686 frames per second. The architecture has also been synthesized using UMC 90 nm technology where 103 mW power is consumed at a clock frequency of 100 MHz with a gate count of 2.3K (NAND2) including two memory buffers.

Index Terms—Impulse noise, image denoising, filtering algorithms, image quality, PSNR, field programmable gate arrays.

I. INTRODUCTION

DIGITAL images are commonly corrupted with impulse noise due to faulty sensing, acquisition, transmission, and reception, leading to loss of information [1], [2]. The impulse noise is randomly distributed in the image, and it can either be fixed-valued or random-valued in nature. The fixed-valued impulse noise is referred to as salt-and-pepper noise, for which the image pixels can be altered into either black (graylevel = 0) or white (graylevel = 255) [3], [4]. Median filtering is widely used to eliminate such disturbances while maintaining image quality [5], [6]. The Standard Median Filter (SMF) alters both corrupted as well as uncorrupted pixels resulting in blurring of the output image [6]–[12]. To minimize blurring an improve image quality, the noisy pixels are detected and restored, while other pixels are kept unchanged. The Adaptive Median Filter (AMF) adjusts the processing window depending upon the noise density, however bigger window sizes result in increased blurring in the output image [13], [14]. Switching Median Filters (SWMF) restore noisy pixels by an immediate neighbour or the median value of the pixel neighbourhood while keeping the remaining pixels intact [15]–[20]. They perform well under various conditions, but are unable to restore images at high noise densities due to difficulties in determining impulse magnitude and impulse strength at sharp edges in the image. SWMFs

TABLE I
3 × 3 PROCESSING WINDOW

$P(i-1, j-1)$	$P(i-1, j)$	$P(i-1, j+1)$
$P(i, j-1)$	$P(i, j)$	$P(i, j+1)$
$P(i+1, j-1)$	$P(i+1, j)$	$P(i+1, j+1)$

with boundary discriminative noise detection (BDND) tackles this problem at the cost of increased execution time [21]–[23]. On the other hand, decision based algorithms (DBA) have lower complexity and can be executed in lesser time, but result in streaking effects at high noise densities [24]–[31]. A variety of techniques such as fuzzy based adaptive median filtering [32]–[39], genetic programming based filters [40], hypergraph-based algorithms [41], partial differential equation based filters [42], linear mean-median filters [30], noise elimination and edge preservation filters [43], interpolation based approaches [44], classified regularization approaches [45], and non-uniform sampling and autoregressive modelling based super-resolution techniques [46], Min-Max Average Pooling Based Filter for Impulse Noise Removal [47], An Adaptive Weighted Min-Mid-Max Value Based Filter [48], & Conditional Min Pooling and Restructured Convolutional Neural Network [49] have been used for impulse noise removal. These techniques involve significant computation on the corrupted images which is performed offline, making it difficult to get denoised images in real time. For real-time embedded applications involving impulse noise reduction from images, hardware implementations of such filtering algorithms is essential. Existing hardware implementations for impulse noise reduction are effective, but are not efficient occupying significantly large area [50]–[59]. We propose a modified decision based real time impulse noise removal (RTINR) algorithm along with its architecture to realize an efficient real time hardware design for portable systems. The algorithm and its implementation are evaluated based on parameters, such as, Peak Signal to Noise Ratio ($PSNR$), Mean Square Error (MSE), Image Enhancement Factor (IEF), Edge Keeping Index (EKI), Figure of Merit (FOM), Structural Similarity Index between Images ($SSIM$), and Mean Structural Similarity Index between Images ($MSSIM$) [1], [60], [61]. Above all the computation complexity is optimized and the processing time for the proposed algorithm is executed on a system with an Intel Core i7 processor at 3.1 GHz, equipped with 8GB RAM, running MATLAB 2012a. The remaining paper is organized as follows: Section II describes the proposed algorithm for the detection and reduction of impulse noise. The hardware architecture of the proposed algorithm and its VLSI implementation is detailed in Section III, Section IV deals with the performance analysis of the reconstructed image and Section IV draws the conclusions.

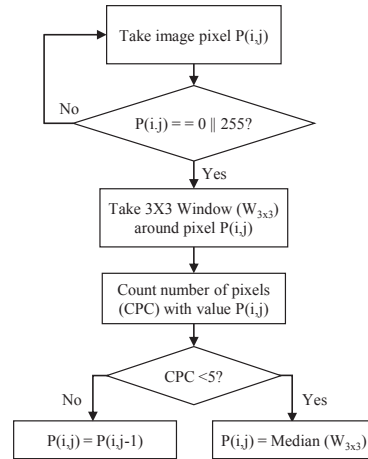


Fig. 1. Flowchart of the proposed algorithm (RTINR) for removing impulse noise from an image.

II. PROPOSED RTINR

Assume pixel values are changed in image I due to salt and pepper noise. For each pixel $P(i, j)$ in I , RTINR first detects the noisy pixel. If it is noisy, a window of size 3×3 ($W_{3 \times 3}$) is considered in Table I. Noisy pixel value is restored by the proposed RTINR depending on the number of corrupted pixel equals to $P(i, j)$ in $W_{3 \times 3}$. The algorithm and the flow chart of RTINR are given in Algorithm. 1 and in Fig. 1 respectively.

Algorithm 1 The RTINR Algorithm

- 1: **for** each pixel in the image **do**
 - 2: **if** pixel is noisy ($P(i, j) == 255 \parallel P(i, j) == 0$) **then**
 - 3: Take 3×3 Window ($W_{3 \times 3}$) around $P(i, j)$;
 - 4: Initialize Corrupted Pixel Count (CPC)
 $CPC \leftarrow 0$;
 - 5: **for** each pixel in $W_{3 \times 3}$ equal to $P(i, j)$ **do**
 - 6: Increment CPC
 $CPC \leftarrow CPC + 1$;
 - 7: **end for**
 - 8: **if** ($CPC < 5$) **then**
 - 9: Replace Pixel with Median Value
 $P(i, j) \leftarrow Median(W_{3 \times 3})$;
 - 10: **else**
 - 11: Replace Pixel with Previous Value
 $P(i, j) \leftarrow P(i, j - 1)$;
 - 12: **end if**
 - 13: **else**
 - 14: Leave Pixel Value Unchanged
 $P(i, j) \leftarrow P(i, j)$;
 - 15: **end if**
 - 16: **end for**
-

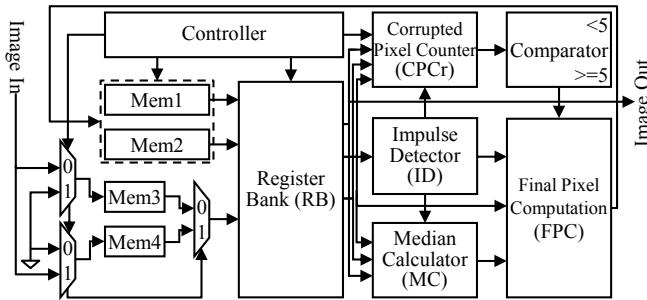


Fig. 2. Block level diagram of the VLSI architecture of RTINR.

III. VLSI IMPLEMENTATION OF THE PROPOSED ALGORITHM

Toward the objective of fast image processing for real time embedded applications, VLSI implementation of the RTINR architecture is proposed. Fig. 2 shows the block level diagram of the VLSI architecture of the proposed algorithm. The main building blocks of the architecture consist of Register Bank (RB), Impulse Detector (ID), Corrupted Pixel Counter (CPCr), Final Pixel Computation (FPC), Median Calculator (MC) and Controller. In the proposed RTINR three clock cycles are initially used to buffer the nine registers (R1 to R9) with image pixel values for $W_{3 \times 3}$ forming. After the latency of three clock cycles noise detection and restoration of each pixel in the corrupted image are taken place at each clock pulse. During the operation of noise cleaning, every processing pixel is read from, buffered with restored value into memory at each clock pulse. The hardware of RTINR requires only 21 comparators (CMP) and 4 adders along with two memory line buffer of size 512×8 each. Arithmetic operations like comparison and addition make it notably less complex in computation as compared to low-complexity noise removal (LCNR) technique [54]. A brief description of each block of RTINR for removing salt and pepper noise from image is presented below.

A. Register Bank (RB)

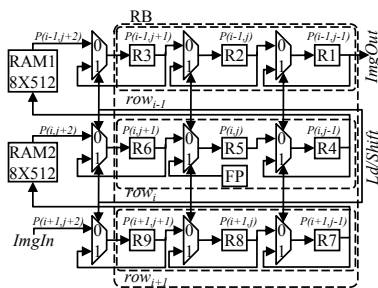


Fig. 3. Architecture of register bank in RTINR.

The register bank is made up of nine registers, R1 - R9, that are connected serially storing the pixel values, $P(i-1,j-1)$

- $P(i+1,j+1)$ of the current processing window $W_{3 \times 3}$. The architecture of the register bank is shown in Fig. 3 to process the entire image corrupted with salt and pepper noise. Denoising process starts when last pixel value $P(i+1, j+1)$ is buffered into the register R9. Current processing pixel $P(i, j)$ stored in register R5 is used to detect the noisy pixel and the type of corruption caused by either salt or pepper noise is also confirmed with the help of impulse detector (ID). The remaining pixels, stored in registers are then used simultaneously to determine the number of corrupted pixel as same type of $P(i, j)$ by subsequent corrupted pixel counter (CPCr). The replacement of the corrupted pixel $P(i, j)$ at register R5 with either median value or previous pixel $P(i, j-1)$ value is performed by final pixel computation (FPC).

In general, the restoration of corrupted pixel values in row_i for the subsequent image denoising process is performed by simultaneous read and write operations of RAM1 and RAM2. Image pixels are simultaneously read out from RAM1 as row_{i-1} and the values of row_i are written back into RAM1. In parallel to this, the values of pixel row_i are read out from RAM2 and at the same time the value of row_{i+1} are written back into RAM2. The data of third row is coming directly from input image as row_{i+1} .

B. Impulse Detector (ID)

Fig. 4 presents the architecture for impulse detector (ID) designed to detect the current processing pixel $P(i, j)$ disturbed with either salt or pepper noise in the proposed RTINR. Logical circuit diagrams for G1 and G2 in Fig. 4b and Fig. 4c are designed to determine grey level value of "255" and "0" for salt and pepper noise detection respectively. The type of noise is also determined from the output SN and PN of G1 and G2 respectively. Image pixel affected with salt noise is determined if all the bits of the register R5 is logically 1 i.e., the output SN of G1 is logic high. On the other hand image pixel is considered to be corrupted with pepper noise if all the bits of the register R5 attains logic value of 0 i.e., the output PN of G2 is logic high. One 5 input AND gate, one 4 input AND gate, one 5 input OR gate, one 4 input NOR gate and a 2 input OR gate are used to design the above architectures.

C. Corrupted Pixel Counter (CPCr)

The architecture for counting number of corrupted pixels (CPCr) or corrupted pixel counter (CPCr) is shown in Fig. 5. The output SN and PN from ID is used to determine the current processing pixel $P(i, j)$ corrupted with either salt or pepper noise. All the registers except R5 in RB are combined individually to perform bitwise AND and NOR operation for salt and pepper noise detection. Four adders are used to count the total number of pixels corrupted with either salt if SN is logic 1 or pepper if PN is logic 1.

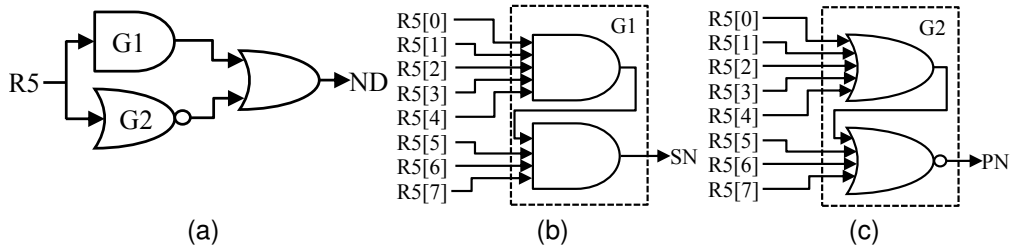


Fig. 4. (a) Architecture for impulse detector, (b) salt noise (SN) detector and (c) pepper noise (PN) detector in RTINR.

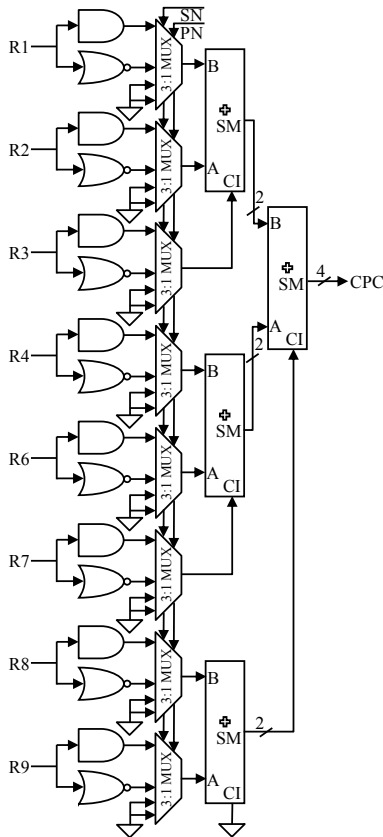


Fig. 5. Architecture for counting number of corrupted pixels (CPCr) or corrupted pixel counter (CPCr).

D. Final Pixel Computation (FPC)

The value to be restored to the current processing pixel, $P(i, j)$, is determined based on the number of corrupted pixels (CPC) in the processing window. A simplified boolean expression is implemented to realize the output (RPSel) which selects either median or neighbour pixel value. If the number of corrupted pixels is less than 4 ($CPC < 4$), the processing pixel is replaced with the median pixel value ($P(i, j) = Median(W_{3 \times 3})$) otherwise ($CPC \geq 5$) by

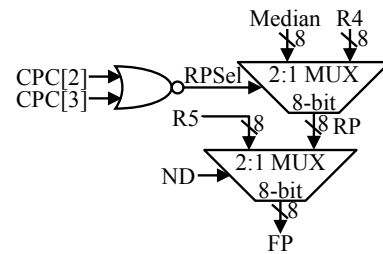


Fig. 6. Architecture for Final Pixel Computation (FPC)

Input Sorted Combination	Output of the Comparators						Select Line		
	CMP1 agb	alb	CMP2 bgc	blc	CMP3 agc	alc	agb	bgc	agc
$A > B > C$	1	0	1	0	1	0	1	1	1
$A > C > B$	1	0	0	1	1	0	1	0	1
$B > A > C$	0	1	1	0	1	0	0	1	1
$B > C > A$	0	1	1	0	0	1	0	1	0
$C > A > B$	1	0	0	1	0	1	1	0	0
$C > B > A$	0	1	0	1	0	1	0	0	0

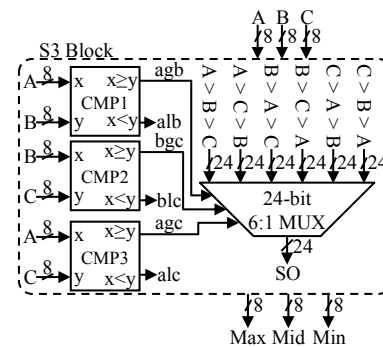


Fig. 7. Truth Table and architecture for sorting three numbers

previous value ($P(i, j) = P(i, j-1)$) in the processing window. Ultimately the final pixel value (FP) for each processing pixel is selected depending upon the value of ND which is shown in Fig. 6.

E. Median Calculator (MC)

Median value calculation is performed if CPC value is less than four ($CPC < 4$). Three steps are followed sequentially

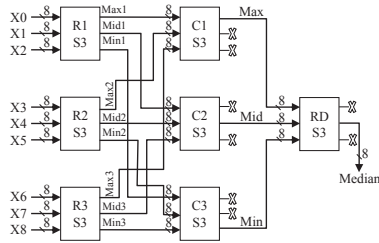


Fig. 8. Architecture for median calculation in RTINR.

to calculate median value of 9 pixels stored in RB . First step is the sorting of rows of $W_{3 \times 3}$, columns of $W_{3 \times 3}$ are sorted in second step and in third step right diagonal sorting of 3 pixels [24] are performed to get the middle (Mid) value which is median of the 9 pixels of the processing window $W_{3 \times 3}$. Sorting of 3 numbers (Sorting Block or S3) is the basic building block shown in Fig. 7 required to determine median value. To achieve this a look up table based six sorting combinations for three numbers are presented in truth table Table III-D and is given as inputs to the 24-bits 6:1 MUX. Three selection lines of this MUX are the three binary outputs of three comparators namely CMP1, CMP2 and CMP3 used to determine the sorted output (SO). Therefore, seven S3 blocks are required to calculate the median value of current processing window $W_{3 \times 3}$. The architecture is shown in Fig. 8 implemented for calculating the median value of $W_{3 \times 3}$ in RTINR.

F. Controller

The controller module is designed to control the overall operations inside the architecture. The controller governs the read and write instructions for the data in the register bank is governed by the controller. The schedule for writing into and reading from memories RAM1 and RAM2 is guided by the controller module. In the architecture for proposed RTINR, an input reset signal initiates the process of impulse noise removal by activating the controller which directs all the steps from taking in a noisy image *ImageIn* to giving out a restored image *ImageOut*.

IV. IMPLEMENTATION RESULTS AND COMPARISONS

The qualitative and quantitative performances of the proposed algorithm RTINR is examined and analysed based on the following parameters: *PSNR*, *IEF*, *MSE*, *EKI*, *SSIM* and *FOM*. Five different gray scale images of size 512×512 such as Lena, Baboon, Boat, Goldhill, and Barbara are considered as reference images for testing.

To validate our proposed work we have taken five reference images like Lena, Baboon, Boat, Goldhill, and Barbara of size 512×512 with 8-bits gray value for simulation. The performance in terms of above mentioned parameters of the proposed algorithm has been evaluated and compared with the

existing noise cleaning algorithms at different noise densities in Fig. 9, Fig. 10, Fig. 11, Fig. 12 and Fig. 13 respectively. It is clear from the graph of Fig. 9 that upto 45% noise density, the *PSNR* value of our RTINR is greater than reported *PSNR* of existing methods. Other parameters like *IEF* and *SSIM* are also significantly higher compared to other noise removal techniques at 45% and 80% noise density respectively. It is also observed that proposed RTINR technique has lower *MSE* value with respect to other methods even upto 80% noise density. However, in comparison to other algorithms, there is no significant difference is observed in *FOM* at upto 60% noise density except CWMF method but at higher noise density RTINR outperforms the other methods. *EKI* of RTINR is very high with respect to SMF only. It also reveals that the RTINR outperforms the median filter based design not only at low noise density but also at higher noise density. It is observed from Fig. 10, Fig. 11, Fig. 12 and Fig. 13 that RTINR produces better quality of image even from different varieties of images corrupted with impulse noise. The performance of the proposed algorithm (RTINR) has also been verified with zero noise density of Lena image which is compared with different filtering algorithms in Table III. It is observed that RTINR has the capability even to restore noise free image without disturbing its pixel value. Fig. 9 shows the results of *PSNR* of 8-bits gray image of Lena (512×512) corrupted by impulse noise with a noise density varied from 10% to 90% after simulation in hardware of RTINR is compared with the reported results of SMF [7], decision based algorithm (DBA) [24], decision based adaptive filtering method [30], decision based switching median filtering (DBSMF) [31], fuzzy based adaptive switching median filtering [36], novel decision based adaptive weighted and trimmed median filter [28], iterative adaptive fuzzy filter using alpha-trimmed mean (IAFFUATM) [39], enhanced edge preserving impulse noise removal technique (EEPINRT) [56] and advanced modified decision based unsymmetric trimmed median filtering method [29]. In [30] the reported *PSNR* is greater than RTINR but the execution time of 20s is quite high. The architecture reported in [59] suffers from initial latency of 25 clock cycles and each pixel restoration is performed after 5 clock cycles. In the proposed RTINR architecture each pixel in the test image is restored in each clock cycle with a initial delay of 3 clock cycles. As the noise density increases IAFFUATM [39] gives better result but the iterations consume more time during filtering. Hardware implementation by Chuang *et al.* [53] for removal of impulse noise in RSEPD and SEPD is reported with a minimum number of false detection of corrupted image whereas in the proposed RTINR, the occurrence of false detection is zero. Although the platforms are different still two buffer lines of size 512×8 each and comparatively less area are required to implement it on hardware. The speed of the operation is also higher compared to other image denoising hardwares are tabulated in Table II. Execution times

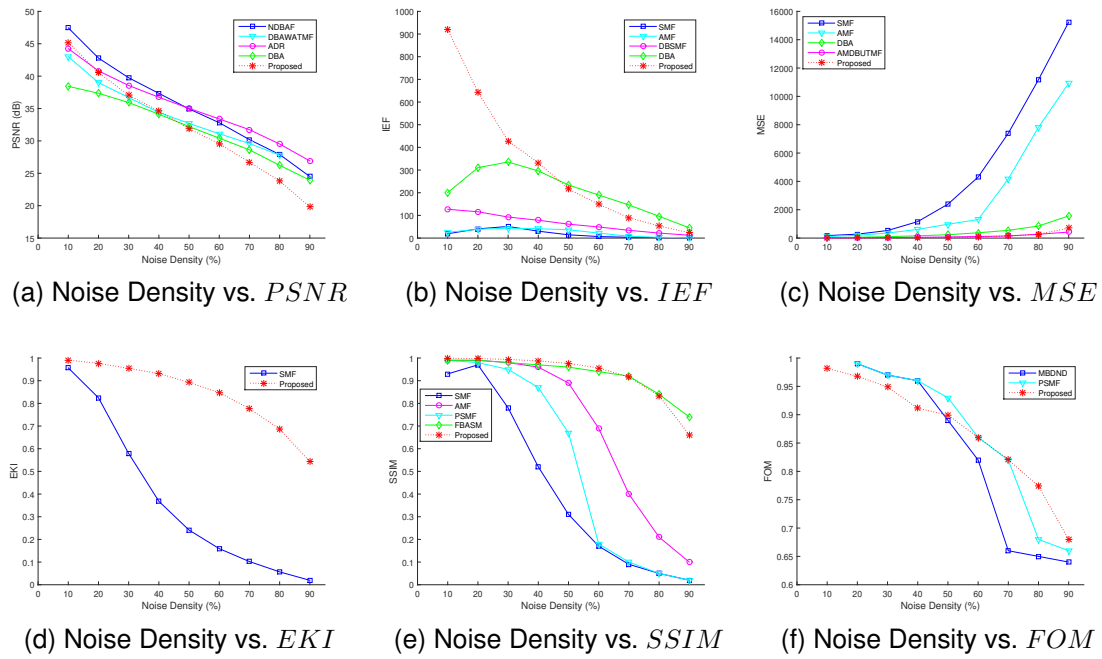


Fig. 9. Comparison of features of restored image in terms of $PSNR$, IEF , MSE , EKI , $SSIM$ and FOM for the 512×512 gray images of Lena with different noise densities using Matlab

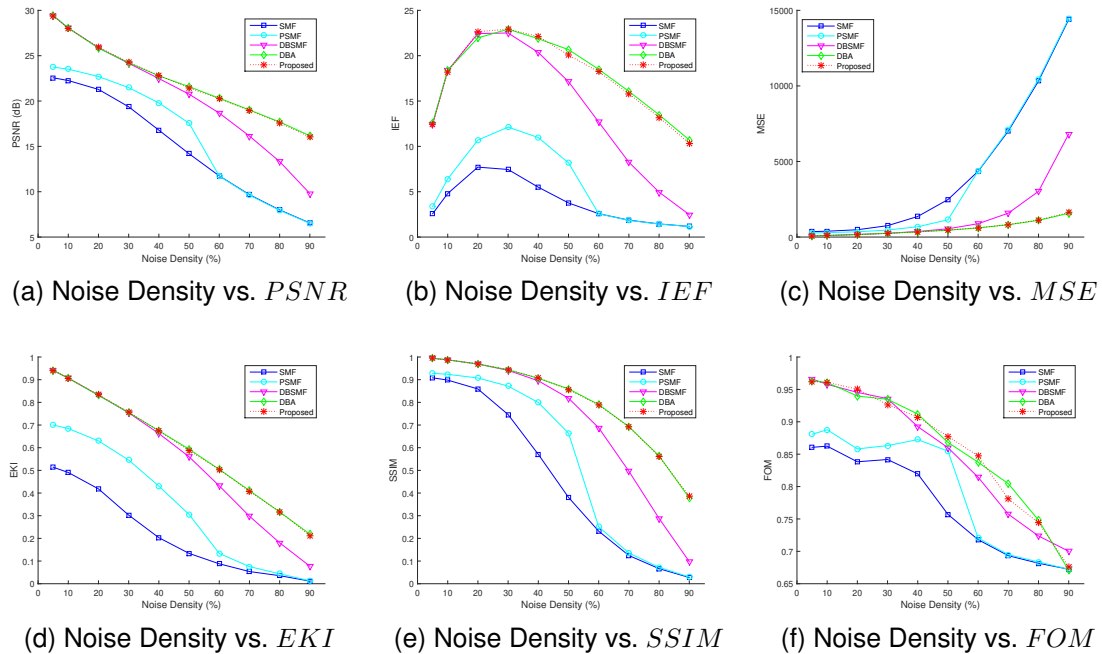


Fig. 10. Comparison of features of restored image in terms of $PSNR$, IEF , MSE , EKI , $SSIM$ and FOM for the 512×512 gray images of Baboon with different noise densities using Matlab

of different denoising algorithms are plotted with respect to different noise densities in Fig. IV which proves that RTINR can produce faster result. It is observed from Fig. IV that RT-

INR has higher $PSNR$ value compared to existing hardware based denoising techniques upto 30% noise density. Table IV shows the superiority of our RTINR in terms area and speed.

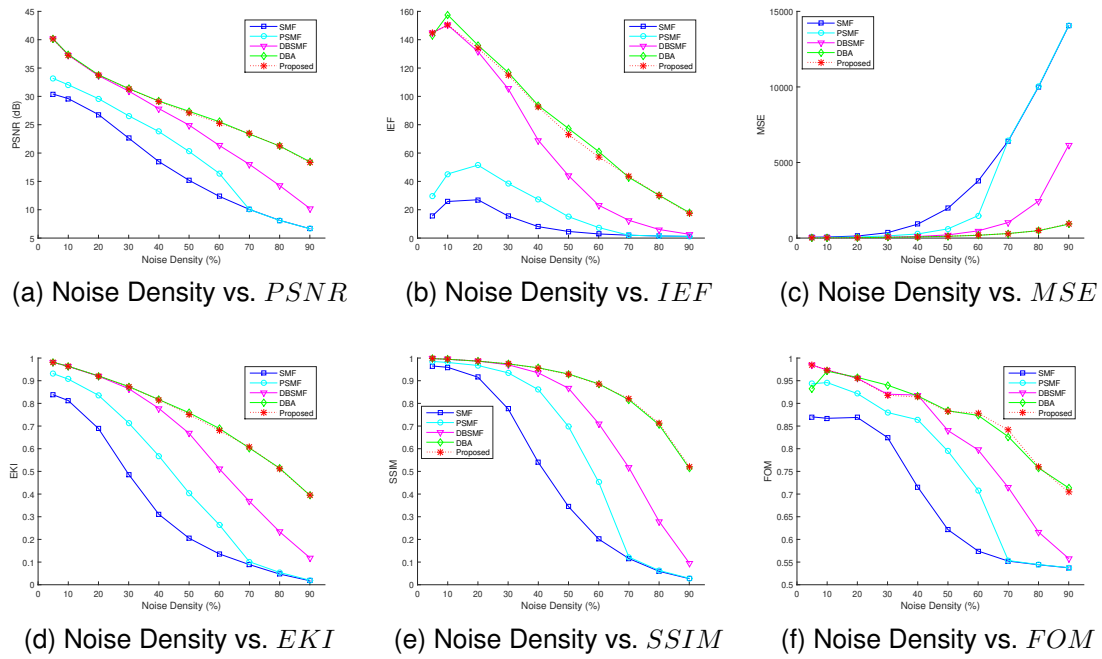


Fig. 11. Comparison of features of restored image in terms of $PSNR$, IEF , MSE , EKI , $SSIM$ and FOM for the 512×512 gray images of Boat with different noise densities using Matlab

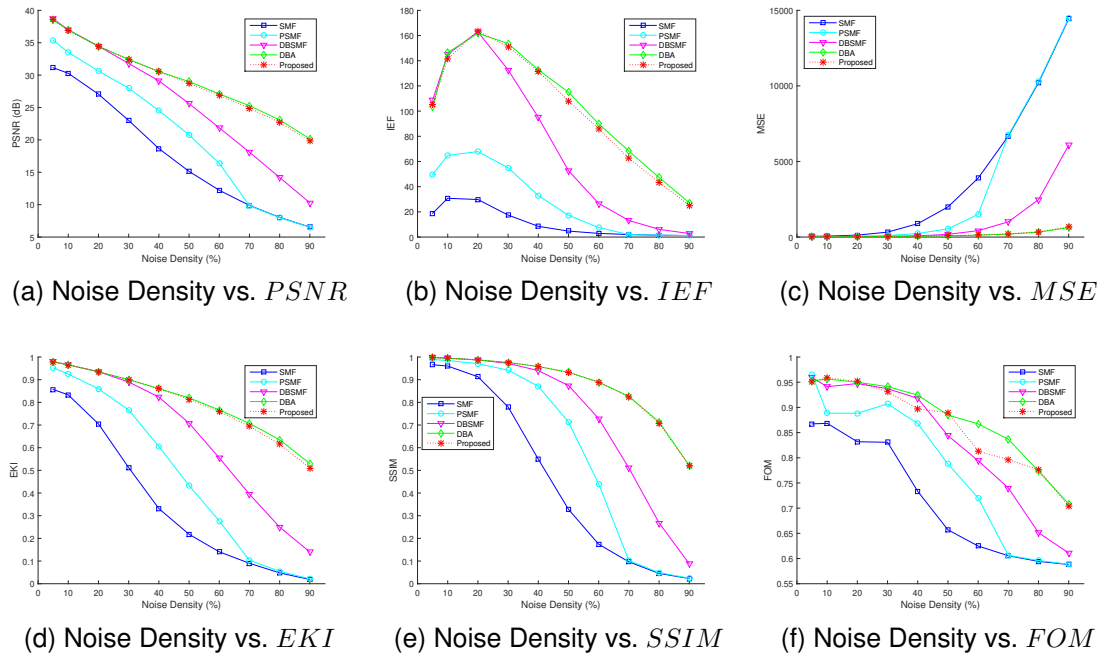


Fig. 12. Comparison of features of restored image in terms of $PSNR$, IEF , MSE , EKI , $SSIM$ and FOM for the 512×512 gray images of Goldhill with different noise densities using Matlab

The proposed architecture has also been simulated using XC7VX330T-FFG1761 VIRTEX7 device and the reported maximum frequency and device utilization after post place

and route synthesis using Xilinx ISE tool with version 14.7 has been tabulated in Table IV. The total power consumption of RTINR using XC7VX330T-FFG1761 FPGA is 178.52 mW

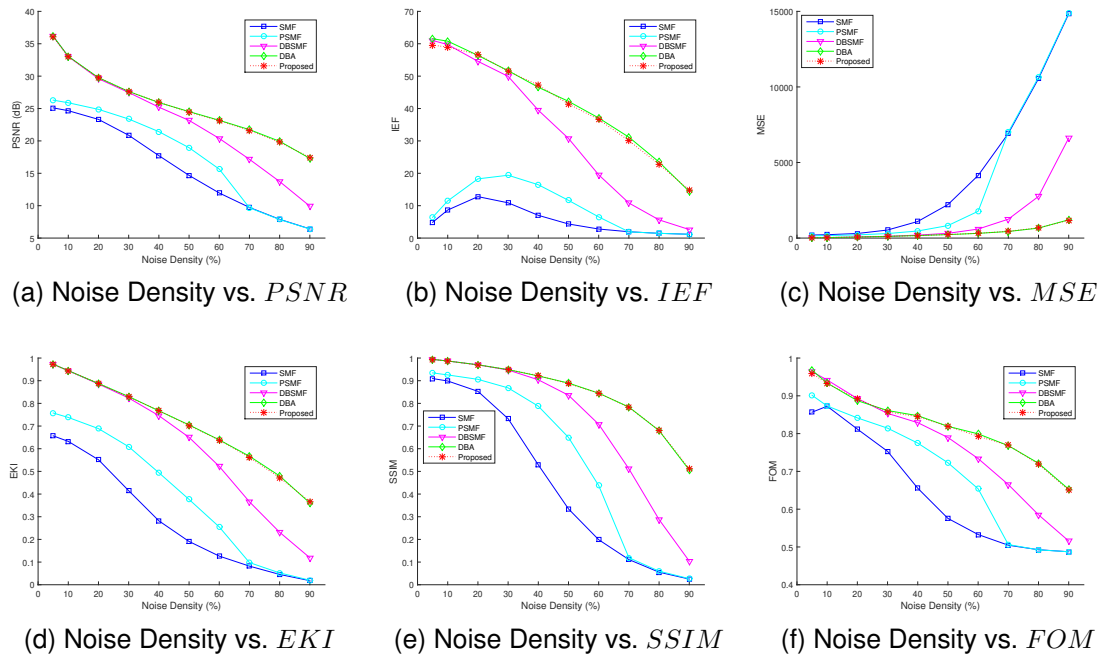


Fig. 13. Comparison of features of restored image in terms of $PSNR$, IEF , MSE , EKI , $SSIM$ and FOM for the 512×512 gray images of Barbara with different noise densities using Matlab

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT IMPULSE NOISE DETECTION AND REDUCTION HARDWARE ARCHITECTURES

Algorithm	Chen <i>et al.</i> [53]	Chen <i>et al.</i> [53]	Matsubara <i>et al.</i> [54]		Lien <i>et al.</i> [57]		Proposed RTINR	
Image Size	512×512	512×512	640×480		512×512		512×512	
Window Size	3×3	3×3	3×3	5×5	3×3		3×3	
Device (Altera)	STRATIX EP1S25	STRATIX EP1S25	CycloneII EP2C20F484C7	CycloneII EP2C20F484C7	STRATIX EP1S25	CycloneII EP2C20F484C7	STRATIX EP1S25	CycloneII EP2C20F484C7
Area Logic Cells	709	1487	513	1397	1743	1709	197	200
Maximum Frequency	162.6 MHz	72.3 MHz	129.58 MHz	93.76 MHz	144.11 MHz	140.37 MHz	187.9MHz	333.33 MHz

TABLE III
COMPARATIVE RECONSTRUCTED RESULTS OF LENA IMAGE CORRUPTED WITH IMPULSE NOISE WITH ZERO NOISE DENSITY

Algorithm	Number of False Detected Pixels	PSNR(dB) of Reconstructed Image
Zhang <i>et al.</i> [17]	284	43.1
Hsia <i>et al.</i> [51]	271	42.81
Andreadis <i>et al.</i> [52]	1	74.34
Luo <i>et al.</i> [62]	226	48.83
Ng <i>et al.</i> [21]	0	No Change
Luo <i>et al.</i> [32]	0	No Change
Srinivasan <i>et al.</i> [24]	0	No Change
Crnojevic <i>et al.</i> [40]	3238	43.38
Chen <i>et al.</i> [53]	1	74.34
Chen <i>et al.</i> [53]	1	74.34
Proposed RTINR	0	No Change

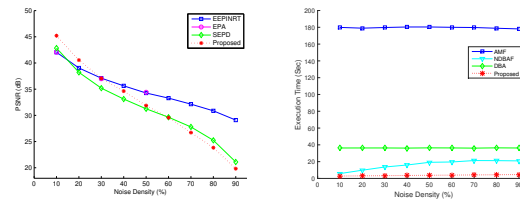


Fig. 14. Comparison of features of restored image in terms of $PSNR$ and execution time for the 512×512 gray images of Lena with different noise densities using Hardware

The architecture has also been implemented using UMC 90 nm CMOS process in Synopsys Design Compiler where 103 mW power is consumed at clock frequency of 100 MHz. The

total gate count of RTINR is 2.3K gates (NAND2) along with two memory buffer of size 512×8 each.

V. CONCLUSION

In this paper, a real time modified decision based algorithm and its architecture for removing fixed-valued impulse noise from the corrupted images has been implemented. The architecture has been simulated using Virtex7 device and the reported post-place and route simulation frequency is 205 MHz. During the process of noise elimination impulse detection of the current processing pixel is performed first. Corrupted pixel is then replaced with median or nearby value of the noise free pixel depending on the number of remaining eight pixels in 3×3 window. The execution time for noise cleaning of our proposed hardware at worst case supersedes with other reported denoising hardware. Comparative study of the contemporary techniques shows better performance in terms of *PSNR*, *IEF*, *EKI*, *MSE*, *SSIM* and *FOM* of our proposed work specifically at up to 50% noise density. Performance degradation of our proposed hardware is observed as the noise density increases from 60% to 90%. The speed, hardware cost of the proposed RTINR is also better in comparison to existing architectures.

REFERENCES

- [1] W. K. Pratt, *Introduction to Digital Image Processing*. CRC Press, 2013.
- [2] A. K. Boyat and B. K. Joshi, "A Review Paper : Noise Models in Digital Image Processing," *Signal & Image Processing: An International Journal*, vol. 6, no. 2, pp. 63–75, 2015.
- [3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2008.
- [4] M. Tlich, H. Chaouche, A. Zeddami, and F. Gauthier, "Impulsive Noise Characterization at Source," *Wireless Days*, pp. 1–6, 2008.
- [5] L. Shapiro and G. Stockman, *Computer Vision*. Prentice Hall, 2001.
- [6] N. C. Gallagher and G. L. Wise, "A Theoretical Analysis of the Properties of Median Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1136–1141, 1981.
- [7] O. Yli Harja, J. Astola, and Y. Neuvo, "Analysis of the properties of median and weighted median filters using threshold logic and stack filter representation," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 395–410, 1991.
- [8] T. Chen, K.-K. Ma, and L.-H. Chen, "Tri-state median filter for image denoising," *IEEE Transactions on Image Processing*, vol. 8, no. 12, pp. 1834–1838, 1999.
- [9] T. C. Lin, "A new adaptive center weighted median filter for suppressing impulsive noise in images," *Information Sciences*, vol. 177, pp. 1073–1087, 2007.
- [10] C. T. Lu and T. C. Chou, "Denoising of salt-and-pepper noise corrupted image using modified directional-weighted-median filter," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1287–1295, 2012.
- [11] J. J. Priestley, T. Anusuya, R. Pratheepa, and V. Elamaran, "Salt and Pepper Noise Reduction with a Novel Approach of Noise Models using Median Filter," *IEEE International Conference on Computational Intelligence and Computing Research (ICCIIC)*, pp. 1–4, 2014.
- [12] H. Hosseini, F. Hesar, and F. Marvasti, "Real-Time Impulse Noise Suppression from Images Using an Efficient Weighted-Average Filtering," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1050–1054, 2015.
- [13] H. Hwang and R. A. Haddad, "Adaptive median filters: new algorithms and results," *IEEE Transactions on Image Processing*, vol. 4, no. 4, pp. 499–502, 1995.
- [14] H. Ibrahim, N. S. P. Kong, and T. F. Ng, "Simple adaptive median filter for the removal of impulse noise from highly corrupted images," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1920–1927, 2008.
- [15] Z. Wang and D. Zhang, "Progressive switching median filter for the removal of impulse noise from highly corrupted images," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 1, pp. 78–80, 1999.
- [16] H. L. Eng and K. K. Ma, "Noise adaptive soft-switching median filter," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 242–251, 2001.
- [17] S. Zhang and M. A. Karim, "A new impulse detector for switching median filters," *IEEE Signal Processing Letters*, vol. 9, no. 11, pp. 360–363, 2002.
- [18] F. Duan and Y. J. Zhang, "A highly effective impulse noise detection algorithm for switching median filters," *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 647–650, 2010.
- [19] S. Akkoul, R. Ledee, R. Leconge, and R. Harba, "A new adaptive switching median filter," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 587–590, 2010.
- [20] S. J. Horng, L. Y. Hsu, T. Li, S. Qiao, X. Gong, H. H. Chou, and M. K. Khan, "Using sorted switching median filter to remove high-density impulse noises," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 956–967, 2013.
- [21] P. E. Ng and K. K. Ma, "A Switching Median Filter With Boundary Discriminative Noise Detection for Extremely Corrupted Images," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1506–1516, 2006.
- [22] A. Tripathi, U. Ghanekar, and S. Mukhopadhyay, "Switching median filter: advanced boundary discriminative noise detection algorithm," *IET Image Processing*, vol. 5, no. 7, pp. 598–610, 2011.
- [23] I. F. Jafar, R. A. Alna'mneh, and K. A. Darabkh, "Efficient improvements on the BDND filtering algorithm for the removal of high-density impulse noise," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1223–1232, 2013.
- [24] K. S. Srinivasan and D. Ebenezer, "A new fast and efficient decision-based algorithm for removal of high-density impulse noises," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 189–192, March, 2007.
- [25] K. Aiswarya, V. Jayaraj, and D. Ebenezer, "A new and efficient algorithm for the removal of high density salt and pepper noise in images and videos," *International Conference on Computer Modeling and Simulation (ICCMS)*, vol. 4, pp. 409–413, 2010.
- [26] V. Jayaraj and D. Ebenezer, "A new switching-based median filtering scheme and algorithm for removal of high-density salt and pepper noise in images," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, pp. 1–11, 2010.
- [27] S. Esakkirajan, T. Veerakumar, A. N. Subramanyam, and C. H. Prem-Chand, "Removal of high density salt & pepper noise through a modified decision based unsymmetric trimmed median filter," *IEEE Signal Processing Letters*, vol. 18, no. 5, pp. 565–570, 2011.
- [28] M. Nooshyar and M. Momeny, "Removal of high density impulse noise using a novel decision based adaptive weighted and trimmed median filter," *Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp. 387–391, 2013.
- [29] N. K. Chaitanya and P. Sreenivasulu, "Removal of salt and pepper noise using Advanced Modified Decision based Unsymmetric Trimmed Median Filter," *International Conference on Electronics and Communication Systems (ICECS)*, pp. 1–4, 2014.
- [30] Kamarujjaman, M. Mukherjee, and M. Maitra, "A New Decision-Based Adaptive Filter for Removal of High Density Impulse Noise from Digital Images," *International Conference on Devices, Circuits and Communications (ICDCCom)*, pp. 1–4, 2014.
- [31] J. J. Priestley and V. Nandhini, "A Decision based Switching Median Filter for Restoration of Images Corrupted by High Density Impulse Noise," *International Conference on Robotics, Automation, Control and Embedded Systems (RACE)*, pp. 1–5, 2015.
- [32] W. Luo, "Efficient Removal of Impulse Noise from Digital Images," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 523–527, 2006.
- [33] M. Saedi, L. C. Anzabi, and M. Khaleghi, "Image Sequences Filtering Using a New Fuzzy Algorithm Based On Triangular Membership Func-

TABLE IV
COMPARATIVE STUDY OF DEVICE UTILIZATION SUMMARY AND MAXIMUM OPERATING FREQUENCY

Work	Mukherjee <i>et al.</i> [59]		Proposed	Deepa <i>et al.</i> [56]		Jayanthi <i>et al.</i> [55]		Proposed
FPGA Device	XC5VLX50T		XC5VLX50T	XC3S500E		XC3S500E		XC3S500E
Image Size	512 × 512		512 × 512	512 × 512		128 × 128		512 × 512
Window Size	5 × 5		3 × 3	5 × 5		3 × 3		3 × 3
Algorithm	Median Filter	Adaptive Median Filter	RTINR	Without Pipelining	With Pipelining	Without Pipelining	With Pipelining	RTINR
Number of Slices	1296	1352	186	8147	14371	2762	3705	340
Number of Flip Flops	192	192	229	8148	8710	2274	2251	229
Number of LUTs	2400	2504	400	7468	28262	3743	7242	658
Maximum Frequency	-	-	274.65 MHz	20.602 MHz	42.475 MHz	21.66 MHz	40.97 MHz	95.974 MHz

TABLE V
SYNTHESIS REPORT OF THE PROPOSED ARCHITECTURE (DEVICE
SELECTED: XC7VX330T - FFG1761 VIRTEX7)

Proposed Architecture : RTINR			
Max. Operating Frequency: 360.88 MHz			
Resource	Available	Utilized	Utilization
No. of Slice Registers	408000	229	1%
No. of Slice LUTs	204000	397	1%

tion," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 2, no. 2, pp. 75–90, 2009.

- [34] K. K. V. Toh and N. A. M. Isa, "Cluster-based adaptive fuzzy switching median filter for universal impulse noise reduction," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2560–2568, 2010.
- [35] M. S. Nair and G. Raju, "A new fuzzy-based decision algorithm for high-density impulse noise removal," *Signal, Image and Video Processing*, vol. 6, pp. 579–595, 2012.
- [36] V. Thirilogasundari, V. S. Babu, and S. A. Janet, "Fuzzy based salt and pepper noise removal using adaptive switching median filter," *Procedia Engineering*, vol. 38, pp. 2858–2865, 2012.
- [37] C. M. Own and C. S. Huang, "On the Design of Neighboring Fuzzy Median Filter," *LNCS - Intelligent Information and Database Systems*, vol. 7802, pp. 99–107, 2013.
- [38] M. Sultana, M. S. Uddin, and S. Farhana, "High Density Impulse Denoising by A Novel Adaptive Fuzzy Filter," *International Conference on Informatics, Electronics & Vision*, pp. 1–5, 2013.
- [39] F. Ahmed and S. Das, "Removal of High-Density Salt-and-Pepper Noise in Images With an Iterative Adaptive Fuzzy Filter Using Alpha-Trimmed Mean," *IEEE Transaction on Fuzzy Systems*, vol. 22, no. 5, pp. 1352–1358, 2014.
- [40] V. Crnojevic and N. I. Petrovic, "Universal Impulse Noise Filter Based on Genetic Programming," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1109–1120, 2008.
- [41] R. Dharmarajan and K. Kannan, "A hypergraph-based algorithm for image restoration from salt and pepper noise," *International Journal of Electronics and Communications (AEU)*, vol. 64, no. 12, pp. 1114–1122, 2010.
- [42] J. Wu and C. Tang, "PDE-Based Random-Valued Impulse Noise Removal Based on New Class of Controlling Functions," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2428–2438, 2011.
- [43] Z. M. Ramadan, "A New Method for Impulse Noise Elimination and Edge Preservation," *Canadian Journal of Electrical and Computer Engineering*, vol. 37, no. 1, pp. 2–10, 2014.
- [44] J. Tan and T. Bai, "Automatic detection and removal of high-density impulse noises," *IET Image Processing*, vol. 9, no. 2, pp. 162–172, 2015.
- [45] C. L. P. Chen, L. Liu, L. Chen, Y. Y. Tang, and Y. Zhou, "Weighted Couple Sparse Representation With Classified Regularization for Impulse Noise Removal," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4014–4026, 2015.
- [46] X. Wang, G. Shi, P. Zhang, J. Wu, F. Li, Y. Wang, and H. Jiang, "High quality impulse noise removal via non-uniform sampling and autoregressive modelling based super-resolution," *IET Image Processing*, vol. 10, no. 4, pp. 304–313, 2016.
- [47] P. Satti, N. Sharma, and B. Garg, "Min-max average pooling based filter for impulse noise removal," *IEEE Signal Processing Letters*, vol. 27, pp. 1475–1479, 2020.
- [48] S. N., S. P.J.S., and G. B, "An adaptive weighted min-mid-max value based filter for eliminating high density impulsive noise," *Wireless Pers Commun*, no. 119, p. 1975–1992, 2021.
- [49] P. Jun, K. Jun-Yeong, H. Jun-Ho, L. Han-Sung, J. Se-Hoon, and S. Chun-Bo, "A novel on conditional min pooling and restructured convolutional neural network," *Electronics*, vol. 10, no. 19, 2021.
- [50] C. J. Juan, "Modified 2D median filter for impulse noise suppression in a real-time system," *IEEE Transactions on Consumer Electronics*, vol. 41, no. 1, pp. 73–80, 1995.
- [51] S. C. Hsia, "Parallel VLSI Design for a Real-Time Video-Impulse Noise-Reduction Processor," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, pp. 651–658, 2003.
- [52] I. Andreadis and G. Louverdis, "Real-Time Adaptive Image Impulse Noise Suppression," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 3, pp. 798–806, 2004.
- [53] P. Y. Chen, C. Y. Lien, and H. M. Chuang, "A Low-Cost VLSI Implementation for Efficient Removal of Impulse Noise," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 3, pp. 473–481, 2010.
- [54] T. Matsubara, V. G. Moshnyaga, and K. Hashimoto, "A FPGA Implementation of Low-Complexity Noise Removal," *IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, pp. 255–258, 2010.
- [55] S. Jayanthi Sree, S. Ashwin, and S. Aravind Kumar, "Edge preserving algorithm for impulse noise removal using FPGA," *International Conference on Machine Vision and Image Processing (MVIP)*, pp. 69–72, 2012.
- [56] P. Deepa and C. Vasanthayaki, "VLSI implementation of enhanced edge preserving impulse noise removal technique," *Proceedings of the IEEE International Conference on VLSI Design*, pp. 98–102, 2013.
- [57] C. Y. Lien, C. C. Huang, P. Y. Chen, and Y. F. Lin, "An efficient denoising architecture for removal of impulse noise in images," *IEEE Transactions on Computers*, vol. 62, no. 4, pp. 631–643, 2013.
- [58] L. C. Koshy, "Real time wavelet based denoising technique for liquid level system on fpga platform," *IEEE International Conference on Green Computing, Communication and Electrical Engineering (ICGC-CEE)*, pp. 6–8, 2014.
- [59] M. Mukherjee, Kamarujjaman, and M. Maitra, "Reconfigurable Architecture of Adaptive Median Filter - An FPGA Based Approach for Impulse Noise Suppression," *International Conference on Computer, Communication, Control and Information Technology (C3IT)*, pp. 1–6, 2015.
- [60] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images*. SciTech Publishing, 2004.
- [61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [62] W. Luo, "An efficient procedure for removing random-valued impulse noise in images," *IEEE Signal Processing Letters*, vol. 15, no. 7, pp. 922–925, 2006.

Envisioning A Light-Based Quantum-Computational Nano-Cyborg

Dr Pravir Malik
 Deep Order Technologies
 San Francisco, USA

pravir.malik@deepordertechnologies.com

Abstract— By viewing light as a symmetrical, multi-layered construct, it is possible to envision a new genre of nano-cyborgs. In general cyborgs can be considered as a portmanteau of cybernetic and organic, and therefore as an entity consisting of both organic and mechatronic parts. However, in the multi-layered model of light subtle information existing in antecedent layers of light can be thought of as materializing through a process of quantization, that subsequently would require a nano-cyborg to interface with it, make sense of it, and act on it. Further quantum-level dynamics would necessitate nano-cyborgs of a tunneling type, annealing type, superposition type, and entanglement type, amongst others, that would have practical applications at nano-levels of different granularity. Such nano-cyborgs are envisioned to be built leveraging computational stratum at different levels of granularity ranging from the electromagnetic level, to the quantum particle level, to the level of atoms, to the level of molecular plans in cells. Immediate application areas of such light-based quantum-computational nano-cyborgs are envisioned to be in medical technology, material sciences, and alteration of genetic-type information, amongst others.

Keywords— *Quantum Computation, Cyborg, Symmetry of Light, Mechatronic, Nano*

I. INTRODUCTION

A cyborg is portmanteau of the words ‘cybernetic’ and ‘organic’ and is envisioned to be made up by both organic and mechatronic parts. This paper builds on the notion of ‘organic’ and ‘mechatronic’ to propose a new genre of cyborgs based on light-based quantum-computation.

Cyborgs themselves, are commonly categorized into mechanical cyborgs, intelligent cyborgs, and adaptive cyborgs, that leverage technologies such as robotics, algorithms, and machine learning respectively, that are therefore by definition limited by the current limits of AI.

As pointed out in ‘Limits of AI as Established by a Multi-Layered Symmetry-Based Model of Light’ [12] contemporary AI is based on mind-based processes such as memory, computation, sensing, and learning, that operate within a defined and limited conceptual space and are positioned to be substrate independent. This conceptual space and the operating principles that define it limit AI, which in its current incarnation will not

perceive the possibilities inherent in a multi-layered symmetry-based model of light, unless the notion of ‘substrate independence’ is anchored in the multi-layered symmetry-based model of light itself.

Note further, that even robots and androids as conceptualized to date are at best also based on similar mind-based processes, where a robot can be thought of as mechanical device or software program designed to do a specific task that would take more time by a human, and an android can be thought of as a robot designed to look like a human.

The light-based quantum-computational cyborg is going to be based on an enhanced notion of light as explored in a number of previous IEEE papers [8], based on the foundational Cosmology of Light work elaborated in a series of 10 books [3]. Such cyborgs, envisioned to operate at the scales of the electromagnetic spectrum, quantum particles, atoms, and molecular plans in cells, respectively, are further qualified to be nano-cyborgs.

Section I, A Quantum-Computation Model of Light, introduces a symmetrical, multi-layered light-based model. In this model quantization is proposed to connect the layers of light together so that as the layers cascade the vast amount of subtle information in the initiating layer progressively materializes in each subsequent layer.

Such cascading dynamism that is positioned to inform all materializations, as will be explored further, therefore also extends the notion of what it means to be ‘organic’, and Section II, Organic Surfacing in Matter and Life, suggests how layers of matter and life can be thought of as emergence due to a persistent quantum-computation. Such light-based quantum-computation pushes to the surface more of the information in light in forms of matter and life.

As such, an interface with these surfacings and an understanding of their structure will be central to the development of a light-based quantum-computational nano-cyborg being proposed here, and Section III, Light-Based Quantum-Computational Mechatronics, will suggest some technological mechanisms that must exist to access the information that is contained in light and that will be central to

the creation of this new genre of nano-cyborgs. It will also summarize the notion of alternative computational stratum at the electromagnetic, quantum particle, atom, and cellular levels that would be the foundation of the envisioned nano-cyborgs.

Section IV, Technological Extrapolations, will suggest a series of different kinds of nano-cyborgs and possible application areas.

Section V, Summary & Conclusion, will offer a summary and conclusion.

II. A QUANTUM-COMPUTATIONAL MODEL OF LIGHT

Light has a profound impact on how the nature of reality is experienced. Light traveling at speed c , 186,000 miles per second in a vacuum influences fundamentals such as space, time, and how the movement of objects may change [4]. It is possible to construct a multi-layered light-based model [10] [11] which can provide significant insight into the nature of quanta, by extrapolating on the necessity for light to move at a constant speed of c .

In this model the infinite information conceived were light to travel infinitely fast, is depicted by Equation (1), ‘Information When Light Travels Infinitely Fast’:

$$R_{C_\infty}: [Pr, Po, K, H] \quad (1)$$

This information (R_{C_∞}) is conceived as a set of four properties such that the element (Pr) represents the state of ‘presence’ formed because light is instantaneously present in whatever volume is being considered due to its infinite speed. The element (Po), ‘power’, is formed by the ability of light to overpower any other emergence not of the nature of light within the conceived volume. The element (K) represents ‘knowledge’, formed by the fabric of light being able to record any appearance or disappearance of an event in its substance. The element (H), ‘harmony’, is formed by everything that appears or disappears, being connected in the nature of such all-present light.

It is further conceived that the information summarized by (1) precipitates into material reality, represented by R_{C_U} and signifying a reality where light travels at speed c , referred to here as c_U , via intermediate realities where light is envisioned to exist at speeds slower than infinity, but faster than c_U . These intermediate realities are specified by R_{C_K} and R_{C_N} in (2) and (3) respectively, such that $c_U < c_N < c_K < c_\infty$, where c_x represents the speed of light at layer ‘x’. Note that while Einstein’s Theory of Relativity emphasizes that the acceleration to speed c from a slower speed that is not possible [16], it does not disallow speeds of light greater than c . Further, spaces with light speeds greater than c should be viewed as conceptual spaces made to vary were light to travel at different speeds. This is akin to the notion of property spaces, being separate from but influencing physical space as explored by Nobel Physicist Frank Wilczek [19].

Such precipitation itself is envisioned to take place via a series of quantization functions. The first quantization (\downarrow) is specified by ($\downarrow R_{C_K} = f(R_{C_\infty})$) and suggests that reality (R)

at c_K, R_{C_K} , is a function (f) of reality at c_∞, R_{C_∞} . Hence, the set specified by (1) is mathematically transformed into four large sets as specified by Equation (2), ‘Transformation into Four Sets’:

$$R_{C_K}: [S_{Pr}, S_{Po}, S_K, S_H] \quad (2)$$

In (2), S_{Pr} is the set of ‘presence’, S_{Po} is the set of ‘power’, S_K is the set of ‘knowledge’, and S_H is the set of ‘harmony’, respectively.

A further quantization takes place via ($\downarrow R_{C_N} = f(R_{C_K})$), and suggests that reality (R) at c_N, R_{C_N} , is a function (f) of reality at c_K, R_{C_K} .

Hence, elements from each of the four sets combine in unique combinations to create the basis of ‘seeds’, specified by Equation (3), ‘Creation of Unique Combinations’:

$$R_{C_N}: f(S_{Pr} \times S_{Po} \times S_K \times S_H) \quad (3)$$

In (3) hence, bases for a practically infinite number of unique seeds or functions are specified.

A final quantization, specified by ($\downarrow R_{C_U} = f(R_{C_N})$), suggests that reality (R) at c_U, R_{C_U} , is a function (f) of reality at c_N, R_{C_N} . This quantization results in material reality, specified by Equation (4), ‘Formation of Space, Time, Energy, and Gravity’:

$$R_{C_U}: [S, T, E, G] \quad (4)$$

In (4), ‘S’ refers to *Space*, ‘T’ refers to *Time*, ‘E’ refers to *Energy*, and ‘G’ refers to *Gravity*. Specifically, “Space” – is envisioned to be the arena for unique seeds specified by the subtle-seeds in (3); “Time” – ensuring that the possibilities in the seeds are worked out; “Energy” – associated with the conversion of seeds into concrete matter; and “Gravity” – specifying relationship between seed and seed and seeds and seeds thereby being the bases of future collectivities at different levels of granularity. Further, *Space*, being a repository of unique seeds - and therefore archetypes - represents light’s property of knowledge (K); *Time*, assuring maturity regardless of opposition represents light’s property of power (Po); *Energy*, fundamental to the process by which seeds accumulate presence, represents light’s property of presence (Pr); and *Gravity*, allowing relationship between seed and seed, represents light’s property of harmony (H) [9] [11].

Hence the multi-layered fourfold light-based model is summarized by Equation (5), ‘Multi-Layered Fourfold Light-Based Model’:

$$\left[\begin{array}{l} R_{C_\infty}: [Pr, Po, K, H] \\ (\downarrow R_{C_K} = f(R_{C_\infty})) \\ R_{C_K}: [S_{Pr}, S_{Po}, S_K, S_H] \\ (\downarrow R_{C_N} = f(R_{C_K})) \\ R_{C_N}: f(S_{Pr} \times S_{Po} \times S_K \times S_H) \\ (\downarrow R_{C_U} = f(R_{C_N})) \\ R_{C_U}: [S, T, E, G] \end{array} \right]_{\text{Light}} \quad (5)$$

In this model it is assumed that the deeper nature and activity at the quantum veil before matter is formed, is of the substance of *space-time-energy-gravity*. This is a logical outcome following the emergence of *space-time-energy-gravity* when light precipitates to speed *c*. Further, as proposed in ‘The Origins and Possibilities of Genetics’ [9], *space-time-energy-gravity* can also be thought of as the script used to write a “law” about any specific emergence. The aggregation of these “laws” determines the overall dynamics of *Space, Time, Energy, and Gravity* as experienced at the macro level. Note that the script writes more and more complex fourfold “law” into existence as evidenced by subsequent emergences of the electromagnetic spectrum, quantum particles, atoms, molecular plans, and so on [9] [13].

Note further that this model is entirely symmetrical, with the properties of light of *presence, power, knowledge, and harmony* conceived in the conceptual space were light to travel infinitely fast, showing up in different configurations of the same four properties in related conceptual spaces determined by proposed variation in speed of light. Such fourfold law is the basis of all that emerges including layers that are foundational to organic manifestation, as will be illustrated in the next section. Further, Such a view of organic manifestation necessitates a different kind of technology, that then becomes the basis of the mechatronics explored later in this paper.

III. ORGANIC SURFACINGS IN MATTER AND LIFE

The inherent fourfoldness of Light can be seen to form the foundation for the emergence of different layers of complexity of matter and life, starting at the field level and moving through the quantum, atomic, and cellular levels respectively, thereby also being the basis of organic manifestation.

Taking the first layer of emergent complexity, the Field Level, the four characteristics of light are expressed by the EM Spectrum (knowledge), its implicit energy-gradient (power), the speed *c* with which is propagated (harmony), and the potential for mass (presence) also implicit in it, respectively.

It can be surmised that the EM Spectrum itself is a repository of knowledge of a vast range of phenomena experienced in the Universe. This follows from the range of different types of waves from radio waves to gamma rays, determined as a function of frequency and wavelength of light, that form a set of archetypes. This relationship is indicated by Equation (6), ‘Knowledge Correlation in the EM Spectrum’:

$$Knowledge \propto [f(\nu) \& f(\lambda)] \tag{6}$$

It can also be seen that there is a large variation in energy or power implicit in the EM Spectrum. This is directly related to the large variance in the frequency of light in the EM Spectrum. This relationship is summarized by Equation (7), ‘Power Correlation in the EM Spectrum’, where ν is the frequency of the EM Spectrum:

$$Power \propto h\nu \tag{7}$$

The principle of harmony, as already suggested, is related to the speed of light at *U, c_U*. This is highlighted by Equation (8), ‘Harmony Correlation in the EM Spectrum’:

$$Harmony \propto c_U \tag{8}$$

Finally, Equation (9), ‘Presence Correlation in the EM Spectrum’, summarizes that mass-potential can be expressed as indicated, by combining $E = mc^2$ and the equivalence of *hν* with *E*, as in:

$$Presence \propto h\nu / c^2 \tag{9}$$

At the Quantum level it is suggested that quarks are a precipitation of the characteristic of knowledge. This is so because, protons, which determine atomic number, are composed of two “up” quarks and one “down” quark, and atomic number in turn uniquely identifies the element from the periodic table. So, an atomic number of 47, for example, specifies that the element is Silver. This implies that the unique properties of an element, the knowledge of what it is and how it will behave in the universe, is related to the quark. This is summarized by Equation (10), ‘Knowledge Correlation at the Quantum Particle Level’:

$$Knowledge \propto f(quarks) \tag{10}$$

Further, it can be surmised that electrons are a precipitation of the characteristic of energy. While quarks only exist in composite particles with other quarks, leptons are solitary particles. Leptons appear to be point-like particles without internal structure [14] with the best-known lepton being the electron. Arabatzis [1] in his article, “Representing Electrons: A Biographical Approach to Theoretical Entities”, details the characteristics of electrons. The electron may be considered as a surrogate for the lepton class. Importantly, the electron is associated with the flow of energy and power, amongst other properties. This is summarized by Equation (11), ‘Power Correlation at the Quantum Particle Level’:

$$Power|Energy \propto f(leptons) \tag{11}$$

Bosons can be thought of as force-carriers. It is due to them that all known matter particles interact. The three fundamental bosons in this category are the gluon, the photon, and the W and Z bosons. The gluon is the carrier particle of the strong nuclear force that holds quarks together. The carrier particle of the electromagnetic force is the photon. The carrier particle for the weak interactions, responsible for the decay of massive quarks and leptons into lighter quarks and leptons, are the W and Z bosons. Bosons, therefore, can be thought of as the precipitation of what creates relationship and harmony at the quantum level. This is why they can be thought of as the precipitation of harmony as summarized by Equation (12), ‘Harmony Correlation at the Quantum Particle Level’:

$$Harmony \propto f(bosons) \tag{12}$$

This leaves the Higgs-Boson: the other discovered fundamental particle. In ordinary matter the mass of an atom resides in the nucleus, made of protons and neutrons which are each made of three quarks. But it is the quarks that get their mass by interacting with the Higgs field [14]. Therefore, the Higgs-Boson can be thought of as the mass-giver, or that which gives presence to the quarks. This can be summarized by Equation (13), ‘Presence Correlation at the Quantum Particle Level’:

$$Presence \propto f(Higgs_boson) \quad (13)$$

Note that recent research at CERN [15] indicates that just as there are multiple particles in each of the other ‘families’ it is likely that there will be multiple particles in the Higgs-Boson family.

At the level of atoms, all atoms in the Periodic Table can be classified by the p-Group, d-Group, s-Group, and f-Group.

The p-Group of elements are those with a valence shell specified by the p-orbital. This indicated that the probability of the existence of an electron is equally likely on either side of the nucleus. Beyond elements such as Carbon, Nitrogen, Oxygen, and Silicon, there are elements of high significance in this group that are part of the metal, metalloid, non-metal, halogen, and noble gas sub-groupings. This grouping, therefore, can be thought to summarize many element possibilities within it. Perhaps it is the case that the possibility of ideas behind all elements has precipitated in this group. Hence one can hypothesize that this group may be a carrier or precipitation of light’s property of knowledge, forming archetypes from which all other elements are created.

Philosophically, the one probability cloud (s), to be discussed shortly, becoming two (p) signifies an essential polarity created within a unit space. Consider the hypothesis that the form is a ‘switching’ function that attracts function into form. If this is the case, then this dual manifestation may be viewed as the prerequisite condition by which a larger number of such ‘switches’ also come into being. This ‘essential two’ existing in three dimensions of space may then allow a threshold meta-function experimentation to come into being. But being the first instance of such variability in space it could therefore become an attractor for all the essential element-archetypes to precipitate.

To further reinforce this proposition, consider that the essential elements - Carbon and Silicon – both contained within this group, are what allows both thinking and virtual thinking machines to come into being. Carbon, after all, is the basis of DNA and of all life. The fact the Silicon (Si), directly below it in the periodic table and therefore sharing essential qualities, is considered the basis of all virtual thinking machines is therefore perhaps significant and may reinforce the notion that the p-Group is a precipitation of knowledge, as summarized by Equation (14), ‘Knowledge Correlation at the Atom Level’:

$$Knowledge \propto f(p_orbital) \quad (14)$$

In thinking about the d-Group, we know that it comprises the Transition Metals. These metals exhibit corrosive resistance, are generally hard and strong, and can be thought of as workhorse elements. Many industrial and well-known elements sit in this group: Copper, Zinc, Silver, Platinum Titanium, Chromium, Manganese, Iron, Cobalt, Nickel, and Gold, amongst others.

The d-orbital itself is a probability space characterized by four lobes around the nucleus. It is likely that four lobes occurring in 5 possible planes around the nucleus will create a space of stability, since there is a possibility of four lobes creating the four vertices of a tetrahedron that has been implicitly positioned as one of the most stable shapes [2]. Work

done in Crystal Field Theory [18] reinforces this concept. The general stability of the transition metals is reinforced by the d-orbital arrangement.

Note that most of the elements in this group easily lose one or more electrons to form a vast array of compounds. Further, much of the constructed world around us is created from these elements. It can therefore be surmised that these metals exist for service, to help bring about perfection in the constructed world, to help much of the machinery in which they are used, and to assist the processes dependent on them to be completed with diligence. We can conclude then that these transition metals appear to be a precipitation of Presence, as summarized by Equation (15), ‘Presence Correlation at the Atom Level’:

$$Presence \propto f(d_orbital) \quad (15)$$

Exploring the s-Group, one sees that it consists primarily of alkali earth metals and alkali metals. These groups are highly electropositive, easily losing electrons and forming positive ions, and releasing a lot of energy while doing so. Tweed [17] refers to these groups as the “violent world of the s-block”. Gray [5] points out that stars shine because they are transmuted vast amounts of hydrogen into helium. Note that both are s-block elements. This characteristic of easily released energy that the elements of this group share suggests that the s-Group is a precipitation of Power, as summarized by Equation (16), ‘Power Correlation at the Atom Level’:

$$Power \propto f(s_orbital) \quad (16)$$

Philosophically the s-orbital as a probability cloud indicates the equal likelihood that an electron can be anywhere in a symmetrical sphere around a nucleus. Since all other orbitals can be thought of as occurring within some fractal-like cloud as specified by the s-orbital, in therefore gives the sense of a space being created within which varied meta-functions can more easily precipitate to the level of U. In other words, the elements that are part of the s-Group may be thought of as the adventurers who venture forth to create some foundation by which all other element-creations can follow. The fact that H and He constitute 98% of the Universe [7] relative to other elements is significant in this view, since H and He provide the fuel with which the star-furnaces manufacture all other elements.

The f-Group comprises the Lanthanides and Actinides. Philosophically, the f-orbital, consisting of 6 probability lobes around the nucleus in 7 different planes suggests the attempt to build larger and larger bonds within a small space. This is clearly a dynamic reinforcing extended relationship and collectivity at the level of the atom. It is likely, therefore, that this group, continuing to draw the link with the quaternary architecture, is a precipitation of Harmony, as summarized by Equation (17), ‘Harmony Correlation at the Atom Level’:

$$Harmony \propto f(f_orbital) \quad (17)$$

At the cellular level, there are molecular machines that do the myriad things that distinguish living organisms. These machines are identical in all living cells [6] and uses four basic molecular plans with unique chemical personalities: nucleic acids, proteins, lipids, and polysaccharides.

Nucleic acids encode information. They store and transmit the genome, the hereditary information needed to keep the cell alive. They function as the cell's librarians and contain information on how to make proteins and when to make them. They are the keepers of a cell's knowledge, its wisdom, its ability to make laws. They are the vehicle to spread knowledge within cells and to the next generation of cells. It can be surmised then that nucleic acids are a precipitation of knowledge at the cellular level. This is summarized by Equation (18), 'Knowledge Correlation at the Cellular Level':

$$S_{K(cell)} \ni [Knowl., Wisdom, Law Making, Spread of Knowl. ...] \quad (18)$$

This relationship may also be summarized more simply by Equation (19), 'Knowledge Correlation (Simple) at the Cellular Level':

$$Knowledge \propto f(nucleic\ acids) \quad (19)$$

Proteins are known to be the cells work-horses. They are found in any part of the cell. They are built in thousands of shapes and sizes, each performing a different function. As described by Goodsell [6], some proteins are built simply to adopt a defined shape, assembling into rods, nets, hollow spheres, and tubes. Some proteins are molecular motors, using energy to rotate, or flex, or crawl. Many proteins are chemical catalysts that perform chemical reactions to transfer and transform chemical groups atom by atom, exactly as needed. With their wide potential for diversity, proteins are constructed to perform most of the tasks of the cells. In fact, there are as many as 30,000 different kinds of proteins in the human cell to execute on the diverse array of required cellular level tasks.

It may therefore be said that proteins exist for service. They bring about perfection at the level of the cell. They are characterized by extreme diligence and perseverance. Proteins, hence, can be thought of as a precipitation of presence at the cellular level as summarized by Equation (20), 'Presence Correlation at the Cellular Level':

$$S_{Pr(cell)} \ni [Service, Perfection, Diligence, Perseverance, ...] \quad (20)$$

This relationship may also be summarized more simply by Equation (21), 'Presence Correlation (Simple) at the Cellular Level':

$$Presence \propto f(proteins) \quad (21)$$

Lipids by themselves are tiny molecules. When they are grouped together though, they form the largest structures of the cell. When placed in water lipid molecules easily aggregate to form large waterproof sheets. These sheets naturally form boundaries at multiple levels, to allow concentrated interactions and work to be performed within a cell. The nucleus and the mitochondria, for example, are contained within lipid-defined compartments. Similarly, each cell itself is contained within a lipid-defined boundary.

Lipids can therefore be thought of as the promoters of relationship and of harmony in the cell. They can also be thought of as nurturing the cell-level division of labor and of allowing specialization and uniqueness to emerge. Even perhaps of exhibiting early forms of compassion and love. This

function of harmonization suggests that lipids are a precipitation of harmony at the cellular level as summarized by Equation (22), 'Harmony Correlation at the Cellular Level':

$$S_{H(cell)} \ni [Love, Compassion, Harmony, Relationship ...] \quad (22)$$

This relationship may also be summarized more simply, by Equation (23), 'Harmony Correlation (Simple) at the Cellular Level':

$$Harmony \propto f(lipids) \quad (23)$$

In contrast to lipids, polysaccharides are long branched chains of sugar molecules. Further, these are covered with hydroxyl groups, which associate to form storage containers. Because of this polysaccharides function as the storehouse of cell's energy. But also, polysaccharides are used to build some of the most durable biological structures. The stiff shell of insects is made of long polysaccharides, for example.

Polysaccharides therefor function to create energy, power, courage, and strength. They also ready the cell for adventure, amongst similar functional traits. Providing energy and strength, it can be surmised that polysaccharides are a precipitation of power at the cellular level as summarized by Equation (24), 'Power Correlation at the Cellular Level':

$$S_{Po(cell)} \ni [Power, Courage, Adventure, Justice, ...] \quad (24)$$

This relationship may also be summarized more simply by Equation (25), 'Power Correlation (Simple) at the Cellular Level':

$$Power \propto f(polysaccharides) \quad (25)$$

IV. LIGHT-BASED QUANTUM-COMPUTATIONAL MECHATRONICS

Equation (5), Multi-Layered Fourfold Light-Based Model, suggests a model at the core of the dynamic emergence of matter and life. This equation can be generalized to form Equation (26), Generalized Form of Light-Based Emergence, where x_U can be thought of as an initial output of (5) that iterates to create x_T such that the next x_U is the previous x_T :

$$x_T \leftarrow \left(\begin{array}{c} R_{C_{\infty}}: [Pr, Po, K, H] \\ (\downarrow R_{C_K} = f(R_{C_{\infty}})) \\ R_{C_K}: [S_{Pr}, S_{Po}, S_K, S_H] \\ (\downarrow R_{C_N} = f(R_{C_K})) \\ R_{C_N}: f(S_{Pr} \times S_{Po} \times S_K \times S_H) \\ (\downarrow R_{C_U} = f(R_{C_N})) \\ R_{C_U}: [S, T, E, G] \end{array} \right)_{x_U} \leftarrow_{<x_U: x_T>} \quad (26)$$

Hence the application of (26) will iteratively produce the fourfold light-based outputs as illustrated in the previous section: the electromagnetic spectrum, quantum particles, atoms, and molecular plans.

Such light-based output contrasts with the random, probabilistic output imagined to be true of the quantum-level in general. By a process of reverse extrapolation, and more importantly by a process of induction assuming Equations (5)

and (26) to be models of emergence, a different fundamental activity existing in the minutia of space must be considered a possibility.

Such innate activity will also necessitate that this fundamental core of organic-ness must be read differently, necessitating therefore, a different genre of mechatronic devices to read and respond to this activity. Such a device would need to be based on a different “sight” of the dynamics in the minutia of space, as suggested by (5) and (26).

As opposed to traditional computational devices based on binary-based logic, or mainstream quantum-computational devices as envisioned by industry leaders based on qubit logic, such a mechatronic device will need to be based on logic that recognized fourfoldness and have circuitry that can freely manipulate and build on the mathematical possibilities suggested by (5) and (26).

Mechatronic devices could leverage any of the fourfold light-based emergences as elaborated in Section III - the electromagnetic spectrum, quantum particles, atoms, and molecular plans. These emergences have been previously explored as possible computational strata [13] – with the ability to house memory, create logical gates, and seed wide functionality - and here the gist of it will be summarized.

Considering the electromagnetic spectrum level:

- **Memory:** The ability to house memory is suggested by the arising of many potential stable states manifest as different frequencies or wavelengths of light. This is further reinforced by the ability to easily change from one stable state to another.
- **Logical Gates:** The creation of logical gates can be enabled by constructive and destructive interference of wave. Conceptually, this will allow calculations and logic to be implemented.
- **Arbitrating Functionality:** From a consideration of (3) in combination with (6-9) it also becomes apparent how unique seeds or functional states can be created through leveraging the four aspects that architect the electromagnetic field - knowledge, presence, power, and harmony.

Considering the quantum particle level:

- **Memory:** Memory may be operationalized through the action of electrons in an atom. Hence, an electron in a particular orbit around an atom may represent one memory state while its existence in another orbit may represent another memory state.
- **Logical Gates:** Gates of various kinds may be constructed through combinations of orbits in atoms and the action of photons to flip the status of gates, to begin to implement broader functionality.
- **Arbitrating Functionality:** Stratum specific seeds housing unique functionality are envisioned to exist through the seed formulation introduced by Equation (3) in combination with (10 – 13).

Considering the level of atoms:

- **Memory:** Memory can be conceived by a mechanism involving some reversible reaction in which a particular atom in a molecular conglomerate has dominance, and therefore stability, subsequently yielding to another atom in the same conglomerate, and vice-versa.
- **Logical Gates:** Function may exist by implementing logical gates formed by a combination of different groups of atoms and stimuli, such that the atoms will always react in a predictable way under the influence of the same stimulus.
- **Arbitrating Functionality:** Equation (3) in combination with (14 – 17) suggests the potentially vast variety of function-based seeds that are possible, also hinting at the implicitly creative as opposed to constructive aspect of computation possible at this level.

Considering the level of molecular plans:

- **Memory:** Memory can be conceived as a function of one of the myriad types of cellular proteins that might exist in one or more stable states. The influence of an enzyme can predictably flip the protein from one state to another, thus allowing the operation of memory to come into existence.
- **Logical Gates:** Function may exist by implementing logical gates formed by the combination of different proteins and stimuli, such that a protein will always react in a predictable way under the influence of some stimulus.
- **Arbitrating Functionality:** Here too the creative as opposed to constructive aspect of computation becomes apparent when considering the formation of the potentially vast variety of function-based seeds as suggested by (3) in combination with (18 – 25).

V. TECHNOLOGICAL EXTRAPOLATIONS

The envisioned nano-cyborgs, leveraging off one or more of the mechatronic schemes suggested in the previous section, would be sensitive to information at the light-based interface that is positioned to exist in all emergences regardless of level, as suggested by Equation (26).

Hence such nano-cyborgs could be electromagnetic based, quantum particle based, atom based, or molecular plan based, in nature. Being open to the fundamental light-based dynamism that is positioned to be at the base of organic life, would potentially make these nano-cyborgs “organic” – provided they have the mechanism to make choice in real-time that does not in any way impede the natural organic-ness seeking to emerge. Fourfold logic emanating from (5) and leveraging flexible circuitry as suggested in the previous section, could assure this.

It is perhaps easiest to envision applications that may exist at the molecular-plan level in cells first, and then work backward from there.

At the cellular level hence, it can be posited that all health is based on the right functioning of a cell. In the cell not only do all four molecular plans need to play their part, but further all four molecular plans need to work with each other. The light-based quantum-computational nano-cyborgs, being essentially quantum mechanical in nature since they are operating at the level of the smallest fragments or packets of light, could be designed to generate quantum phenomenon such as superposition, entanglement, tunnelling, and annealing [20]. It would be possible to have such nano-cyborgs leverage one or all of the four quantum phenomenon to influence the health of a cell.

For example, a target set of molecules in an unhealthy cell could be urged into a state of health by superposition-causing nano-cyborgs so that the presiding seed-function, as summarized by Equation (3), could be activated in the target cells to thereby change their functioning. Or the action of a well-functioning protein could be tunneled into a set of cells by a tunnelling-causing nano-cyborg to alter growing dysfunction into function. Further, the effect of a group of healthy cells, using an annealing-causing nano-cyborg, could be made to change the functioning of unhealthy cells. Finally, an entanglement-causing nano-cyborg could combine the effect of two specific nutrients that would then be targeted at specific sets of cells or areas of the body.

At the level of atoms, specific light-based quantum computational entanglement-causing nano-cyborgs could entangle a set of meta-functions such as preside over the action of diverse elements, for example - gold, silicon, helium, and cerium - amongst other possibilities, targeting this at a light-based interface, to influence a process of crystallization that could be organized to take place on the other side interface. Such processes may also influence the field of material sciences to ground new properties or even new combinations of elements into unique materials. A similar process may also be applied to the way that quantum particles come together to create stable atoms.

At the level of the electromagnetic spectrum, light-based quantum-computational sensing nano-cyborgs may be designed to sense the rate of change in particular subtle frequencies and wavelengths to predict or even change the space-time-energy-gravity scripts, aka genetic-type information, of particular environments.

VI. SUMMARY & CONCLUSION

This paper has suggested a new genre of light-based quantum-computational nano-cyborgs. Such cyborgs are the natural outcome of modeling light as a quantum-computational, symmetrical, multi-layered construct. The quantum is seen to be a bridge mechanism between layers of light, and quantum-computation is seen as being a persistent process in which different possibilities originating in different layers of light are arbitrated into existence at every instant.

Since space-time-energy-gravity is envisioned as being a first outcome caused by light precipitating to speed c , and further, as the very script by which genetic-type information is encoded, cyborg sensitivity to particular kind of fourfoldness, perhaps at the subtle electromagnetic level, may imply that such cyborgs also be able to influence genetic-type script by sensing and shifting space-time-energy-gravity information in targeted micro-phenomenon.

In such a model organic-ness can be positioned as ubiquitous since even what is perceived as inanimate matter would have emerged through several quantum-computational iterations of the multi-layered light-model resulting in fourfold emergence at the space-time-energy-gravity level, the electromagnetic spectrum level, the quantum particle level, and the atom level, respectively. Something of the information in antecedent layers of light, is dynamic in everything which therefore offers all that emerges a sense of organic-ness. To be sensitive to such fourfold dynamism perhaps even implies that the sensing mechanism itself becomes organic.

But a particular kind of sensing mechanism and architecture that leverages natural computational stratum at different levels of granularity – the electromagnetic, quantum particle, atom, molecular plans in cells - would be required to interface with such organic-ness. Various mechatronic devices would encapsulate this ability.

The resulting nano-cyborgs would be of different kinds: Tunneling-nano-cyborgs, annealing-nano-cyborgs, entanglement-nano-cyborgs, superposition-nano-cyborgs. But then there would also be sensing-nano-cyborgs. Nano-cyborgs could also be enabled by machine learning and neural networking to learn and therefore sense differently.

Application areas of such nano-cyborgs would foreseeably be in the areas of cellular healing and medical technology, in material sciences, and in alteration of genetic-type information, amongst other possible areas.

REFERENCES

- [1] Arabatzis, T. Representing Electrons: A Biological Approach to Theoretical Entities. University of Chicago. Chicago. 2006
- [2] Fuller, B. 1982. Synergetics: Explorations in the Geometry of Thinking. MacMillan Publishing Co.: New York
- [3] Malik, P. Amazon Author Page. https://www.amazon.com/Pravir-Malik/e/B002JVAEZE%3Fref=dbs_a_mng_rwt_scns_share. 2022
- [4] Einstein, A. Relativity: The Special and General Theory. New York: Broadway Books, 1995
- [5] Gray, T. 2009. The Elements: A Visual Exploration of Every Known Atom in the Universe. Black Dog & Levental Publishers. New York.
- [6] Goodsell, David. 2010. The Machinery of Life. New York: Springer
- [7] Heiserman, D. 1991. Exploring Chemical Elements and their Compounds. McGraw-Hill. New York.
- [8] Malik, P. IEEE Author Page. <https://ieeexplore.ieee.org/author/37086022058>. 2022
- [9] Malik, P. The Origin and Possibilities of Genetics. Google Books. 2019.
- [10] Malik, P. Pretorius, L. An Algorithm for the Emergence of Life Based on a Multi-Layered Symmetry-Based Model of Light. 2019 IEEE 9th

- Annual Computing and Communication Workshop and Conference (CCWC). 10.1109/CCWC.2019.8666554, 2019.
- [11] Malik, P. "A Light-Based Quantum-Computational Model of Genetics," *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2020, pp. 1-8, doi: 10.1109/IEMTRONICS51293.2020.9216451. 2020a.
- [12] Malik, P. "Limits of AI as Established by a Multi-Layered Symmetry-Based Model of Light," *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020, pp. 0151-0158, doi: 10.1109/IEMCON51383.2020.9284955. 2020b.
- [13] Malik, P. The Emergence of Quaternary-Based Computational-Strata from a Symmetrical. IJSSST. <https://ijsst.info/Vol-22/No-1/paper4.pdf>, 2021
- [14] Olive, K.A et al. 2014. Particle Data Group. *Chin. Phys. C*, 38, 090001
- [15] Overbye, D. 2015. Physicists in Europe Find Tantalizing Hints of a Mysterious New Particle. *New York Times*. Dec 15, 2015
- [16] Perkowitz, S. *Slow Light*. London: Imperial College Press, 2011
- [17] Tweed, M. 2003. *Essential Elements: Atoms, Quarks, and the Periodic Table*. Walker & Copmany. New York.
- [18] UCDAVIS-CFT. 2015. *Description of Orbitals*.
- [19] Wilczek, F. *A Beautiful Question: Finding Nature's Deep Design*. New York: Penguin Books, 2016.
- [20] QIQuantum. <https://www.deepordertechnologies.com/qiquantum>. 2022

Smart Irrigation Systems: Soil Monitoring and Disease Detection for Precision Agriculture

Prem Sai Peddi

Department Electrical and Electronics
Engineering
Birla Institute of Technology and
Science Pilani, Dubai Campus
Dubai, UAE
premsaip@rocketmail.com

Anuragh Dasgupta

Department Electrical and Electronics
Engineering
Birla Institute of Technology and
Science Pilani, Dubai Campus
Dubai, UAE
anuraghdasguptaofficial23@gmail.com

Vilas H. Gaidhane

Department Electrical and Electronics
Engineering and APPCAIR
Birla Institute of Technology and
Science Pilani, Dubai Campus
Dubai, UAE
vilasgd612@gmail.com

Abstract— Smart farming is an evolving concept in the field of information and communications technology. In this, the IoT sensors and image processing is used to establish transparent mechanisms of feedback about the growth and productivity of crops and the environmental surrounding conditions. In this paper, the solution of the aforementioned problem statement in the form of an accountable live information system of the cultivated crops to yield efficiency has been presented. The feedback mechanism consists of monitoring parameters like temperature, humidity, weather, soil and crop moisture, crop health, etc. It provides the information between the planting phase and the harvesting phase to facilitate soil management and climate forecasting in real time. The proposed paper suggests the use of an open data platform, namely Adafruit IO, for visualizing and analyzing real-time in the IoT integrated system. Further, image processing approach has been used for crop remotely health monitoring for 2 widespread diseases namely, *Glomeralla Cingulata* and *Phaeoisariopsis Bataticola*. Owing to the economical nature and the ergonomic design of the proposed system, it has the feasibility of being implemented on a large scale in water scarce economies aiming to build a sustainable smart farming infrastructure by automating existing irrigation systems.

I. INTRODUCTION

The internet of things (IoT) comprises of all the physical electronic devices that are connected to the internet, all of which are actively collecting and sharing information. These devices are likely to be found embedded with sensors, software, and processors that enable them to connect with other devices and systems over numerous communication networks. IoT has been considered a misnomer, given that no device needs to be connected to the public accessed internet.

Today's smart computing is mostly based on the Internet of Things and over the 46% of the world population is using this technology. It plays a vital role in transforming conventional forms of technology into next generation technology. IoT has already gained a critical role in areas of research globally and specifically in the area of advanced wireless communication technology. It has seen exponential growth in usage in a very short period of time. In the view of

a normal user, IoT has laid the foundation for products that uses wireless technology extensively. For example, it is being used in products like smart living, smart education, automation, and process controls [1-3]. Commercially, it is being used in manufacturing, transportation, agriculture, and business management as well [4-7].

The future of agricultural technology is precision agriculture. The data that is being generated from multiple sensors on the field can be used for data analytics. Therefore, assisting farmers in improving crop yield. The aim of this paper is to design a working product which will enable farmers to access real time soil, crop, and environmental data.

The anticipated advantages of smart farming include remote monitoring for farmers, handling water supply and natural resource conservation. Real time data allows for necessary manipulations of variables that can be handled by man. Integrating an image processing mechanism guarantees information about crop diseases as well. This would allow the farmer to take quick action and stop the disease from spreading to other crops in the field.

Some disadvantages of smart farming are the requirements. Full-time availability of the internet is a major challenge in the rural areas. Unfortunately, most of the farming in India happens in rural areas. However, there has been a lot of improvement in last few years. Multiple segments of Indian states have been getting access to the public internet. The smart farming-based equipment require farmers to understand and learn the use of technology. A mobile application with a sophisticated user interface however might enable farmers to understand how this concept works.

Integration of image processing and computer vision enhances the potential of this system. With computer vision, the agricultural industry greatly benefits by further productivity along with lower capital costs surrounding production capacities [8]. This is done via the detection and analysis of objects and presenting valid hypotheses based on meaningful interpretations out of a sequence of images. Computer vision AI models have immeasurable uses in the

fields of planting, harvesting, analysis of weather, weeding and crop health detection and real time feedback for monitoring [9].

This paper presented an agricultural system with an IOT environment which requires adequate manpower. It employs IOT and cloud computing globally to remove the inadequacy and lack of management, which are considered to be the key factors responsible for the decline in quality agriculture.

II. METHODOLOGY

Varying crops require different levels of water levels for cultivation between the plantation and the harvesting phase depending upon the week of harvest. Based on the standardized template of the sample crops, information about the minimum threshold and maximum capacity of water required along with live feed of additional parameters like pesticides, seed monitoring, sunshine and humidity on the proposed automated irrigation network would be fed into the system. The water supply from the submersible water pump would be released at regular intervals based on the inputs from the sensors in the soil. Moreover, the quantity and the type of pesticides can be decided using the concepts of image processing approaches. The image processing techniques such as morphology, binarization and segmentation can be used for the identification of anomalies over a crop area to estimate and monitor plant growth.

III. PROPOSED SYSTEM DESIGN

The proposed system for smart irrigation is shown in Fig. 1. It consists of Sensors, Node MCU, LDR and camera. The principal framework comprises of a Wi-Fi Module, specifically, an ESP8266 Node MCU configured with numerous sensors such as the DHT11 humidity sensor, DS18B20 temperature sensor probe, soil moisture sensor, a light dependant resistor (LDR) and a water pump as shown in Fig. 1. The descriptions of each sensor implemented in this design are given below.

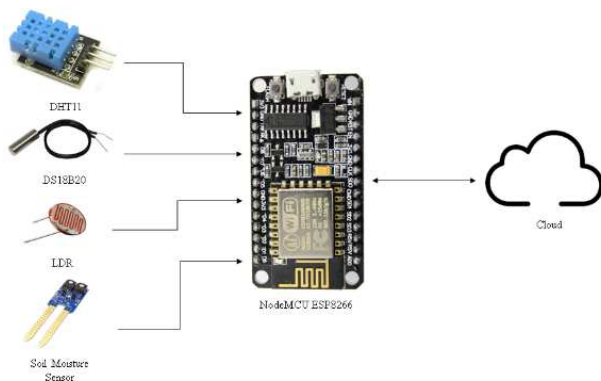


Fig. 1. Hardware Diagram

A. NodeMCU ESP8266

The ESP8266 is a low-cost Wi-Fi microchip, with built-in TCP/IP networking software, and microcontroller capability. This Wi-Fi module is used in the design to connect all the sensors to an online IO, Adafruit to share, collect and analyse the data.



Fig. 2. NodeMCU Module

B. Soil Moisture Sensor

Soil moisture sensors are globally used to estimate the content of water in the soil. The moisture sensor used in this system is a capacitive sensor. It calculates the change in capacitance caused due to the dielectric. It cannot measure moisture directly as pure water does not conduct electricity. Some advantages of using a capacitive sensor are that corrosion is avoided and gives an accurate reading of the moisture content in the soil.

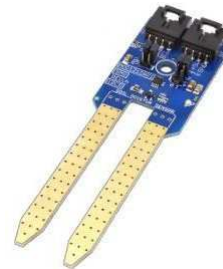


Fig. 3. Capacitive Moisture Sensor

C. DHT11 Humidity Sensor

The DHT11 Sensor is the most frequently used temperature and humidity sensors in the field of IoT, owing to its extremely low price. It uses a capacitive humidity sensor and a thermistor to measure the air temperature. It also directly produces a digital signal without the need of an ADC.

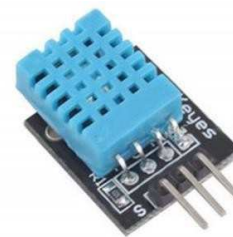


Fig. 4. DHT11 Sensor

D. Light Dependent Resistor(LDR)

LDRs are also called photoresistors since the resistance produced is dependent on the amount of light. Hence, this module is used in our system to monitor the amount of sunlight during the day. The resistance of this LDR is indirectly proportional to the intensity of light, hence, when the light intensity increases, the resistance offered by the LDR decreases.

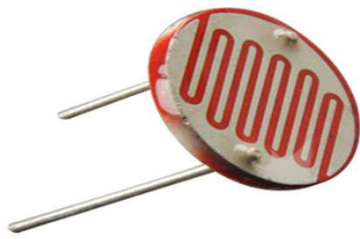


Fig. 5. Light Dependent Resistor

E. Water Pump

The mini water pump that is used in this model is a 3-5V DC Pump. It is programmed to turn on when the moisture content in the soil is lower than the configured value.



Fig. 6. Mini Water Pump

This model also employs an open-source Input/Output (I/O) cloud service, namely, Adafruit I/O. It is a platform that permits aggregation, visualisation, and analyzation of live information on the cloud. Adafruit I/O also enables motor controls and reading data via sensors. The cloud service has multiple feeds that are used to monitor various data being captured by the sensors. A minima and maxima are predefined during configuration, below or over which the farmer is notified for corrective measure.

The ESP8266 NodeMCU governs the communication between the sensors on the board and behaves as the IOT Gateway to the cloud. All the sensors detect the physical parameters and convert the analogue value into a digital value. This is achieved using an in-built Analog-Digital-Converter (ADC) in the sensor modules. The humidity sensor, DHT11, is used to compute the environmental humidity. The temperature sensor probe and the OpenWeatherAPI is used to monitor the soil temperature and get live environmental temperature, respectively. The soil moisture sensor is a capacitive sensor which estimates the amount of water in the soil. It works on the principle of open and short circuit. In simpler words, it acts as a switch with an ON/OFF mechanism [11]. Whether the output is high or low indicated by the in-built LED. When the soil is dry, the current is not conducted, hence, acting as an open circuit, with the output being high. When the soil is wet, the current flows from one terminal to the other and the circuit is shorted. Consequently, the output is low. Therefore, when the moisture sensed is below

threshold, the water pump turns on and provides continuous water flow till the threshold value is met [10].

The cloud service, Adafruit I/O, deployed in this system will provide a dashboard of multiple feeds depending on the parameters that have to be analysed. In this case, a total of 8-9 feeds are set up which include all the data acquired from the installed sensors. This will also comprise of a system which will alert the farmer or the user when the environmental factors are extreme. For example, whenever the temperature measured is above the set point, an output from a decision logic notifies the farmer. A model Adafruit I/O dashboard from an run is shown in Fig. 7.

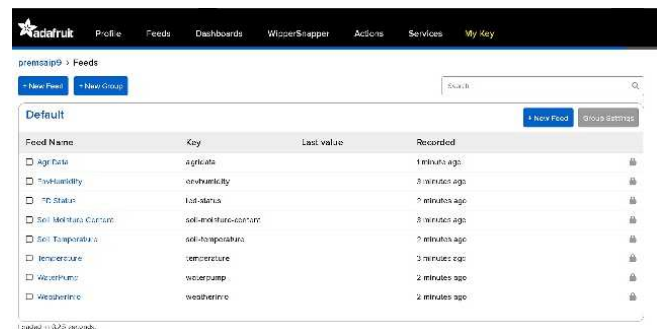


Fig. 7. Adafruit Dashboard

The IoT based system is implemented by using Arduino IDE. The sensing phase is concerned with the estimates of all the physical parameters which comprise of humidity, moisture, temperature, and light. Although the ESP8266 module acts as the IoT gateway, Arduino is used to program the sensors to it. The basic flow of the program is as follows:

Start

- Initialise all sensors
- Continuous sensing of data over regular time intervals
- Analogue to Digital conversion of data
- Set threshold values for physical parameters
 - If the value of acquired data is above threshold
 - LED glows on particular sensors
- Read weather forecast from OpenWeatherAPI
- Establish data on Adafruit dashboard

End

F. Image Processing

The following parameters can be broadly classified as the major criteria involved surrounding the productivity of crops and plants in 4 respective areas:

- Identifying Plant/Crop diseases
- Monitoring the growth of crop/plant
- Monitoring the health aspects of crop/plant throughout its plantation timeline

The following flowchart in Fig.8. lays down the foundations of the steps involved ranging from acquiring our image to the final classification of the disease detected:

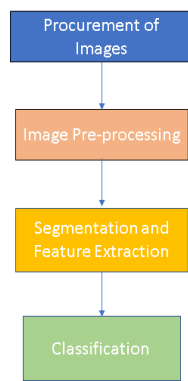


Fig. 8. Basic Steps of Image Processing in Leaf Disease Detection

The Images of the target plants during their harvesting phase are captured via a webcam. The images are pre-processed in order to eliminate any distortions or impurities and noise which might be present in the images extracted and are prepared for the upcoming processing methods like extraction of features required in the later stages. Segmentation is primarily done to separate the respective area of interest by filtering out from the image captured. The basic purpose of segmentation is to create a collection of segments that are combined overall to represent the entire image into a set of contours which are obtained from the captured image.

Feature Extraction is mainly used to extract features from the processed image after which we use the respective features for classification purposes. Its main use is to reduce dimensions in the image and compress the data which is to be processed in order to target the specific features which help us in disease classification. This is done in order to filter out the input data and eliminate redundant data.

Classification is done on the basis of spectral features in the features created. And classification aids in the process of dividing the feature target space into various classes according to the input decision rule.

Initially after high resolution images of plants/crops are captured via a webcam, image pre-processing and processing techniques are implemented in order to get features which would be needed for analysis at a later time. The image, which is in the form of Red, Green and Blue (RGB) is converted to a Hue, Intensity, Saturation (HIS) model for increasing luminance of every frame of the image [11]. Further for the purpose of smoothening and filtering out noise. This is done by enhancing contrast in the image for increasing the accuracy of output and better implementation of segmentation on the image. Furthermore, Image segmentation is carried out using thresholding which is an efficient method in order to separate the background from the foreground and also masking pixels which are green indicative of healthy parts of a leaf or crop [12].

Segmentation is done using k means clustering, with standard Euclidean distance as the measuring parameter for

calculating extent of similarity, which is an unsupervised machine learning algorithm where k denotes the number of centroids (which represents the center of the cluster) that are present in our dataset. For our respective algorithm, a structure was created for color transformation which is based on the model proposed by Oo and Htun[13] in his research. This is done mainly to mask the green pixels of the leaf image which is to separate out parts of the image. Our proposed algorithm creates $k = 3$ clusters in no particular order- One cluster for separating out the leaf from the background, one cluster to segment out the healthy part of the leaf and the final cluster dedicated to segment out the infected and diseased part of the leaf, if there is any. Since k is small, the computational speed of the algorithm increases exponentially than hierarchal clustering. The Infected cluster is converted to HSI from the RGB format. In this work, the SVM (Support Vector Machine) classification method is used because it is much easier due to less processing time [14-17].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Irrigation System

The various reading has been taken from the design system to facilitate and validate the proposed system. It has been observed that the water pump starts automatically at the particular soil conditions. A prototype for monitoring the soil condition is shown in Fig. 9. The setup acquires the values of moisture, humidity and temperature and transmits it to the cloud via the NodeMCU Module.

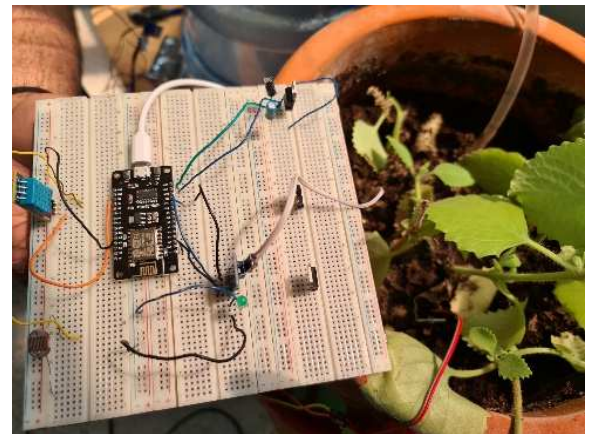


Fig. 9 Physical Setup

Table 1 shows the status of the pump corresponding to the threshold value that has been set, along with other measured data. A tulsi plant was used for the purpose of testing and threshold range coded for moisture content used was between 60 to 63, i.e below 60% moisture, the motor is on and above 63%, the motor is turned off, between 60 and 63, there is no change in the motor status. However, the threshold values for light and moisture can be changed in the program according to the needs during field implementations. Fig. 10 to Fig. 13, the variation of soil Moisture, Soil Temperature, Environment Humidity, Environment Temperature, respectively.

TABLE 1. SENSOR MEASURED DATA DISTRIBUTED

S. No	Time	Soil Moisture (%)	Soil Temperature (Celsius)	Environment Humidity (%)	Environment Temperature (Celsius)	Pump Status
1.	11.00 AM	65.69	22	47	25	OFF
2.	12.00.10 PM	62.76	22	48	25	ON
3.	1.00.10 PM	60.8	22	47	25	ON
4.	2.00.11 PM	60.61	22	47	25	ON
5.	3.00.08 PM	63.21	22	47	25	OFF

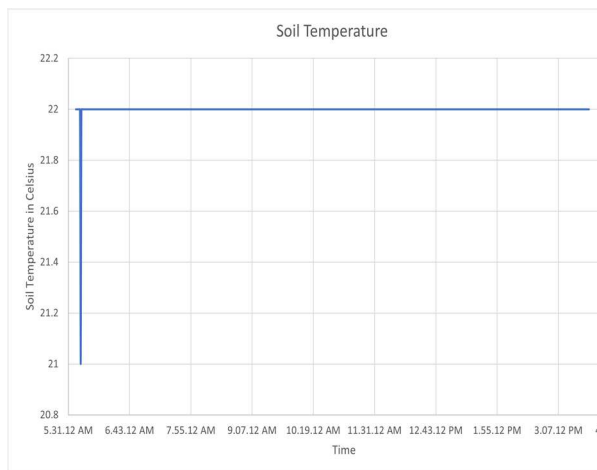


Fig 10. Soil Temperature sensed over a period of 8 hours

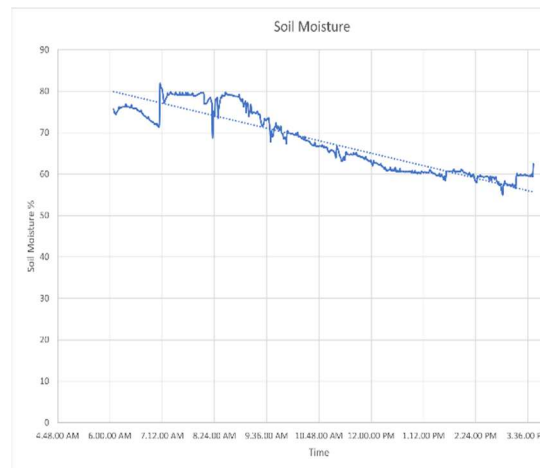


Fig 12. Soil Moisture data of over 8 hours

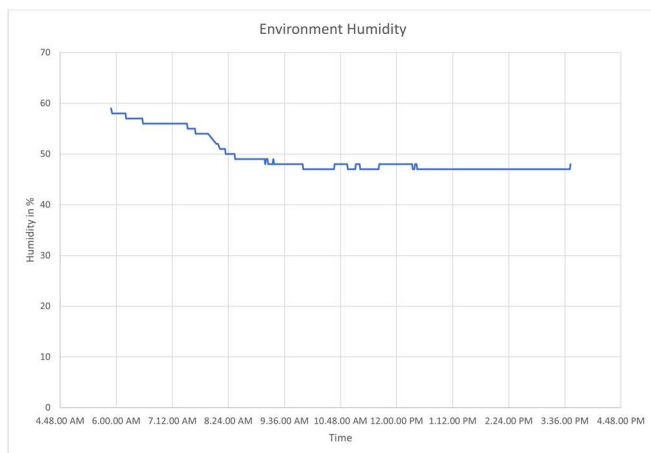
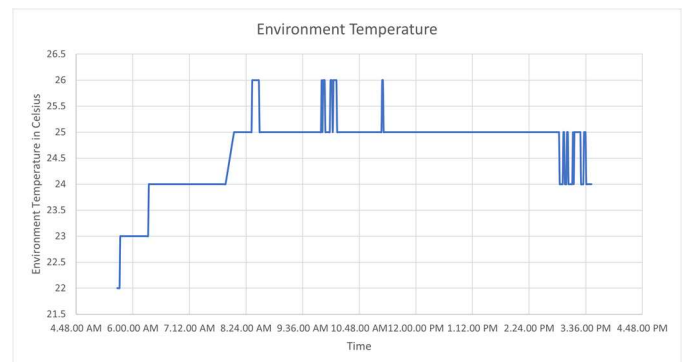


Fig 11. Environmental Humidity sensed over a period of 8 hours



The graphs attached above were obtained for a testing period of 8 hours. All the data was also published on the online service Adafruit. This has been shown in the Fig. 14 below.



Fig 14. Published data on Adafruit IO

B. Disease Detection using Image Processing

The presented model has been validated using the available dataset [18]. The images are initially pre-processed using the filter to remove the noise. After pre-processing edge extraction is carried out using canny edge detection approach to preserve main features and remove the remaining features as shown in Fig. 15. It is observed that Canny edge detection method performs better as compared Sobel approach. The feature extraction on the region of interest gives information whether the plant or crop is healthy or unhealthy. This work has been carried out for two widespread diseases – *Glomeralla Cingulate* and *Phaeoisariopsis Bataticola*. The various feature metrics then calculated and used as an input parameter for further classification. The various average values of parameters are shown in Table I.

TABLE 2. EXTRACTED FEATURES PARAMETERS

Features Parameters	Metric Values
Mean	41.1795
Standard Deviation	69.2234
Entropy	2.9503
RMS	8.50045
Variance	4583.02
Smoothness	1
Kurtosis	3.11504
Skewness	1.30558
IDM	255
Contrast	1.4293
Correlation	0.834972
Energy	0.467692
Homogeneity	0.91613

The leaf disease detection GUI is also shown in the figure below. The images captured have vertical resolution and horizontal resolution of 96 dpi with bit depth equal to 24. The dimensions of the captured image is resized in MATLAB for processing. The webcam to capture query images to test our model has the following specifications: Lenovo 300 FHD Flexible Mount Webcam FHD 1080P 2.1 Megapixel CMOS Camera.

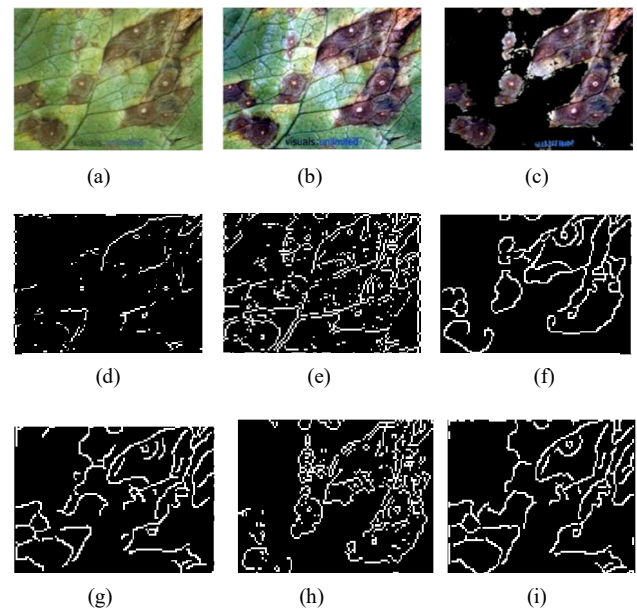


Fig. 15. (a) leaf with anthracnose, (b) contrasted image, (c) segmented image, (d, e, f) edge detected using Canny method for feature extraction, (g, h, i) edge detected using Sobel method for feature extraction

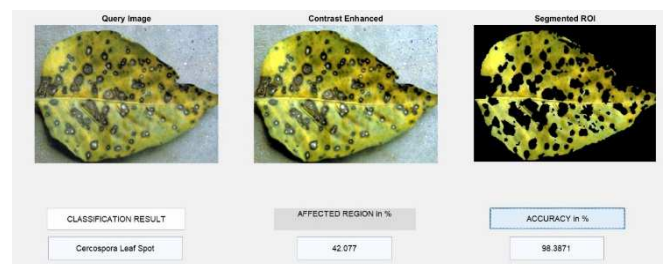


Fig. 16. Sample output of classification from a query image specifying affected region and accuracy of output which does 500 iterations

V. CONCLUSION

This paper proposes and implements the smart farming-soil monitoring and disease detection system. The presented may be useful for automatic irrigation system which increases the efficiency of the harvesting process by tracking real time plant/crop growth. It enables the farmers at all times requiring his intervention in the process only in the case of any anomalies. The plant disease classifier classifies two diseases: *Glomeralla Cingulate* and *Phaeoisariopsis Bataticola*. The design can further be improved by increasing and training the dataset on new diseases and real time images. The proposed system reduces human labor and workload of the farmers. The process is economical owing to its low budget and highly feasible considering the processing of low-resolution images captured via webcam which implies its suitability for farmers’ uses.

REFERENCES

[1] A. L Kor, C. Pattinson, M. Yanovsky, and V. Kharchenko, “IoT-enabled smart living,” *Technology for Smart Futures*, pp. 3-28, 2018, Springer, Cham.

[2] J. Shenoy and Y. Pingle “IOT in agriculture,” In: 3rd international conference on computing for sustainable global development, 2016, pp. 1456-1458.

- [3] R. Dobrescu, D. Merezeanu and S. Mocanu, "Context-aware control and monitoring system with IoT and cloud support," *Computers and Electronics in Agriculture*, vol. 160, pp.91-99, 2019.
- [4] J. Wan, B. Chen, M. Imran, F. Tao, D. Li, C. Liu and S. Ahmad, "Toward dynamic resources management for IoT-based manufacturing," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 52-59, 2018.
- [5] F. Zantalis, G. Koulouras, S. Karabetsos and D. Kandris, "A review of machine learning and IoT in smart transportation," *Future Internet*, vol. 11, no. 4, pp. 94, 2019.
- [6] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N. Lekbangpong, A. Wanichsombat and P. Nillaor, "IoT and agriculture data analysis for smart farm," *Computers And Electronics in Agriculture*, vol. 156, pp. 467-474. 2019.
- [7] M. Del Giudice, "Discovering the Internet of Things (IoT) within the business process management: A literature review on technological revitalization," *Business Process Management Journal*, vol. 22 no. 2, pp. 263-270, 2016.
- [8] A. Na, W. Isaac, "Developing a human-centric agricultural model in the IOT environment," in 2016 International Conference on Internet of Things and Applications (IOTA), 2016, pp. 292-297.
- [9] K. G. Liakos, P. Busato, D. Moshou, S. Pearson and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p.2674, 2018.
- [10] Devika, C. M., K. Bose and S. Vijayalekshmy, "Automatic plant irrigation system using Arduino," in IEEE International Conference on Circuits and Systems (ICCS), 2017, pp. 384-387.
- [11] K. Namee , J. Polpinij and G. M. Albadrani, "Applying image processing and edge computing for plant growth monitoring in smart farm," In 2020 2nd International Conference on Image Processing and Machine Vision, 2020, pp. 47-52.
- [12] N. Kaushik, K. G. Nikhil and S. Sarkar, "Plant leaf disease detection and soil condition monitoring system using CNN and IoT" *Journal Of Xi'an University of Architecture and Technology*, vol. 12, no. 6, pp.1006-7930, 2020.
- [13] Y. M. Oo, and N.C. Htun, "Plant leaf disease detection and classification using image processing," *International Journal of Research and Engineering*, vol. 5, no. 9, pp. 516-523, 2018.
- [14] V. H. Gaidhane, Y. V. Hote, and Vijander Singh, "Emotion recognition using eigenvalues and Levenberg–Marquardt algorithm-based classifier," *Sādhanā*, vol 41, no. 4, pp. 415-423, 2016.
- [15] N. Kumar, V. H. Gaidhane, and R. K. Mittal, "Cloud-based electricity consumption analysis using neural network," *Int. J. Comput. Appl. Technol.*, vol. 62, no. 1, pp. 45-56, 2020.
- [16] V. H. Gaidhane, N. Kumar, R. K. Mittal and J. Rajevenceltha, "An efficient approach for cement strength prediction," *Int. Journal of Comput. Appl.*, pp. 1-11, 2019.
- [17] V. H. Gaidhane, and Y. V. Hote, "An efficient edge extraction approach for flame image analysis" *Pattern Anal. Appl.*, vol. 21, no. 4 pp. 1139-1150, 2018.
- [18] <https://github.com/anuragcoder23/Plant-Leaf-Disease->

Bimetal (Au-Pd, Au-Pt) loaded WO₃ hybridized graphene oxide FET sensors for selective detection of acetone

Radha Bhardwaj
Dept. of Electrical & Electronics
Engineering,
Birla Institute of Technology and
Science (BITS)-Pilani,
radikabhardwaj.rb@gmail.com

Arnab Hazra
Dept. of Electrical & Electronics
Engineering,
Birla Institute of Technology and
Science (BITS)-Pilani,
arnabhazra2013@gmail.com ,
arnab.hazra@pilani.bits-pilani.ac.in,
Tel: +91-1596-255724

Abstract— Efficient detection of acetone is important for a variety of applications in pharmaceutical, automotive industries, medical diagnosis etc. Surface modification is one of the potential method to enhance the sensitivity as well as selectivity of any sensors. In recent days, surface functionalization with bimetallic nanoparticles become attractive because of its enhanced catalytic properties and the possibility to form discrete heterojunctions. In this study, WO₃ flowered morphology was prepared by one step acid precipitation method and bimetallic nanoparticles of Au-Pd and Au-Pt were deposited on WO₃/GO hybrid layer by one-step dip-coating process and fabricated a back gated field effect transistor (FET) structured sensor. Various morphological and structural characterizations were performed to study the various properties of the hybrid sensing layer. I_D-V_{GS} characteristics and the acetone sensing performance were measured for both the sensors i.e., Au-Pd/WO₃/GO and Au-Pt/WO₃/GO at room temperature. Among the two sensors, Au-Pt/WO₃/GO FET sensor exhibited an appreciably high sensitivity of 56% towards 80 ppm acetone at room temperature under applied gate voltage (V_{GS}) of 1.2V. The lower detection limit of the Au-Pt/WO₃/GO FET sensor was 400 ppb of acetone where it showed a 3 % response. The sensing mechanism envisages that the bimetallic loading in the ternary form of the nanocomposite enhanced sensitivity significantly by the spill-over effect. Also, the application of an optimized gate voltage amplified the sensitivity of the FET structured sensors.

Keywords— noble metals/WO₃/GO, FET, acetone sensing, improved sensing properties

I. INTRODUCTION

A volatile organic compound like acetone is used in a variety of applications like purifying paraffin, dissolved plastic, laboratories and most attractively in noninvasive disease detection [1,2]. But, continuous inhalation of acetone causes illness and affects the nervous system so a highly sensitive acetone sensor is required to detect the traces of acetone [3]. Semiconducting Metal oxide nanostructure-based composites are most widely used in VOC sensing applications

due to their large specific surface area, high carrier mobility and good sensing properties [4,5]. Among them, WO₃ is n-type semiconducting material considered a promising candidate for VOC sensing because of its unique chemical and physical properties [5,6]. Pure WO₃ gas sensing properties were optimized by different research groups [7-10]. Chong wang et. Al. reported pure WO₃ flowered nanostructures for NO₂ sensing and showed a good response (152) at 80 ppm concentration at a slightly high operating temperature (90 °C) [9]. Marco Righettoni et al. reported a pure WO₃ sensor for breath acetone detection. The acetone sensor response was (4.63 at 600 ppb) at a high operating temperature (400 °C) in dry air [24]. Still, metal oxide sensors show some shortcomings such as high operating temperature, low response speed and low selectivity [2,11]. On the other hand, at low operating temperature metal oxide sensor shows poor sensitivity and low response and recovery time [12]. High operating temperature sensors require high power consumption and can cause device damage. In this scenario, 2D graphene Oxide (GO) emerges as an ideal channel material for gas sensing application due to its outstanding electrical and physical properties including large specific area, strong gas adsorption capacity, high electrical mobility and conductivity at room temperature [1,6,13-17]. GO composites with semiconducting metal oxides emerges with good sensing properties [2,4]. GO and WO₃ based nanocomposites attracted wide attention in VOC sensing applications [5,6,15]. Jinniu Zhang et al introduced GO-WO₃ composite nanofibers and showed the highest response of 35.9 to 100 ppm acetone at 375 °C, which was 4.3 times higher than pure WO₃ nanofibers [32]. However, the sensitivity and other sensing properties of developed nanostructures can be improved by the functionalization with noble metal nanoparticles [14, 15, 18-21]. Lu Chen and coworkers found that noble Pt nanoparticles decorated sensor exhibits a good selectivity and response (12.2) to 10 ppm acetone with a very fast gas response/recovery time (14.1/16.8 s) at the operating temperature of 200 °C [3].

In this work, we are reporting bimetallic (Au-Pd, Au-Pt) loaded WO₃ hybridized GO FET sensor systems. We synthesized flowered WO₃ nanostructures with one step chemical method and FET sensor systems were synthesized by spray coating. FESEM (field emission scanning electron microscopy), Raman spectroscopy and UV-Vis spectroscopy results were used to characterize the samples. Transfer

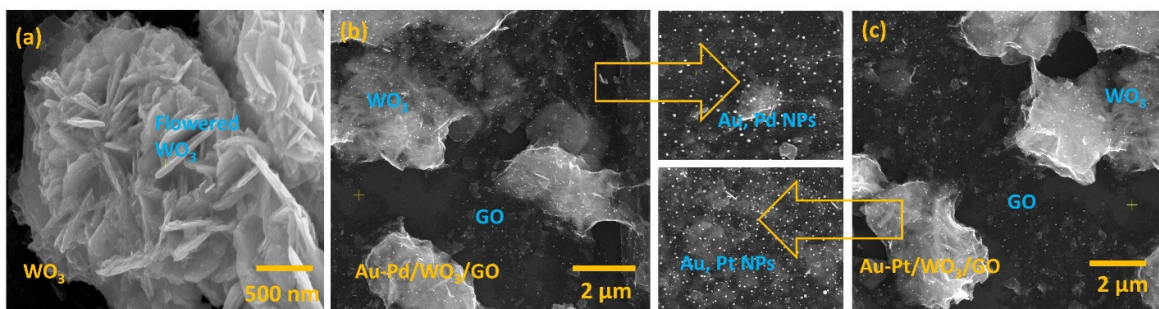


Fig. 1 Microscopic images of pure WO₃ (a), Au-Pd/WO₃/GO (b) and Au-Pt/WO₃/GO (c) and magnified image of nanoparticles distribution on samples.

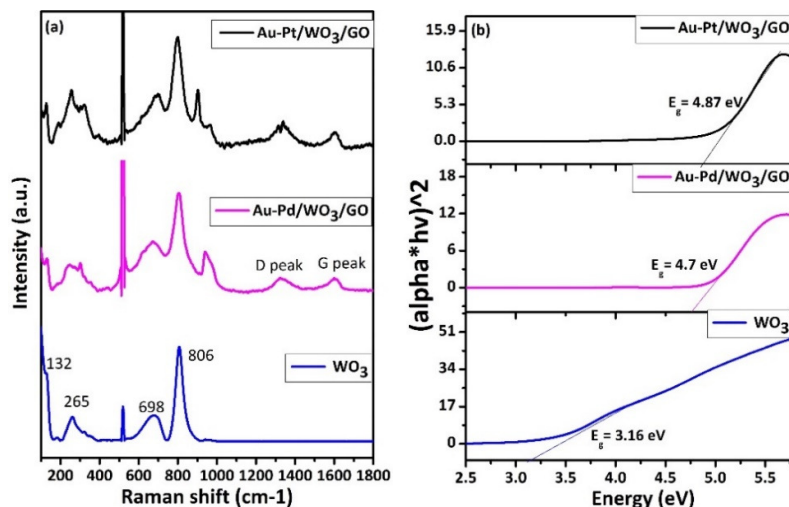


Fig. 2 (a) Raman spectra and (b) UV-Vis plot of pure WO₃, Au-Pd/WO₃/GO and Au-Pt/WO₃/GO samples.

characteristics of prepared FET sensors were optimized. The sensing properties of pure WO₃ and bimetallic (Au-Pd, Au-Pt) loaded WO₃/GO FET sensors were studied by gas measurement setup.

II. EXPERIMENTAL

A. Synthesis of WO₃ nanostructures

One step acid precipitation route was used to prepare WO₃ nanoflowers. The synthesis process includes 16.5 g Na₂WO₄·2H₂O dissolved in 125 ml HNO₃ and 125 ml deionized water to form a homogenous mixture. The resulting suspension was aged for 24 h followed by 4 h stirring at 25 °C. The resulting yellow precipitate was undergone the gravity filtration method to remove impurities and was cleaned multiple times with water and ethanol and then dried at 60 °C. Annealing of material was carried out at 500 °C for 3.5 h to remove the acidic impurities. Green-colored homogenous WO₃ solution was obtained by adding 600 mg WO₃ powder to 100 ml of DI water. The detailed procedure is given in our previous reports [15].

B. Synthesis of nanocomposites

All the chemicals are analytical grade with high purity. 0.2 wt% GO suspension was obtained by adding graphene oxide powder in DI water at room temperature with continuous stirring to maintain homogeneity. Similarly, 1MM noble metal nanoparticles (Au, Pd and Pt) solutions were prepared

by incorporating metallic salts (AuCl₃, PdCl₂ and H₄PtCl₆·xH₂O) in DI water at a stirring state and with this drop by drop hydrochloric acid was also added to the solution to reduce the aggregation of nanoparticles at room temperature. Boron-doped, ~500 μm thick <100> SiO₂/Si was used as a substrate for the sensor preparation. Nanocomposite samples were prepared by mixing nanoparticle solutions (Au-Pd and Au-Pt) separately with continuous sonication and then spray coated on previously deposited WO₃ and GO layer on SiO₂/Si substrate and dried at room temperature as shown in table 1. Thermal annealing was performed at 250 °C for 4 h in the air to maintain the thermal stability and crystallinity in samples.

Table 1. Specifications of all three samples.

Sample no.	Specification
S1	Pure WO ₃
S2	(1MM Au + 1 MM Pd) + WO ₃ + GO
S3	(1MM Au + 1 MM Pt) + WO ₃ + GO

C. Device fabrication

Au-Pd and Au-Pt/WO₃/GO FET sensors were fabricated by depositing top 150 nm thicker gold electrodes by using electron beam evaporation unit and Cu physical mask. HF etching was used to create back gate contact in FET devices

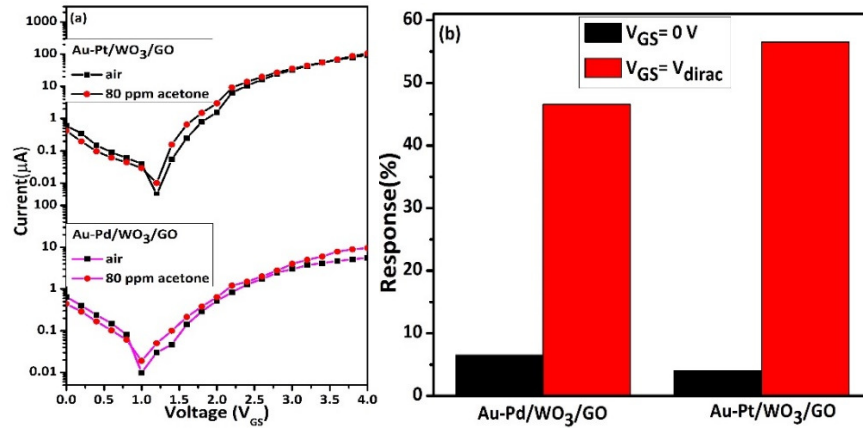


Fig. 3 I_D - V_{GS} characteristics of both FET sensors; Au-Pd/WO₃/GO and Au-Pt/WO₃/GO in air and 80 ppm acetone (a) and Response of all FET sensors in 80 ppm acetone measured at $V_{GS} = 0$ V and $V_{GS} \approx V_{dirac}$ (b) under $V_{DS} = 1$ V at room temperature.

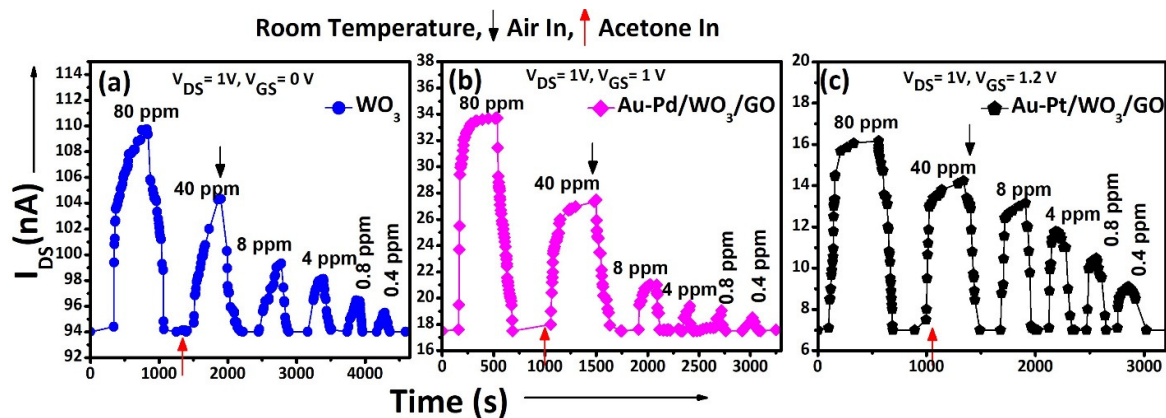


Fig. 4 The transient curve of all sensors i.e., pure WO₃ (a), Au-Pd/WO₃/GO (b) and Au-Pt/WO₃/GO (c) in the acetone concentration range from 80 ppm to 400 ppb with applied individual gate voltages ($V_{GS} \approx V_{dirac}$) at room temperature.

by etching a window on the backside of (0.1cm×0.1cm) SiO₂ layer assisted by a negative photoresist.

D. Material characterizations

Prepared samples were optimized by different characterization techniques. FESEM (field effect scanning electron microscopy) was used to identify the morphology of all three samples at high magnifications. Defects and structural properties of samples were analyzed by Raman spectroscopy and the bandgap of prepared samples was calculated from UV-Vis spectroscopy data.

E. Sensing characterizations

A static gas sensing setup was used to analyze the sensing performance of prepared samples where the sensor was mounted inside a 600 ml volume sealed chamber. The VOC was injected by micro syringes (Hamilton 705RN 50UL SYR). The ppm level of injected VOC was measured by the formula; C (ppm) = $2.46 \times (V_i D / VM) \times 10^3$, where D (gm/mL) is the density of VOC, M (gm/mol) is the molecular weight of injected VOC and V (Lit) volume of sensing chamber [22,23]. The mass flow controller (MFC) was maintaining a 450 SCCM continuous flow of synthetic air to recover the sensors. Two source meters (Keithley 6487) were mounted to measure the FET characteristics and sensing behavior of fabricated

sensors. The response value of FET sensors towards acetone was calculated by RM (%) = $[(\Delta I) / I_v] \times 100$ where ΔI ($I_v - I_a$) is the change in the current of the sensor in VOC and air. The FET devices were optimized at a constant bias voltage ($V_{DS} = 1$ V) with varied gate voltages (V_{GS}). 90 % of the total change in the current value of the sensors was marked as a response and recovery time.

III. RESULTS AND DISCUSSION

A. Material Characterizations

Figure 1 is showing FESEM images of prepared all three samples. Pure WO₃ is showing a layered structure and these layers get aggregated with each other and show flowered morphology as shown in Fig. 1 (a). Fig. 1b is displaying Au and Pd nanoparticles distribution on WO₃ and GO film. GO was showing layered morphology and aggregation of these layers from the continuous film (Fig. 1b and c). Fig. 1c is showing the arrangement of Au and Pt nanoparticles on WO₃ and GO. We have found the uniform, compact and continuous arrangement of nanoparticles (Fig. 1b and c). The average size, 200 nm of the Au metal particle was the largest and Pt nanoparticles with 5 nm of the average size were found smallest in diameter and the Pd nanoparticles were 12 nm in

Table 1. Literature survey on the status of the work on GO based acetone sensor.

Sensing materials	Gas/VOC conc.	Operating temperature	Response (%)	Ref.
Gd-WO ₃ /RGO	50 ppm	200 °C	54	1
ZnO-GO	100 ppm	240 °C	35.8	2
Pt- WO ₃	10 ppm	200 °C	12.2	3
Au-Pt/WO ₃ /GO	80 ppm	RT	56.5	This work

in diameter as shown in Fig. 1b and c.

Raman spectroscopy scan is displayed in (Fig. 2a) to identify the presence of all materials and defect density in samples. Four characteristic peaks of WO₃ were found in all three samples and 2 peaks of GO in metal loaded WO₃/GO samples as shown in Fig. 2a. The four prominent Raman peaks at wavelength values of 131.7, 267.5, 697.9 and 806.2 cm⁻¹ were detected in the resulting samples (Fig. 2a). The peak at a lower wavelength of 131.7 cm⁻¹ is due to the lattice vibrations of WO₃. Other than peak at 267.5 cm⁻¹ is originated due to the δ (O-W-O) bending mode and the peaks at higher wavelength values 697.9 and 806.2 cm⁻¹ are corresponding to the W-O-W stretching vibration modes [6, 13, 15]. The samples with presence of graphene oxide showing the characteristic peaks of material at wavelengths 1336.9 and 1591.2 cm⁻¹ are related to the D band (defect density of GO) and G band (shows the restoration of pi conjugation) [16, 17, 24] as shown in Fig. 2a. The intensity ratio (D peak/G peak) was 1.1 and 1.4 for Au-Pd/WO₃/GO and Au-Pt/WO₃/GO, respectively.

UV-Vis spectroscopy results of all three samples is shown in Fig. 2b. The bandgap of the resulting hybrids were 3.16, 4.7, 4.87 eV for WO₃, Au-Pd/WO₃/GO and Au-Pt/WO₃/GO respectively which is measured from the UV-Vis data by the help of Tauc model and the final equation, αhv = A(hv - E_g)^{1/2} where, A is the material property coefficient, hv is photon energy and E_g is the bandgap. The schematic representation of calculated the bandgap was as shown in Fig. 2b.

B. I-V Characteristics of FET sensors

Fig. 3a is showing I_{DS}-V_{GS} characteristics of both the Au-Pd/WO₃/GO and Au-Pt/WO₃/GO FET sensor in air and 80 ppm acetone at room temperature. *p*-GO and *n*-WO₃ both are semiconducting materials and form discrete Schottky junctions [9, 16, 25]. Change in the baseline current and dirac point position from Au-Pd/WO₃/GO to Au-Pt/WO₃/GO FET sensor was due to the work function difference between materials (Fig. 3a). GO shows ambipolar behavior and when the positive gate bias is applied the device current start decreasing and at dirac point the charge carrier density of GO is very low and the minimum device current in both air and 80 ppm acetone ambient was observed as shown in Fig. 3a [14, 16, 26]. After the dirac point conductivity gets changed from *p* to *n*-type in FET sensors and due to this in I_D-V_{GS} of 80 ppm acetone before the dirac point current was low and after the dirac point current was high compared to I_D-V_{GS} of air as shown in Fig. 3a. Moreover, Au-Pd/WO₃/GO FET sensor was showing dirac point at 1 V and Au-Pt/WO₃/GO FET sensor at 1.2 V (Fig. 3a).

Fig. 3b is showing sensitivity curve of Au-Pd/WO₃/GO and Au-Pt/WO₃/GO at V_{GS} = 0 V and V_{GS} ≈ V_{dirac}. The sensitivity value of the sensors in 80 ppm acetone was ~6.52 % and ~4 % at V_{GS} = 0 V and 46.6 % and 56.5 % at V_{GS} ≈ V_{dirac} for Au-Pd/WO₃/GO and Au-Pt/WO₃/GO, respectively, at room temperature. Moreover, ~7 and ~14 times amplification in

the response magnitude was noted for Au-Pd/WO₃/GO and Au-Pt/WO₃/GO FET sensors at 80 ppm acetone under the influence of optimized gate voltages (V_{GS} ≈ V_{dirac}).

C. Transient behavior study of sensors

Transient behavior of all three sensors was examined towards the acetone concentration range 400 ppb to 80 ppm at room temperature shown in Fig. 4. WO₃ and two FET sensors; Au-Pd/WO₃/GO and Au-Pt/WO₃/GO transient study were performed under constant V_{DS} of 1 V and V_{GS} ≈ V_{dirac}. With the decrease in acetone concentration, the device current also gets decreased in all three sensors. The response for the WO₃ sensor for highest ppm (80) and lowest ppm (0.4) of acetone was 14.1 % and 2 %, respectively as shown in Fig. 4a. While Au-Pt/WO₃/GO FET sensor was showing 56.5 % and 3 % response for 80 ppm and 400 ppb at room temperature and at applied V_{GS} (Fig. 4c). Noble metal functionalized/WO₃/GO FET sensor was showing ~4 times higher response than pure WO₃ at room temperature. All three sensors showed good sensing responses with good stability and repeatability in pulses for a minimum amount of acetone concentration (0.4 ppm) (Fig. 4). Moreover, noble metal nanoparticle functionalization improved the overall sensing performance of FET devices by the electronic and chemical sensitization behavior [14,15,18,20].

Table 1 is the comparison of sensing properties of GO nanocomposites with different sensing materials for acetone detection [1-3]. The reported Au-Pt/WO₃/GO FET sensor have good comprehensive performances including high response and low operating temperature compared to other ones.

IV. CONCLUSION

In this work, we are proposing three sensor systems marked as; pure WO₃, Au-Pd/WO₃/GO FET and Au-Pt/WO₃/GO FET. Among all three sensors, the best acetone sensing properties were observed in Au-Pt/WO₃/GO FET sensor in the influence of noble metal functionalization and applied gate voltage at room temperature. FESEM images confirm the uniform and discrete decoration of noble metal nanoparticles on WO₃ Nanoflowers and GO layer. GO as a channel material was deposited in sheets without any aggregation and discontinuity. Raman spectroscopy results signify the presence of WO₃ and GO in both the Au-Pd/WO₃/GO and Au-Pt/WO₃/GO samples and UV-Vis spectroscopy result was showing a bandgap shift from WO₃ to Au-Pd/WO₃/GO and Au-Pt/WO₃/GO after incorporation of noble metal nanoparticles and GO. The I_D-V_{GS} characteristic study was showing the ambipolar nature of GO at room temperature. An amplified sensitivity was found in Au-Pd/WO₃/GO and Au-Pt/WO₃/GO FET sensors at V_{GS} = 1V and 1.2 V, respectively. Au-Pt/WO₃/GO FET as the best acetone sensor was showing ~14 times higher response for 80 ppm acetone than the response obtained in two-terminal configurations and ~4

times higher than the pure WO₃ sensor at the same operating conditions.

ACKNOWLEDGMENT

This work was supported in part by Department of Biotechnology grant (Letter No. BT/PR28727/NNT/28/1569/2018) and SPARC grant (SPARC/2018-2019/P1394/SL).

REFERENCES

[1] J. Kaur, K. Anand, A. Kaur, and R. C. Singh, *Sensitive and selective acetone sensor based on Gd doped WO₃/reduced graphene oxide nanocomposite*, vol. 258. Elsevier B.V., 2018.

[2] P. Wang *et al.*, “ZnO nanosheets/graphene oxide nanocomposites for highly effective acetone vapor detection,” *Sensors Actuators B Chem.*, vol. 230, pp. 477–484, Jul. 2016, doi: 10.1016/j.snb.2016.02.056.

[3] L. Chen *et al.*, “Fully gravure-printed WO₃/Pt-decorated rGO nanosheets composite film for detection of acetone,” *Sensors Actuators, B Chem.*, vol. 255, pp. 1482–1490, 2018, doi: 10.1016/j.snb.2017.08.158.

[4] F. Liu, X. Chu, Y. Dong, W. Zhang, W. Sun, and L. Shen, “Acetone gas sensors based on graphene-ZnFe₂O₄ composite prepared by solvothermal method,” *Sensors Actuators, B Chem.*, vol. 188, pp. 469–474, 2013, doi: 10.1016/j.snb.2013.06.065.

[5] C. Dong, R. Zhao, L. Yao, Y. Ran, X. Zhang, and Y. Wang, *A review on WO₃ based gas sensors: Morphology control and enhanced sensing properties*, vol. 820. Elsevier B.V., 2020.

[6] J. Zhang *et al.*, “Fabrication of conductive graphene oxide-WO₃ composite nanofibers by electrospinning and their enhanced acetone gas sensing properties,” *Sensors Actuators, B Chem.*, vol. 264, pp. 128–138, 2018, doi: 10.1016/j.snb.2018.02.026.

[7] X. X. Zou *et al.*, “A precursor route to single-crystalline WO₃ nanoplates with an uneven surface and enhanced sensing properties,” *Dalt. Trans.*, vol. 41, no. 32, pp. 9773–9780, 2012, doi: 10.1039/c2dt30748k.

[8] C. Wang, X. Li, C. Feng, Y. Sun, and G. Lu, “Nanosheets assembled hierarchical flower-like WO₃ nanostructures: Synthesis, characterization, and their gas sensing properties,” *Sensors Actuators, B Chem.*, vol. 210, pp. 75–81, 2015, doi: 10.1016/j.snb.2014.12.020.

[9] C. Wang *et al.*, “Hierarchical flower-like WO₃ nanostructures and their gas sensing properties,” *Sensors Actuators, B Chem.*, vol. 204, pp. 224–230, 2014, doi: 10.1016/j.snb.2014.07.083.

[10] M. Righettoni, A. Tricoli, and S. E. Pratsinis, “Si:WO₃ sensors for highly selective detection of acetone for easy diagnosis of diabetes by breath analysis,” *Anal. Chem.*, vol. 82, no. 9, pp. 3581–3587, 2010, doi: 10.1021/ac902695n.

[11] T. Gakhar and A. Hazra, “Oxygen vacancy modulation of titania nanotubes by cathodic polarization and chemical reduction routes for efficient detection of volatile organic compounds,” *Nanoscale*, vol. 12, no. 16, pp. 9082–9093, 2020, doi: 10.1039/c9nr10795a.

[12] H. Gu, Z. Wang, and Y. Hu, *Hydrogen gas sensors based on semiconductor oxide nanostructures*, vol. 12, no. 5. 2012.

[13] A. Esfandiari, A. Irajizad, O. Akhavan, S. Ghasemi, and M. R. Gholami, “Pd-WO₃/reduced graphene oxide hierarchical nanostructures as efficient hydrogen gas sensors,” *Int. J. Hydrogen Energy*, vol. 39, no. 15, pp. 8169–8179, 2014, doi: 10.1016/j.ijhydene.2014.03.117.

[14] R. Bhardwaj and A. Hazra, “Realization of ppb-level acetone detection using noble metals (Au, Pd, Pt) nanoparticles loaded GO FET sensors with simultaneous back-gate effect,” *Microelectron. Eng.*, vol. 256, no. December 2021, p. 111719, 2022, doi: 10.1016/j.mee.2022.111719.

[15] R. Bhardwaj, U. N. Thakur, P. Ajmera, R. Singhal, Y. Rosenwaks, and A. Hazra, “Field-Assisted Sensitivity Amplification in a Noble Metal Nanoparticle Decorated WO₃/GO Hybrid FET-Based Multisensory Array for Selective Detection of Breath Acetone,” *ChemNanoMat*, 2022.

[16] A. Hazra, “Amplified Methanol Sensitivity in Reduced Graphene Oxide FET Using Appropriate Gate Electrostatic,” *IEEE Trans. Electron Devices*, vol. 67, no. 11, pp. 5111–5118, 2020, doi: 10.1109/TED.2020.3025743.

[17] T. Gakhar and A. Hazra, “p -TiO₂ / GO heterojunction based VOC sensors : A new approach to amplify sensitivity in FET structure at optimized gate voltage,” vol. 182, no. January, pp. 0–2, 2021.

[18] R. Bhardwaj, V. Selamneni, U. N. Thakur, P. Sahatiya, and A. Hazra, “Detection and discrimination of volatile organic compounds by noble metal nanoparticle functionalized MoS₂coated biodegradable paper sensors,” *New J. Chem.*, vol. 44, no. 38, pp. 16613–16625, 2020, doi: 10.1039/d0nj03491f.

[19] M. S. Tsai, C. J. Lu, and P. G. Su, “One-pot synthesis of AuNPs/RGO/WO₃ nanocomposite for simultaneously sensing hydroquinone and catechol,” *Mater. Chem. Phys.*, vol. 215, pp. 293–298, 2018, doi: 10.1016/j.matchemphys.2018.05.058.

[20] P. Bindra and A. Hazra, “Electroless deposition of Pd/Pt nanoparticles on electrochemically grown TiO₂nanotubes for ppb level sensing of ethanol at room temperature,” *Analyst*, vol. 146, no. 6, pp. 1880–1891, 2021, doi: 10.1039/d0an01757d.

[21] P. Joshna, A. Hazra, K. N. Chappanda, P. K. Pattnaik, and S. Kundu, “Fast response of UV photodetector based on Ag nanoparticles embedded uniform TiO₂ nanotubes array.”

[22] A. Hazra, “Surface Potential-Based Approach to Estimate Bias Dependent Sensitivity of 1-D Metal Oxide Resistive Gas Sensors,” vol. 20, no. 11, pp. 5766–5775, 2020.

[23] P. Bindra and A. Hazra, “Selective detection of organic vapors using TiO₂ nanotubes based single sensor at room temperature,” *Sensors Actuators, B Chem.*, vol. 290, no. January, pp. 684–690, 2019, doi: 10.1016/j.snb.2019.03.115.

[24] S. J. Choi, B. H. Jang, S. J. Lee, B. K. Min, A. Rothschild, and I. D. Kim, “Selective detection of acetone and hydrogen sulfide for the diagnosis of diabetes and halitosis using SnO₂ nanofibers functionalized with reduced graphene oxide nanosheets,” *ACS Appl. Mater. Interfaces*, vol. 6, no. 4, pp. 2588–2597, 2014, doi: 10.1021/am405088q.

[25] S. Cao, C. Zhao, T. Han, and L. Peng, “Oxalic acid assisted hydrothermal synthesis and optical absorption property of WO₃/TiO₂ nanocomposites,” *J. Mater. Sci. Mater. Electron.*, vol. 27, no. 6, pp. 5635–5639, 2016, doi: 10.1007/s10854-016-4471-z.

[26] U. N. Thakur, R. Bhardwaj, P. K. Ajmera, and A. Hazra, “ANN based approach for selective detection of breath acetone by using hybrid GO-FET sensor array ANN based approach for selective detection of breath acetone by using hybrid GO-FET sensor array,” 2022.

Intelligent Reflecting Surfaces in UAV-Assisted 6G Networks: An Approach for Enhanced Propagation and Spectral Characteristics

Mobasshir Mahbub*, Raed M. Shubair

Department of Electrical and Computer Engineering, New York University (NYU) Abu Dhabi, Abu Dhabi, UAE
 Email: mobasshir@ieee.org*, raed.shubair@nyu.edu

Abstract - Intelligent reflecting surfaces (IRSs) with the ability to reconfigure inherent electromagnetic reflection and absorption characteristics in real-time provide unparalleled prospects to improve wireless connectivity in adverse circumstances. Unmanned aerial vehicles (UAV)-assisted wireless networks are evolved as a reliable solution to combat non-line of sight (NLoS) scenarios. Thereby, the IRS-empowered UAV-assisted cellular networks will be a significant role-player to improve the coverage and user experiences. The paper aimed to minimize the path loss and maximize the achievable data rate in IRS-UAV-assisted networks. In this context, the work analyzed path loss and achievable rate utilizing millimeter wave (mmWave) carrier considering the conventional UAV model and IRS-empowered UAV communication model. The research obtained that the IRS-empowered UAV communications model can significantly minimize path loss and maximize the achievable data rate compared to the conventional UAV-assisted model.

Index Terms – 6G, mmWave, UAV, IRS, Path Loss.

I. INTRODUCTION

The data rate will immensely increase attributed to the prevalence of increased users in hotspot circumstances. To solve the challenge, 6G (sixth generation) systems must migrate to several sophisticated network enhancement techniques e.g. IRS [1], [2], THz communication [3], etc.

UAV technology is regarded as a critical component of 6G mobile networks [4]. In comparison to typical terrestrial cellular connectivity, UAVs can swiftly facilitate connections in coverage region due to their great mobility. It is also cost-effective in regions with insufficient coverage, such as remote or hilly locations. As a result, the design of 6G networks may transition from the stationary terrestrial infrastructure model to aerial mobile connectivity.

IRS [5], [6], also referred to as a metasurface [7], is a newly developed engineered material whose radiation characteristics, such as reflection, phase, absorption, and refraction can be electronically modified in real-time. Because these surfaces are low-cost to produce, they may be distributed globally, giving an unparalleled opportunity to manipulate the wireless multipath scenario both indoors and outdoors. As a result, IRS offers an entirely new avenue in mobile communications research countering the multipath. Indeed, numerous studies have recently proven that IRS-assisted solutions may increase the coverage, capacity, energy, spectral efficiency of contemporary mobile networks dramatically. Fig. 1 visualizes the IRS-empowered UAV-assisted network.

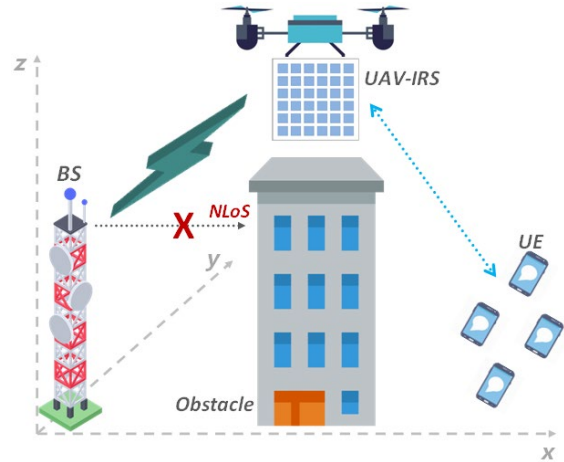


Fig. 1. IRS-empowered UAV-assisted wireless network

UAVs are becoming increasingly used for relaying, data collection, secure communication, and information distribution. Because of their great agility and versatility for on-demand implementation, UAVs have a high likelihood of having line-of-sight (LoS) transmission connections [4]. As a result, UAVs can be used as aerial communication mediums to augment the performance of current ground wireless transmission networks, such as mobile networks.

However, UAV communication confronts several problems, particularly in urban environments. One notable concern is obstruction created by ordinary things including buildings/infrastructure, forests/trees, and human bodies that can exacerbate coverage and connection issues [8]. The movement of both the UAV and the UE (user equipment), on the contrary, causes excessive temporal and spatial changes in the non-stationary communication channels.

To solve these issues, the intelligent reflecting surface (IRS) technology [9] has recently been proposed to avoid impediments and improve connectivity in UAV systems, referred to as IRS-empowered UAV network systems [10]. IRSs are installed in the open network environment to aid the connection between UAVs and users, according to the IRS-assisted UAV architecture. A blocked i.e. non-line of sight (NLoS) transmission channel can be mitigated by deploying the IRS enabling several LoS links, which considerably minimizes channel attenuation.

Therefore, the work targeted minimizing the path loss and maximizing the achievable data rate in IRS-UAV-assisted networks.

II. RELATIVE LITERATURE

The paper in this section briefed several prior works and literature relative to IRS-empowered UAV communication. Since IRS-assisted wireless communication is an emerging topic and research is ongoing, during the literature review the authors find that there is a limitation of literature relative to path loss measurement in the IRS-UAV communication scenario.

Ma et al. [11] analyzed the potential deployment of IRS in cellular communications with UAVs that suffer from degraded signal strength. The work considered the implementation of IRS on walls that can be configured remotely by base stations to coherently transmit the reflected signal towards corresponding UAVs to enhance signal strengths at the user end. Al-Jarrah et al. [12] presented the outage probability and symbol error rate (SER) analysis of multi-layer UAV-empowered wireless communications supported by the IRS. The research of Pang et al. [13] overviewed the amalgamation of UAV and IRS, by illustrating the applications of IRS and the advantages of UAV and describing the advantages of incorporating them in a combined manner in the wireless network. Then, the work investigated case studies namely the UAV trajectory, the passive beamforming in IRS, and the transmit beamforming at the base station are jointly optimized. Jiang et al. [14] proposed and analyzed a three-dimensional (3D) stochastic geometry-based channel model incorporating multiple-input multiple-output (MIMO) for IRS-aided UAV communications. Wei et al. [15] considering the application of IRS in UAV-assisted orthogonal frequency division multiple access (OFDMA) transmission systems first derived the expression to analyze composite channels and proposed an approximation approach to set a lower and an upper bound for the considered problem. Mahmoud et al. [16] investigated the deployment of IRS in UAV-empowered communications networks aiming to enhance the coverage and improve the reliability in terms of spectral efficiency considering the Internet of Things (IoT) paradigm. Specifically, the work first derived tractable analytic expressions and then analyzed the ergodic capacity, achievable SER, and outage probability. Cao et al. [17] proposed and analyzed a 3D non-stationary MIMO channel model for an IRS-assisted UAV communications network. Shafique et al. [18] presented and analyzed a theoretical framework of an IRS-integrated UAV relaying system.

III. MEASUREMENT MODEL

A. Conventional UAV-Assisted Communication Model

Contemplate a 3D geographic coverage region in which terrestrial base stations, UAVs, and user equipment are spanned in a 3D plane. A set of base stations \mathfrak{B} are deployed in the mentioned communication (or coverage) region along with \mathfrak{U} set of UAVs to enhance the coverage (in case of NLoS and obstacle-filled area) for a set of user \mathfrak{R} . Consider $u_i = (x_i^{uav}, y_i^{uav}, z_i^{uav})$ as the location (3D coordinates) of UAV i

$\in \mathfrak{U}$ and $(x_n^{UE}, y_n^{UE}, z_n^{UE})$ as the coordinates denoting the location of user $n \in \mathfrak{R}$. Therefore, according to the ITU (International Telecommunication Union), the path loss is formulated as (Eq. 1) [19],

$$PL = \left(\frac{4\pi f_c d_0}{c} \right)^2 \left(\frac{d_i}{d_0} \right)^2 u_{NLoS} \quad (1)$$

where u_{NLoS} denotes the attenuation factor. c denotes the speed of light in ms^{-1} . f_c indicates the carrier frequency in Hz. d_0 denotes the reference distance (free space). Here, $d_0 = 1$ m. $d_i = \sqrt{(x_n^{UE} - x_i^{uav})^2 + (y_n^{UE} - y_i^{uav})^2 + (z_n^{UE} - z_i^{uav})^2}$ indicates the distance between the UAV i and arbitrary user equipment (UE) located at $(x_n^{UE}, y_n^{UE}, z_n^{UE})$ coordinates. The received power thereby can be obtained as follows (Eq. 2),

$$\wp_r = \frac{\wp_t}{\mathcal{K}_0 d_i^2 u_{NLoS}} \quad (2)$$

where \wp_t is the transmit power and $\mathcal{K}_0 = \left(\frac{4\pi f_c d_0}{c} \right)^2$. The SNR (Signal to Noise Ratio) can be formulated by the following equation (Eq. 3),

$$\zeta = \frac{\wp_r}{\sigma^2} \quad (3)$$

where $\sigma^2 = -90$ dBm denotes the Gaussian noise power. The achievable data rate (bits/s/Hz/m²) by the user equipment can be determined by (Eq. 4),

$$\mathfrak{R} = \log_2 \left(1 + \frac{\wp_r}{\sigma^2} \right) \quad (4)$$

B. IRS-Embedded UAV-Assisted Communication Model

Incorporating the IRS in UAV the far-field beamforming model for determining the path loss in the case of IRS-enhanced UAV-assisted communication model is (Eq. 5) [20], [21],

$$PL_{IRS} = \frac{64\pi^3 (d_1 d_2)^2}{G_t G_r G M^2 N^2 d_x d_y \lambda^2 \cos(\theta_t) \cos(\theta_r) \mathcal{A}^2} \quad (5)$$

where $d_1 =$

$\sqrt{(x^{BS} - x_i^{uav(IRS)})^2 + (y^{BS} - y_i^{uav(IRS)})^2 + (z^{BS} - z_i^{uav(IRS)})^2}$ denotes the separation distance between the transmitter i.e. base station located at (x^{BS}, y^{BS}, z^{BS}) coordinates and IRS-embedded UAV located at $(x_i^{uav(IRS)}, y_i^{uav(IRS)}, z_i^{uav(IRS)})$ coordinates. $d_2 =$

$$\sqrt{(x_i^{uav(IRS)} - x_n^{UE})^2 + (y_i^{uav(IRS)} - y_n^{UE})^2 + (z_i^{uav(IRS)} - z_n^{UE})^2}$$

denotes the separation distance between the IRS and the UE located at $(x_n^{UE}, y_n^{UE}, z_n^{UE})$. G_t and G_r are the gains of the transmitter and receiver and $G = \frac{4\pi d_x d_y}{\lambda^2}$ is the scattering gain. M and N are the numbers of transmission and reception elements of the IRS. d_x and d_y are the length and width of each scattering element. λ denotes the wavelength of the transmitted signal. θ_t and θ_r are the transmitter and receiver angles with the IRS. \mathcal{A} denotes the amplitude relative to the reflection coefficient of unit cells of the IRS. Thereby, the received power at the UE can be formulated by (Eq. 6),

$$\rho_r^{IRS} = \frac{\rho_t G_t G_r G M^2 N^2 d_x d_y \lambda^2 \cos(\theta_t) \cos(\theta_r) \mathcal{A}^2}{64\pi^3 (d_1 d_2)^2} \quad (6)$$

The SNR can be measured by the following formula (Eq. 7),

$$\mathcal{S}_{IRS} = \frac{\rho_r^{IRS}}{\sigma^2} \quad (7)$$

The achievable can be calculated by the following equation (Eq. 8),

$$\mathfrak{R}^{IRS} = \log_2 \left(1 + \frac{\rho_r^{IRS}}{\sigma^2} \right) \quad (8)$$

IV. RESULTS AND DISCUSSIONS

This section includes the measurement results based on the measurement model utilizing MATLAB-based simulation and incorporates discussions on the derived simulation result. The research considered low-altitude UAV since it utilized mmWave carrier.

Fig. 2 (a) and (b) show the path loss in terms of transmitter-receiver separation distance with 2D and 3D figures respectively in the case of conventional UAV-assisted communication networks. $f_c = 100$ GHz, $(x_i^{uav}, y_i^{uav}, z_i^{uav}) = (0, 0, 40)$ m, $(x_n^{UE}, y_n^{UE}, z_n^{UE}) = (0-100, 0-100, 1.5)$ m, $u_{NLoS} = 23$ dB indicate the simulation parameters.

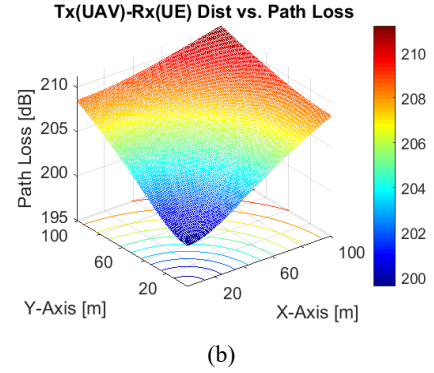
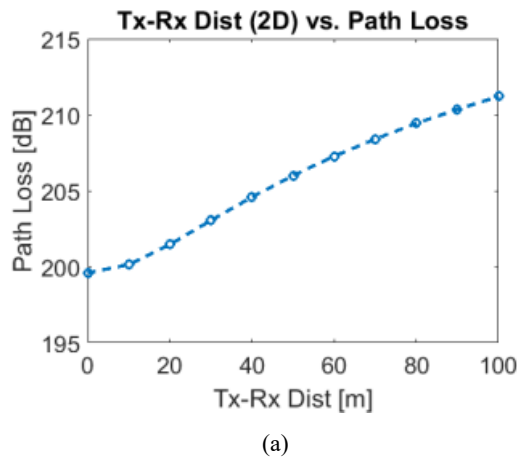


Fig. 2. (a) Tx-Rx separation distance vs. path loss (2D); (b) Tx-Rx separation distance vs. path loss (3D)

Fig. 3 (a) and (b) represent the achievable data rate in terms of transmitter-receiver separation distance with 2D and 3D figures respectively considering the conventional UAV-assisted communication model. For this measurement $\rho_t = 6$ W (other parameters are as same as Fig. 2).

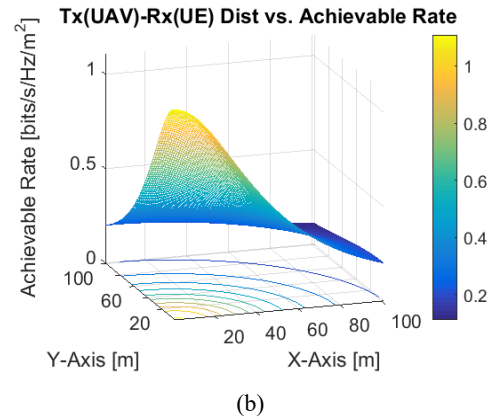
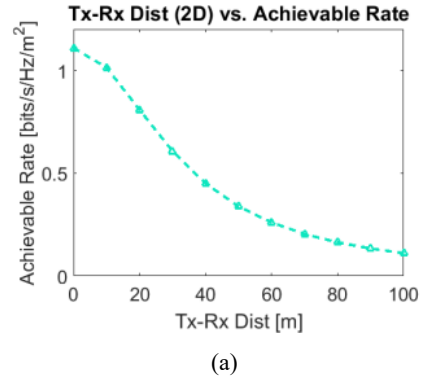
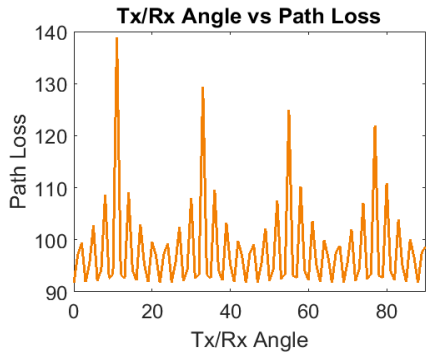


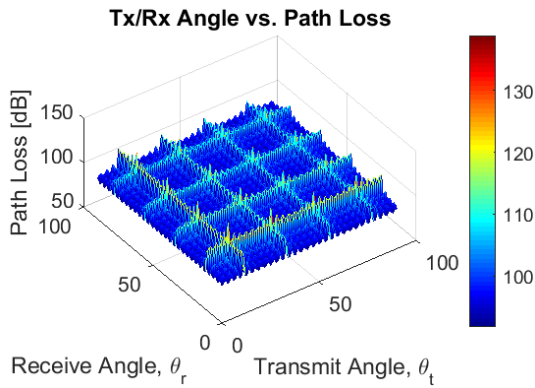
Fig. 3. (a) Tx-Rx separation distance vs. achievable rate (2D); (b) Tx-Rx separation distance vs. achievable rate (3D)

Fig. 4 (a) illustrates the path loss in terms of transmit-receive angle with a 2D figure for better realization and Fig. 4 (b) shows the path loss in terms of transmit-receive angle with 3D figures for 0° to 90° angles (from BS to IRS and IRS to UE) respectively in the context of IRS-empowered UAV communications model. $f_c = 100$ GHz, $(x^{BS}, y^{BS}, z^{BS}) = (0, 0,$

8) m , $(x_i^{uav(IRS)}, y_i^{uav(IRS)}, z_i^{uav(IRS)}) = (50, 50, 40)$ m, $(x_n^{UE}, y_n^{UE}, z_n^{UE}) = (100, 100, 1.5)$ m, $G_t = 20$ dB, $G_r = 20$ dB, $\mathcal{M} = 100$, $\mathcal{N} = 100$, d_x & $d_y = \lambda/2$, $\mathcal{A} = 0.9$ (tx-rx angles are varied).



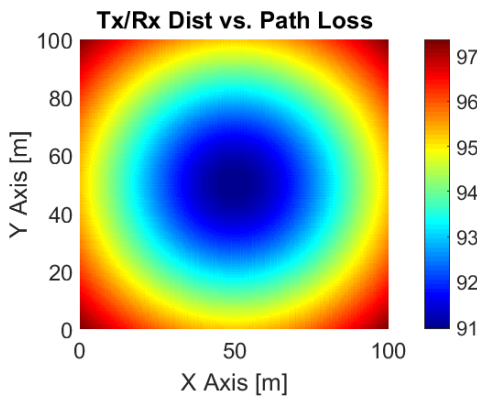
(a)



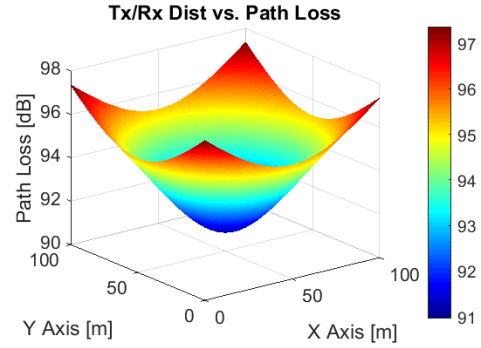
(b)

Fig. 4. (a) Tx-Rx angle vs. path loss (2D); (b) Tx-Rx angle vs. path loss (for angle 0° to 90°)

Fig. 5 (a) and (b) visualize the path loss with top view and 3D figures in terms of transmitter-receiver separation distance (from BS to IRS and IRS to UE) in the case of IRS-empowered UAV communication networks. θ_t & $\theta_r = 45^\circ$, $x_i^{uav(IRS)}$ & $y_i^{uav(IRS)}$ are varied only, (other parameters are as same as Fig. 4).



(a)



(b)

Fig. 5. (a) Tx-Rx separation vs. path loss (top view); (b) Tx-Rx separation vs. path loss (3D)

Fig. 6 represents the path losses in terms of the number of transmit-receive elements of IRS. \mathcal{M} & \mathcal{N} are varied only, (other parameters are as same as Fig. 4 & angles of Fig. 5).

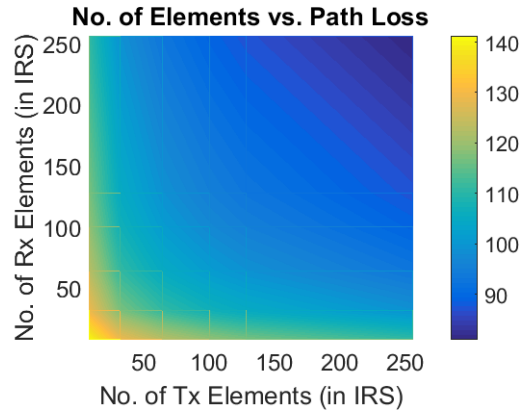
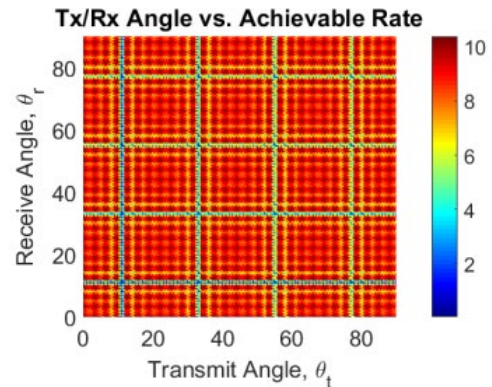


Fig. 6. No. of Tx-Rx elements vs. path loss

Fig. 7 (a) shows the achievable data rate in terms of transmit-receive angle with a top view figure and Fig. 7 (b) shows the path loss in terms of 0° to 90° transmit-receive angles in the context of the IRS-empowered UAV communications model. $\rho_t = 2$ W (parameters are as same as Fig. 4).



(a)

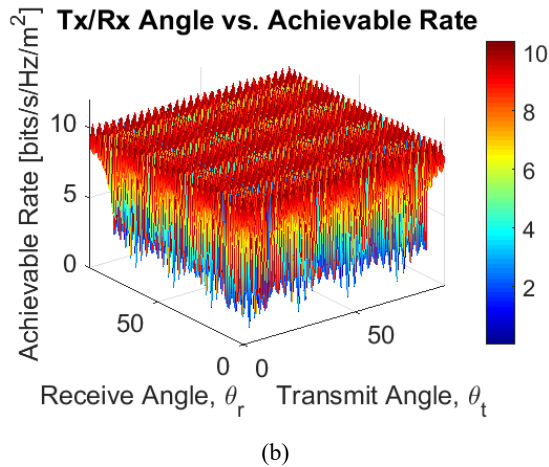


Fig. 7. (a) Tx-Rx angle vs. achievable data rate (top view); (b) Tx-Rx angle vs. achievable data rate (for angle 0° to 90°)

Fig. 8 (a) and (b) represent the achievable data rate with top view and 3D figures in terms of transmitter-receiver separation distance in the case of IRS-empowered UAV communication networks. $\rho_t = 2W$ (parameters are as same as Fig. 5).

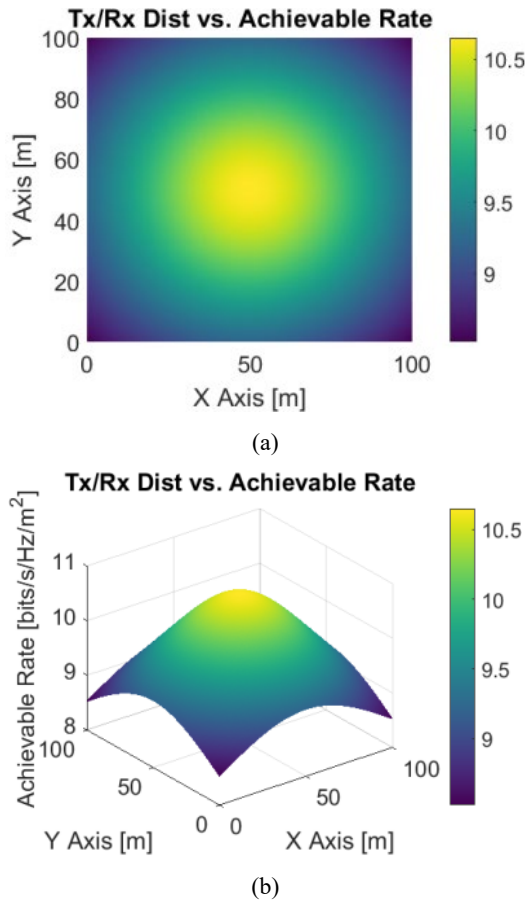


Fig. 8. (a) Tx-Rx separation vs. achievable data rate (top view); (b) Tx-Rx separation vs. achievable data rate (3D)

Fig. 9 illustrates the achievable rate in terms of the number of transmit-receive elements of IRS. $\rho_t = 2W$ (parameters are as same as Fig. 6).

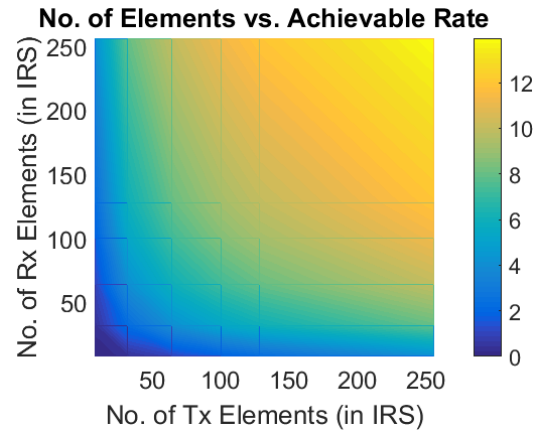


Fig. 9. No. of Tx-Rx elements vs. achievable data rate

Through the observation of Figs. 4 and 7, it is evident that 11°, 33°, 55°, and 77° angles exhibits the highest path loss and lowest achievable data rate. 0°, 22°, 44°, 66°, 88° angles of transmissions and reception produce the lowest path loss and highest achievable data rate. But the path loss is quite lower and the achievable rate is much higher than the case of the conventional UAV communication model. According to the prior research 30° to 60° transmit-receive angle [21], [22] is prominent for IRS communication.

It can be interpreted by examining Figs. 5 and 8 that, with the increase of transmitter-receiver separation distance the path loss increases and the achievable data rate gradually decreases. Since the IRS enables almost perfect reflection of the transmitted signal, reduces the absorption, and establishes an LoS condition a greater number of users can obtain a significantly favorable signal (in the considered coverage region) compared to the conventional UAV communication model (comparing with Figs. 2 and 3).

According to the observation of Figs. 6 and 9 it is comprehensible that, with the increase of the number of transmit-receive elements in IRS the path loss decreases, and the achievable data rate increases gradually.

Observing and comparing the figures finally it is evident that, the deployment of IRS reduces the path loss by around 70 – 100 dB and increases the achievable data rate up to 8 – 13 times approximately (according to the chosen network parameter).

An intriguing observation of this research is that by utilizing a lower transmit power a significantly higher achievable data rate can be obtained by deploying IRS-empowered UAV communication networks (2W) compared to the conventional UAV communication model (6W).

The research considered low altitude UAVs since it utilized mmWave carrier to minimize the significant attenuation of the carrier. Moreover, the work considered a limited interference scenario.

Furthermore, the research considered IRS (UAV) 3D location in the reference formula [20] which makes the measurement more precise.

V. CONCLUSION

The research work targeted minimizing the path loss and maximizing the achievable data rate of conventional UAV-assisted communication by deploying IRS-empowered UAV-assisted wireless networks. The work, therefore, analyzed and compared the deployment of IRS-UAV communication and conventional UAV communication considering the aforementioned measurement model. Performing MATLAB-based measurements the research obtained that IRS-empowered UAV communication networks minimize the path loss and maximize the achievable data rate significantly. Moreover, the IRS-UAV communication model reduces the energy consumption of the entire network. Further research can be performed on the analyzed model considering higher mmWave and THz-band carriers, multiple UAVs, higher altitude, etc. The research will be assistive to the scholars, researchers, and enthusiasts who are engaged in the relative research.

ACKNOWLEDGEMENT

The authors would like to express gratitude to the Department of Electrical and Computer Engineering, New York University (NYU) Abu Dhabi, Abu Dhabi, UAE.

REFERENCES

- [1] S. Gong et al., "Toward Smart Wireless Communications via Intelligent Reflecting Surfaces: A Contemporary Survey," in *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2283-2314, Fourthquarter 2020.
- [2] M. Di Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and The Road Ahead," in *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450-2525, Nov. 2020.
- [3] H. Elayan, O. Amin, B. Shihada, R. M. Shubair and M. -S. Alouini, "Terahertz Band: The Last Piece of RF Spectrum Puzzle for Communication Systems," in *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1-32, 2020.
- [4] A. Masaracchia et al., "UAV-Enabled Ultra-Reliable Low-Latency Communications for 6G: A Comprehensive Survey," in *IEEE Access*, vol. 9, pp. 137338-137352, 2021.
- [5] C. Pan et al., "Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions," in *IEEE Communications Magazine*, vol. 59, no. 6, pp. 14-20, June 2021.
- [6] Q. Wu, S. Zhang, B. Zheng, C. You and R. Zhang, "Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial," in *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313-3351, May 2021.
- [7] S. Zhang et al., "Intelligent Omni-Surfaces: Ubiquitous Wireless Transmission by Reflective-Refractive Metasurfaces," in *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 219-233, Jan. 2022.
- [8] A. A. Khuwaja, Y. Chen, N. Zhao, M. -S. Alouini and P. Dobbins, "A Survey of Channel Modeling for UAV Communications," in *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2804-2821, Fourthquarter 2018, doi: 10.1109/COMST.2018.2856587.
- [9] X. Pei et al., "RIS-Aided Wireless Communications: Prototyping, Adaptive Beamforming, and Indoor/Outdoor Field Trials," in *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8627-8640, Dec. 2021.
- [10] X. Pang, M. Sheng, N. Zhao, J. Tang, D. Niyato and K. -K. Wong, "When UAV Meets IRS: Expanding Air-Ground Networks via Passive Reflection," in *IEEE Wireless Communications*, vol. 28, no. 5, pp. 164-170, October 2021.
- [11] D. Ma, M. Ding and M. Hassan, "Enhancing Cellular Communications for UAVs via Intelligent Reflective Surface," *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1-6.
- [12] M. Al-Jarrah, A. Al-Dweik, E. Alsusa, Y. Iraqi and M. S. Alouini, "On the Performance of IRS-Assisted Multi-Layer UAV Communications With Imperfect Phase Compensation," in *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8551-8568, Dec. 2021.
- [13] X. Pang, M. Sheng, N. Zhao, J. Tang, D. Niyato and K. -K. Wong, "When UAV Meets IRS: Expanding Air-Ground Networks via Passive Reflection," in *IEEE Wireless Communications*, vol. 28, no. 5, pp. 164-170, October 2021.
- [14] H. Jiang, R. He, C. Ruan, J. Zhou and D. Chang, "Three-Dimensional Geometry-Based Stochastic Channel Modeling for Intelligent Reflecting Surface-Assisted UAV MIMO Communications," in *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2727-2731, Dec. 2021.
- [15] Z. Wei et al., "Sum-Rate Maximization for IRS-Assisted UAV OFDMA Communication Systems," in *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2530-2550, April 2021.
- [16] A. Mahmoud, S. Muhaidat, P. C. Sofotasios, I. Abualhaol, O. A. Dobre and H. Yanikomeroglu, "Intelligent Reflecting Surfaces Assisted UAV Communications for IoT Networks: Performance Analysis," in *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1029-1040, Sept. 2021.
- [17] C. Cao, Z. Lian, Y. Wang, Y. Su and B. Jin, "A Non-Stationary Geometry-Based Channel Model for IRS-Assisted UAV-MIMO Channels," in *IEEE Communications Letters*, vol. 25, no. 12, pp. 3760-3764, Dec. 2021.
- [18] T. Shafique, H. Tabassum and E. Hossain, "Optimization of Wireless Relaying With Flexible UAV-Borne Reflecting Surfaces," in *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 309-325, Jan. 2021.
- [19] M. Mozaffari, W. Saad, M. Bennis and M. Debbah, "Optimal Transport Theory for Cell Association in UAV-Enabled Cellular Networks," in *IEEE Communications Letters*, vol. 21, no. 9, pp. 2053-2056, Sept. 2017.
- [20] W. Tang et al., "Path Loss Modeling and Measurements for Reconfigurable Intelligent Surfaces in the Millimeter-Wave Frequency Band," 2021, arXiv: 2101.08607v2. [Online]. Available: <https://arxiv.org/abs/2101.08607>
- [21] W. Tang et al., "Wireless Communications With Reconfigurable Intelligent Surface: Path Loss Modeling and Experimental Measurement," in *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 421-439, Jan. 2021.
- [22] Ö. Özdoğan, E. Björnson and E. G. Larsson, "Intelligent Reflecting Surfaces: Physics, Propagation, and Pathloss Modeling," in *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 581-585, May 2020.

Intelligent Feature Selection on Multivariate Dataset using Advanced Data Profiling

¹ *Abstract*—The differential diagnosis of diseases which share similar clinical features is a real and difficult problem in medicine. This paper demonstrates the use of data mining (DM) techniques to augment standard data profiling methods and establishes an efficient approach for an intelligent feature selection method for disease that share similar features. The results from experiments returned show that by using DM techniques to select features as an additional layer on top of data profiling, there is considerable improvement in the performance of the prediction model built to predict a disease such as “Psoriasis”. A brief comparison between features selected by existing mining tools such as Weka and the proposed approach with respect to predictive accuracy is recorded in this paper. The proposed algorithm works as a promising tool for assisting diagnosis of disease like erythemato-squamous diseases, where the symptoms are overlapping. By combining data cleansing and knowledge discovery techniques, the algorithm aims to be “agnostic” and can be used on a wide variety of data standards with variable data quality.

I. INTRODUCTION

Machine learning models are built to capture relationships in an n -dataset (dataset of size n) between its $n-1$ features and the (n th) target feature in the dataset. Since such models are highly data-driven, their accuracy relies on how “good” the data is. “Goodness” of any dataset can be determined by understanding the data and preparing it so that it can be deemed “good”. Typically, repositories such as Data Lakes [15] store data in its raw form, which is eventually (pre)processed before it is used in a data mining model. To understand any given data in a dataset requires gathering information about that data, commonly known as metadata. Type of metadata depends on the application for which the data mining model is being designed. As an example, let us consider an application that requires educational data, such as student information, the courses they have completed and grades achieved in these courses. Metadata in this example can be as simple as the student’s name and age stored in its raw form, or it could require some calculation such as average age of students in this dataset. If described and managed well by big data repositories such as data lakes, such metadata can be exploited by analysts to discover underlying relationships between different features in a dataset, and thereby allow data mining models to be better informed about the dataset. Data profiling refers to this structured activity of creating small but informative summaries of a database [8]. Data profiling uses scientific methods to

explore, understand and collect statistics on raw data of a given dataset for detecting statistical distributions and structural patterns.

In a typical real-world scenario, data is integrated from various different sources, and this makes it challenging to provide consistent, clean and accurate data for machine learning models. Data profiling can be used as a pre-processing step in the process of data mining to evaluate datasets for consistency, uniqueness and quality of data. It is a well known fact that data mining models spend 70% of their time in the pre-processing step which clearly needs to be improved [15]. This paper proposes a data agnostic algorithm for data profiling. In this algorithm, a full understanding of the entire data set is not required. The algorithm uses data mining techniques to understand the data (we call it as intelligent data profiling) and use that to select features that could prove to have a higher impact on the accuracy of the final model that is being built. One may argue that the high cost of computational complexity of profiling large datasets may deter data enthusiasts from using it. But this can be overcome by leveraging massively parallel clusters and will not be addressed in this project. Our research aims to build an intelligent data profiling algorithm that evaluates and analyzes different attributes or features in the dataset, with an overall objective of selecting those features that are of good quality and that improve the accuracy of the predictive model built using these features.

This paper is organized in the following way: Section II explains the proposed methodology. Sections III illustrates the evaluation of the proposed algorithm. A literature review of the related work and their limitations is discussed in section IV. Section V presents the conclusions and proposes future work.

II. PROPOSED METHODOLOGY

This section presents the dataset used and proposed methodology for intelligent data profiling and feature selection.

A. The Dataset

The dataset used in this research is on dermatology and is taken from an open-source repository [4]. It includes 34 clinical and histopathological features that have been transformed in a variety of ways to suit this research.

¹978-1-6654-8684-2/22/\$31.00 ©2022 IEEE

1) *Rationale for using the dataset:* Diagnosing skin diseases such as 'Psoriasis' and 'lichen planus' is a real problem in dermatology [5]. Firstly, they all share the clinical features with very little differences. Secondly, there are too many features, some of them do not have enough values or have values that are not meaningful. Some diseases have overlapping features from another disease at the beginning stage and may have the characteristic features at the subsequent stages. In terms of data, the significance of the presence or absence of a feature in the dataset makes the analysis challenging. For example, given the features, it is difficult to differentiate between 'Psoriasis' or 'Lichen Planus' as they have overlapping features (for example, itching is a common feature in both). The proposed method intelligently selects only those features that have a significant impact on the accuracy of the predictive model that predicts the disease.

2) *Notation used :*

1) The following notation is used:

D = dataset

2) Each instance n of D is represented as a triplet:

$\forall_1^n i, (ca_i, ha_i, ta_i)$

where ca represents clinical attributes, ha represents histopathological attributes and ta represents the target attribute. $|x|$ indicates the cardinality or number of instances of x

3) D^T – dataset after transformation of attributes

4) $D_{profiled}^T$ - dataset after data profiling of the attributes

5) $D_{reduced}^T$ - optimal set of attributes used for prediction

B. Transformation of selected features in dataset D

In general, preprocessing is used to transform raw input data into appropriate formats for subsequent mining [14]. Real-world data is often inconsistent, has missing values and may have data that has errors. Data preprocessing is a technique that is used to resolve such issues.

The transformations on different attributes of D and the rationale for doing so are listed below.

1. The target attribute (ta) in the original dataset [4] has 6 classes. For simplicity, we transform ta to a binary class code problem, i.e. to 2 classes. We aim to predict if a person with several given symptoms (represented by a vector of clinical and histopathological attributes) has psoriasis (class code = 1) or not (class code = 2).

2. Attribute age is a clinical attribute (ca) that is continuous. In order to see the impact of age categories on the outcome (i.e. presence of psoriasis), we chose to discretize age into 5 different bins. This decision was a result of a few experiments with attribute age. Binning is a method to group a number of continuous values into a smaller number of "bins". This research creates bins using equal-width unsupervised discretization method (except for the last bin that is of a larger size than others) [10]. The bins created using this algorithm and their distribution with the target attribute are as follows:

Bin	Interval
1	[< 12.5]
2	[12.5 - 25]
3	[25 - 37.5]
4	[27.5 - 50]
5	[>= 50]

The third bin that caters to the age group 25 - 37.5 has the maximum count of patients among all bins, whereas age group less than 12.5 has the least. The graph in figure 1 shows the distribution of all age groups.

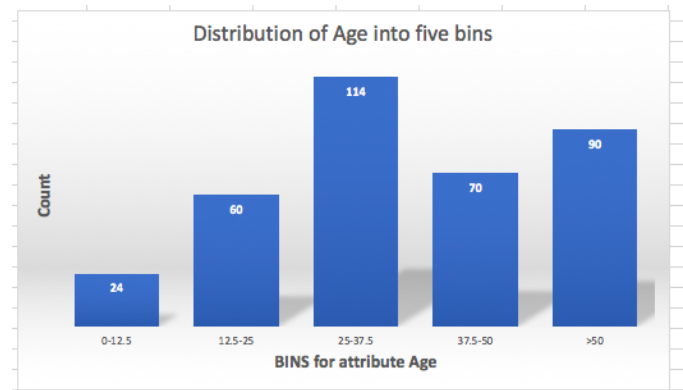


Figure 1. Discretization of Age into bins

3. All other clinical (ca) and histopathological (ha) attributes in the original dataset D can have any of the four values (0, 1, 2 or 3, where 0 indicates the absence of a symptom, 1, 2, 3 indicate the intensity of the symptom present, with 3 representing the highest intensity and 1 the lowest). All these attributes can be categorized as asymmetric attributes.

An asymmetric attribute is defined as an attribute in which the presence of one of the values (e.g. 1) is regarded as more significant than the other (e.g. 0), as opposed to a symmetric attribute in which the value 1 (or higher indicating presence of the attribute) is considered equally significant as its absence (0) [14]. For example, erythema in D is taken to be an asymmetric attribute in this research, where a value of 1 or higher indicates the presence of that symptom and 0 indicates its absence. In an asymmetric attribute, a 1-1 match of erythema in two rows is significant, whereas a 0-0 match has no significance (since 0 implies that the symptom is not present) and is ignored in the proposed method. On the contrary, an example of a symmetric attribute is gender, where 1 represents female and 0 represents male, then a 1-1 match (indicating a female-female match) is as significant as a 0-0 match (indicating a male-male match).

After the above transformation, our dataset now has:

- 11 clinical attributes: asymmetric
- 22 histopathological attributes: asymmetric
- 1 family income: binary symmetric attribute
- 4 binary asymmetric attributes: unique bin for age category

It is worth mentioning here that id numbers of patients are removed from the database in the original dataset D.

C. The proposed algorithm - Predicting Using Intelligent Feature Selection (PIFS)

Algorithm 1 lists the core steps of the proposed algorithm called PIFS (Predicting using Intelligent Feature Selection). Step 1 transforms the original dataset D to D^T . Section B defines the transformation on some of the attributes such as age and the target attribute (ta) and the rationale behind those transformations.

Step 2 performs the data profiling step to convert D^T to $D^T_{profiled}$. We did single-column profiling tasks in terms of the number of rows and uniqueness of the values [1]. Those features or columns that have 80% or more values as 0 (0 indicates absence of that symptom) were eliminated and not included in the new set $D^T_{profiled}$. At the end, the attributes selected for the next step were reduced to a new dataset ($D^T_{profiled}$).

Dataset $D^T_{profiled}$ is certainly more informed than D^T , however, many of its attributes still have a large number of zero values that indicate the absence of a symptom. In order to make our dataset more informed, this research runs an intelligent algorithm that is based on the distribution of different categorical values (0, 1, 2, 3) in each ca or ha attribute in $D^T_{profiled}$. Step 3 of PIFS (Algorithm 1) illustrates its details. Thresholds 1 and 2, selected by combining expert opinion with experiments, are given below. Here, \wedge represents AND; \vee represents the OR symbol.

$$\begin{aligned}
 & threshold_1 : |zeros| \geq 250 \\
 & threshold_2 : 150 < |zeros| < 250 \wedge \\
 & (|ones| \vee |twos| \vee |threes| > 50)
 \end{aligned}$$

These thresholds are determined by the attributes in the current dataset and by the different categorical values that these attributes can have. Figure 2 shows two attributes that do not meet these threshold values and therefore are not selected by PIFS. Figures 3 and 4 illustrate distribution for four attributes that were selected by the proposed algorithm. As can be seen, the number of zeros in both Erythema and Scaling (Figure 3) is negligible. Similarly, figure 3 shows two attributes that were filtered out by our algorithm since they do not meet the two thresholds.

Step 4 creates a predictive model using k-nearest neighbor algorithm to predict the target (psoriasis or not) given a vector from $D^T_{reduced}$. This research chose to use k-nearest-neighbor due to its simplicity. Its results were compared to decision tree model as well (as shown in the experimental section).

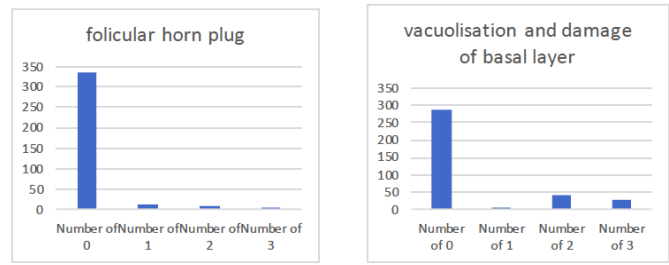


Figure 2. attributes not selected by PIFS

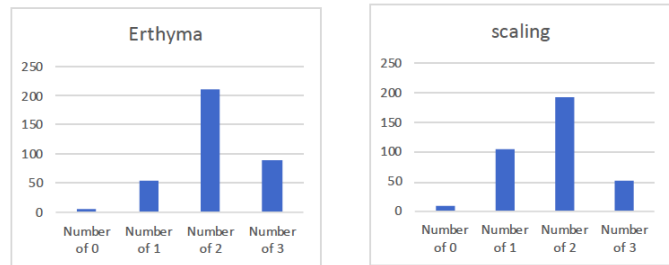


Figure 3. attributes selected by PIFS

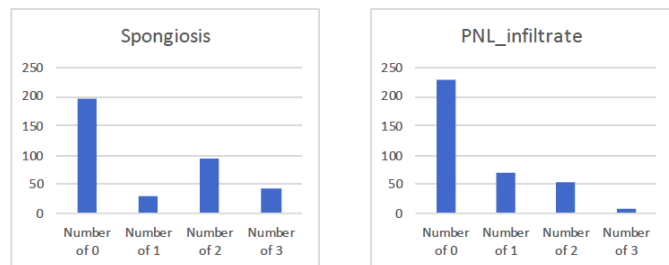


Figure 4. additional attributes selected by PIFS

K-nearest-neighbor (k-nn) algorithm [14] typically takes 4 inputs in order to predict a target attribute: an integer k (k=number of neighbors), a set of training samples whose target attribute y is known, a test vector t and a similarity or distance function. It then predicts test sample of class label t by performing the following steps :

- 1) calculate similarity between a test vector t and all training samples using the chosen similarity function (as explained below)
- 2) sort these similarity values and pick the top k samples - these are the k nearest neighbors of test sample t
- 3) Use the values of ta of each of the k nearest neighbors of t to predict a value for t. If the total number of neighbors that have a 'Yes' as its target attribute is greater than the total number of neighbors that have a 'No', assign 'Yes' as the class label of t; otherwise assign a 'No'.

Jaccard's Coefficient (JC) works best with asymmetric attributes [14] and therefore is more applicable to this research. Jaccard's Coefficient (JC) between two vectors x and y is

measured as

$$JC(x, y) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} \quad (1)$$

where f_{11} is the frequency of 1-1 match in vectors x and y , f_{01} is the frequency of occurrence of 0 and 1 in x and y (non-matching pair) and f_{10} is the frequency of occurrence of 1 and 0 in x and y (non-matching pair). For example, if $x = [1, 0, 0, 1]$ and $y = [1, 0, 1, 0]$, then $JC(x, y) = 1/3$. Since all the attributes in D are categorical and asymmetric, the similarity function that we choose to use in the proposed algorithm PIFS is Jaccard's coefficient of similarity.

Algorithm 1: The proposed algorithm PIFS

1. Preprocess and prepare attributes in the original dataset D (call it D^T) - as explained in section 2.2
2. Perform SCP (single-column profiling) data profiling algorithm on D^T to get a new set of attributes (call it $D_{profiled}^T$) - as explained in section 2.3
3. Use an intelligent feature selection method to eliminate those that are (lets call it $D_{reduced}^T$) - as explained in section 2.3

```

for each attribute  $i$  in  $D_{profiled}^T$  do
    find the distribution of its different categories
    if  $|zeros| < threshold_1$  then
        Add  $i$  to  $D_{reduced}^T$ 
    else
        if  $|two| > threshold_2$  OR  $|three| > threshold_2$ : then
            Add  $i$  to  $D_{reduced}^T$ 
        end
    end
end

```

4. Create a predictive model to predict the target attribute (presence or absence of psoriasis) - as explained in section 2.3
- Divide data into 2 subsets using an 80-20 split: training (80% - $D_{reduced}^{Training}$) and test (20% - $D_{reduced}^{Test}$).
- Training dataset is used to build the model, whereas test is used to test the model.
- Build a model by applying a prediction algorithm such as k-nearest-neighbors on the training dataset $D_{reduced}^{Training}$.
- Apply the model to each vector in the test dataset $D_{reduced}^{Test}$ to predict its ta values.

	Features selected by Weka	Number of zeros
1	Erythema	4
2	Itching	116
3	Scaling	8
4	Follicular-Papules	325
5	PNL-filtrate	229
6	Fibrosis of the papillary dermis	308
7	Clubbing of the rete ridges	248
8	Thinning of the supra-papillary epidermis	249
9	Follicular horn plug	336
10	Perifollicular parakeratosis	337

Table I
FEATURES SELECTED BY WEKA

	Features selected by PIFS	Number of zeros
1	Erythema	4
2	Itching	116
3	Scaling	8
4	Definite Borders	55
5	Exytosis	117
6	Acanthosis	9
7	Parakeratosis	85
8	Inflammatory mononuclear infiltrate	9
9	Spongiosis	195
10	Age	Categorical

Table II
FEATURES SELECTED BY PIFS

III. EVALUATION OF PIFS

The proposed algorithm is evaluated by the following criteria: (1) feature selection (steps 1, 2 and 3 of PIFS) (2) performance of the prediction algorithm (step 4 of PIFS) using measures such as accuracy, recall and fscore [11].

A. Feature selection

Intelligent feature selection is a core step of PIFS, as described in section 2. The selected features are compared with features selected by Weka [16]. Weka is a collection of machine learning algorithms for data mining tasks such as data preparation, feature selection, prediction and clustering.

The features (also referred to as attributes in this paper) selected by Weka using its feature selection algorithm (cfs-SubsetEval + BestFirst) and the number of zeros in each of the selected feature is shown in table I. Note that zeros in this dataset indicate the absence of a feature (or a symptom) of the target disease being predicted (e.g., Psoriasis). All features other than Erythema and Scaling have a very high number of zeros. This implies that Weka is not able to accurately distinguish between the features as they are extremely similar and overlapping – this finding concurs with the claim made by the creators of the dermatology dataset D[4]. Attributes selected by PIFS are shown in table II. The selected features in this table have less number of zeros (0 indicates absence of symptoms) than the features selected by Weka.

B. Performance measurement:

Accuracy measures the ability of the model to match the actual value of the target attribute with its predicted one (e.g.

"Yes" predicted as "Yes"). Accuracy alone is not always a deterministic measure, especially when dealing with target attribute values that are imbalanced or uncommon. For example, let's assume that in a dataset with 100 samples, there are 10 people with target attribute = "Yes" and 90 with a target attribute = "No". Also assume that the model predicts 1 out of 10 people correctly as "Yes", and 90 out of 90 as "No", then the accuracy of the model according to its definition (equation 2) will be calculated as high as 91%. However, this result is misleading as the model is not at all accurate in predicting one of the target attribute values (in this case "Yes"). Other measures that are used to evaluate predictive models are precision and recall. Most models achieve a trade-off between precision and recall, since it is very challenging to keep both the measures high. F-score is a combined measure that assesses this trade-off between precision and recall. Figures 6 and 7 show high values of accuracy and fscore for the proposed model.

$$accuracy = \frac{ActualTrueValuesPredictedCorrectly}{NumberOfPredictions} \quad (2)$$

$$precision = \frac{ActualTrueValuesPredictedCorrectly}{PredictedTrueValues} \quad (3)$$

$$recall = \frac{ActualTrueValuesPredictedCorrectly}{ActualTrueValues} \quad (4)$$

$$f\ score = \frac{2}{\frac{1}{r} + \frac{1}{p}} \quad (5)$$

Performance of PIFS is compared with the models created using Weka. Weka gives over 98% accuracy with classification models such as Naïve Bayes, iBK and J48 (decision tree) using the selected attributes shown in table I. Most likely, this is because weka is unable to distinguish between the presence or absence of a feature (symptom). We also compared the accuracy of PIFS using selected features and all 34 features in the original dataset D. PIFS accuracy with the selected features is as high as 92%, as compared to 60% when all features are used, as shown in figure 5, asserting that PIFS selects its features intelligently. Figure 5 also shows that the accuracy is highest at k = 7 with PIFS features.

IV. RELATED WORKS

In this section, existing work on data profiling and their limitations are discussed. We first describe the different applications of data profiling and thereafter describe the recent works done to cater to these applications. Commonly, data profiling is the set of actions required to determine metadata about a given dataset. Determining metadata requires computation on rows of the given dataset (e.g. finding rows that have missing values such as identifiers) or more commonly on columns in the dataset (single or multiple). In a survey on data

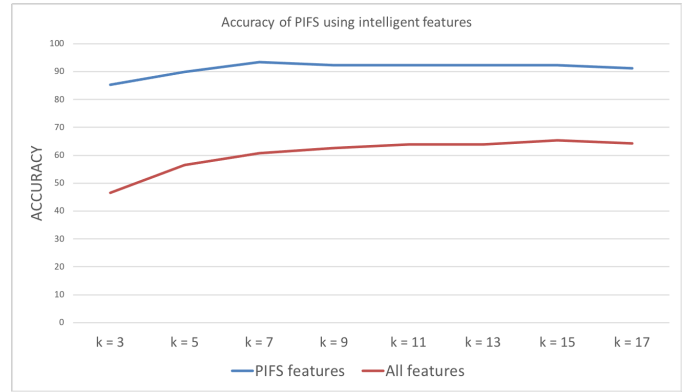


Figure 5. Accuracy of PIFS using intelligent features

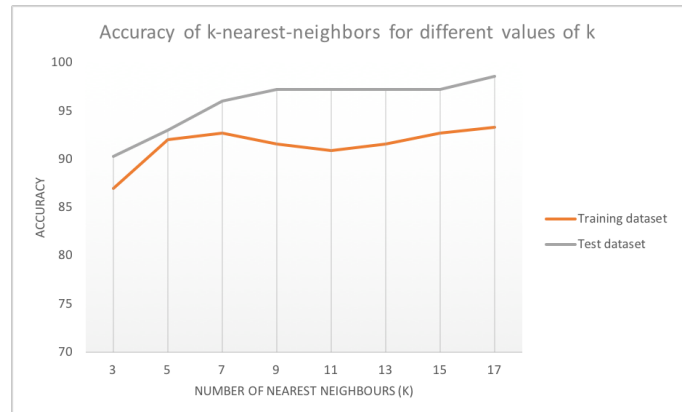


Figure 6. Accuracy of kNN model

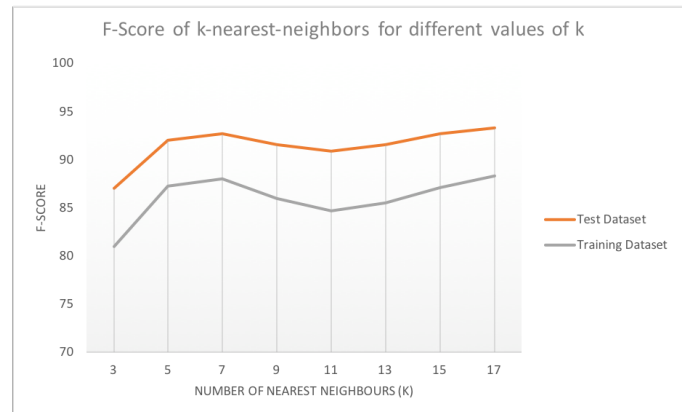


Figure 7. Fscore of kNN model

profiling [1], the authors classify data profiling tasks as single-column and multi-column profiling, depending on the use case of the task. A traditional application of profiling is in data exploration. Most times, database experts including researchers are faced with a collection of data with not much information about the data due to the proliferation of data generated by

machines. Two main challenges in data exploration are as follows: (1) to assume that the data to be explored is in a well-structured format [12] and (2) to use manual methods such as data gazing to explore data. Data gazing [3], involves experts manually skimming through the dataset to understand its characteristics. In order to do this in an automated way, experts need to know databases well, so that they can run SQL queries to get better insight of the data. But this poses an additional constraint that explorers always have to be database experts [7]. Gathering statistics about the data (typically called as metadata in DBMS) is a common task done by DBMS. Metadata could be on single columns (e.g. number of unique values) or on multiple columns (e.g. find pairs or groups of columns that can be used to optimize queries by DBMS). Such metadata is also useful in doing reverse database engineering (e.g. creating ER models from an existing dataset) [6][1]. The most common application of data profiling is data integration, where data from different sources and with possible different formats are integrated into a single source. When performing integration, columns or features of two schema from different sources are compared to find the matching ones [13]. Data cleansing is considered as an essential step in the process of knowledge discovery and data mining and the need to collect insight into data (using data profiling techniques) is inevitable [9]. Besides these applications, data profiling is extremely useful in Big Data analytics [2]. Operations such as storing and fetching big data are much more expensive than traditional structured systems. Since the data here is typically huge and heterogeneous, gaining an insight into the data and cleaning it before using it to store or query gives some hope to big data researchers [1]. Most of the methods of data profiling listed above follow the traditional cycle of extracting metadata and then using it towards an application.

The methods described above fail to address the volume and heterogeneity of datasets today. If traditional data profiling methods are followed, it would take days to achieve any significant information about the data. To keep up with these challenges, keeping in mind the computational cost of profiling is covered, this paper proposes a progressive algorithm that blends the traditional methods with state-of-the-art data mining methods (particularly supervised mining) to extract metadata. The proposed method is extensible and scalable and can be applied to very large datasets to generate results in a timely manner.

V. CONCLUSION AND FUTURE WORK

The importance of data profiling as a pre-processing step of machine learning models is well known. In this paper, we demonstrate that layering supervised data mining techniques to data profiling, thereby building intelligence in the feature selection process, results in higher accuracy of prediction of differential diagnosis of diseases where there are several overlapping features. The intelligent feature selection algorithm PIFS selects features from a dataset based on the data and its

distribution. No prior knowledge of the dataset is needed. The algorithm safely ignores “noisy data” and selects an optimal and reduced set of features that yields a highly accurate predictive model.

As future work, we plan to test and train the algorithm to achieve better accuracy for a variety of multivariate datasets where differential diagnosis of disease pose a challenge due to overlapping features. The proposed algorithm can also be extended to multiple classes in order to discriminate between other diseases.

REFERENCES

- [1] ABEDJAN, Z., GOLAB, L., AND NAUMANN, F. Profiling relational data: a survey. *The VLDB Journal* *The International Journal on Very Large Data Bases* 24, 4 (2015), 557–581.
- [2] AGRAWAL, D., BERNSTEIN, P., BERTINO, E., DAVIDSON, S., DAYAL, U., FRANKLIN, M., GEHRKE, J., HAAS, L., HALEVY, A., HAN, J., ET AL. Challenges and opportunities with big data 2011-1.
- [3] ARKADY, M. Data quality assessment.
- [4] DUA, D., AND KARRA TANISKIDOU, E. UCI machine learning repository, 2017.
- [5] GÜVENIR, H. A., DEMİRÖZ, G., AND ILTER, N. Learning differential diagnosis of erythematous-squamous diseases using voting feature intervals. *Artificial intelligence in medicine* 13, 3 (1998), 147–165.
- [6] HAINAUT, J.-L., HENRARD, J., ENGLEBERT, V., ROLAND, D., AND HICK, J.-M. Database reverse engineering. In *Encyclopedia of database systems*. Springer, 2009, pp. 723–728.
- [7] HANRAHAN, P. Analytic database technologies for a new kind of user: the data enthusiast. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (2012), ACM, pp. 577–578.
- [8] JOHNSON, T. Database reverse engineering. In *Encyclopedia of database systems* (2009), Springer.
- [9] KANDEL, S., PARIKH, R., PAEPCKE, A., HELLERSTEIN, J. M., AND HEER, J. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), ACM, pp. 547–554.
- [10] LIU, H., HUSSAIN, F., TAN, C. L., AND DASH, M. Discretization: An enabling technique. *Data mining and knowledge discovery* 6, 4 (2002), 393–423.
- [11] MARKOV, Z., AND LAROSE, D. T. *Data mining the Web: uncovering patterns in Web content, structure, and usage*. John Wiley & Sons, 2007.
- [12] MORTON, K., BALAZINSKA, M., GROSSMAN, D., AND MACKINLAY, J. Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment* 7, 6 (2014), 453–456.
- [13] NAUMANN, F. Data profiling revisited. *ACM SIGMOD Record* 42, 4 (2014), 40–49.
- [14] TAN, P.-N. *Introduction to data mining*. Pearson Education India, 2018.
- [15] TERRIZZANO, I. G., SCHWARZ, P. M., ROTH, M., AND COLINO, J. E. Data wrangling: The challenging journey from the wild to the lake. In *CIDR* (2015).
- [16] WITTEN, I. H., FRANK, E., HALL, M. A., PAL, C. J., AND DATA, M. Practical machine learning tools and techniques. In *DATA MINING* (2005), vol. 2, p. 4.

Probing the states around the charge neutrality point of reduced graphene oxide with time-resolved gated Kelvin Probe Force Microscopy

1st Ragul S

Department of Electrical Engineering Indian Institute of Technology Madras
Chennai, India

ee16d031@smail.iitm.ac.in

2nd Soumya Dutta

Chennai, India

s.dutta@ee.iitm.ac.in

3rd Debdutta Ray

Chennai, India

dray@ee.iitm.ac.in

Abstract—In this work, we performed gated Kelvin Probe Force Microscopy on reduced graphene oxide thin-film transistors with time transient. This enabled us to probe the electronic density of states around the charge neutrality point in reduced graphene oxide thin film. The charge neutrality point is of significance to know the nature of the intrinsic doping of the thin film and the switching of the majority carriers in the transistor devices. We measured the transfer characteristics of the reduced graphene oxide transistor devices to estimate the intrinsic charge neutrality point. The results were in good agreement with the time-resolved gated surface potential measurements obtained using Kelvin probe force microscopy. We propose that gated time-resolved measurement of these semi-metals can be an effective tool to study the nature of electronic states.

Keywords—reduced graphene oxide, gated KPFM, DFT, density of states

I. INTRODUCTION

The discovery of a process to yield graphene from bulk graphite has led to an immense amount of research in this field [1]. Even after such extensive study, an effective method to grow large-area pristine graphene toward electronic application is not viable till now. In order to compensate for this, a derivative called reduced Graphene Oxide (rGO) with similar electrical properties to that of graphene was investigated [3]. Though the production of graphitic oxide from graphite has been around for about a century, the adoption of the Hummer's method [10] to yield a one-molecule-thick version of the substance known as graphene oxide [2] and its reduction to get graphene-like material was made only in later years after the experimental discovery of graphene. Hence, there is considerable interest in graphene oxide

(GO) and rGO for application in multiple fields [11]–[14]. Though there is a lot of research in this field, a better understanding of the electrical properties of the rGO is needed for its effective application in the field of electronics. The hysteresis in the transfer characteristics of Field Effect Transistor(FET) devices made of these materials as the active layer is one of the undesired properties. In order to understand the source and mechanisms involved with the hysteresis in these graphene-based FETs, various techniques have been explored towards realizing hysteresis-free devices in the future [15], [16]. However, there are very few efforts made towards the improvement of rGO based FETs though it finds application in many areas which demand a hysteresis-free device operation.

We make use of KPFM to probe information around the charge neutrality point, which corresponds to the minimum current point in the transfer characteristics of rGO based transistors. We employ an external gating to the thin film in the KPFM set up to capture the time transients of the change in work function for the applied change in gate voltage which we will call the g-KPFM.

The gate voltage was applied in steps of 1V, and the contact potential difference between the rGO film and the AFM cantilever probe was measured continuously. This is a direct measure of the change in work function due to the applied gate voltage stress. The obtained transients reflect that the charges are not able to respond instantaneously to the applied change in gate voltage which is also a signature of the rGO based transistors. The hysteresis in the transfer characteristics of these transistors has been reported and observed in our rGO

thin film transistor that was studied in this work.

II. EXPERIMENT

The fabrication starts with the deposition of the back aluminum gate after SiO_2 etching on the backside by protecting the top gate oxide with photoresist of $\text{SiO}_2/\text{p-Si}$ substrate. After this, the photoresist was removed with hot acetone and dried. The rGO film was coated using a two-step wet chemical method. The procedure involves the spin coating of GO dispersion in methanol followed by a vapor phase reduction in a closed environment of Hydriodic acid and acetic acid mixture for 12 hours. The rGO channel ($100\mu\text{m} \times 259\mu\text{m}$) formation along with top metal contacts (source/drain) follows the procedure reported elsewhere [24].

For g-KPFM measurements, the SiO_2/Si with aluminum back contact on the common gate silicon side and rGO coating on the SiO_2 gate oxide side was used. On top of rGO, a Chromium/Gold ($20/100\mu\text{m}$) pad was evaporated through a shadow mask onto the rGO surface to get the top contact grounding for the g-KPFM measurements as shown in appendix Fig. A1. The back Al gate is connected to a variable dc voltage source through the common ground. The frequency modulated (FM) KPFM was performed to probe the surface topography and the surface contact potential difference (V_{CPD}) of the rGO film. The working of FM-KPFM requires two ac voltage signals of different frequencies coupled with two lock-in amplifiers. These ac signals were applied along with the dc control voltage to the tip. The control voltage along with the z-piezo scanner was modulated to compensate for the change in the two ac voltages. The corresponding measure obtained at their resonant frequencies gives the topography and that obtained with dc control voltage correction gives the contact potential difference of the sample surface from that of the tip. A simplified schematic of the lock-in circuitry loop for measuring the V_{CPD} is shown in Fig. A1 in the appendix. The measurements were done with the Park NX10 system operated in the non-contact FM-KPFM mode. We used a standard conducting tip made of silicon coated with Chromium and Platinum of thicknesses 5nm and 30nm respectively as procured from Parksystems with the commercial name Multi-75E which has a resonant frequency of 70kHz (used to capture the topography). The second lock-in amplifier for measuring V_{CPD} was operated at a much lower frequency of 17kHz . Initial calibration of the tip work function was done with a freshly cleaved surface of highly oriented pyrolytic graphite standard whose work function is fixed at 4.5eV . The calculated work functions were obtained

using the relation $\phi_{tip} - \phi_{sample} = qV_{CPD}$, where ϕ_{tip} & ϕ_{sample} are the work function of tip and the sample respectively and V_{CPD} represents the contact potential difference obtained from the second lock-in circuitry. All the measurements were done at room temperature and pressure in a controlled Nitrogen ambient inside a glove box setup. The relative humidity levels were maintained below 5% during all the measurements.

III. COMPUTATIONAL METHODS

All the density functional theory (DFT) based calculations for obtaining the density of states were done with Quantum Espresso package [19], [20]. A simplified rGO model consisting of a hexagonal honeycomb carbon sheet with oxygen atom attachments to account for the semi-reduced complex functional group linkages as described elsewhere [5], [6] was adopted for this study. A top view of its ball-stick model is shown in Fig. 1(b). A dense k-point grid of 72×72 along the in-plane direction with Generalised Gradient approximation was considered [21]. The crystal structure was relaxed with Broyden–Fletcher–Goldfarb–Shanno algorithm [18] through self-consistent calculations.

The obtained density of state curves were fitted at and around the experimentally obtained charge neutrality point (V_{CNP}) data (refer Fig. 2) in the range 7V to 8V. The fitting to conserve the charges was performed with the condition describe in (1).

$$\frac{\Delta Q}{\Delta V_g} = \text{constant} \quad (1)$$

Here ΔQ is the change in induced charges due to the change in applied gate voltage ΔV_g . This implies that the total charge obtained by integrating the area under the density of states curve should be a constant for an equal change in gate voltage. The ΔQ obtained by integrating the area between the work functions measured at 7V and 8V under steady-state was kept as the reference. The work function values corresponding to other voltage step were obtained by fitting onto the curve by preserving the value of ΔQ , which equals the area under the curve bounded by the corresponding vertical lines in Fig. 1. The estimated position of the Fermi level as a function of the applied discrete voltage steps is in good agreement with the gate voltage-dependent change in work function measured under the steady-state condition, which confirms the reliability of the DFT calculations. The slopes at these fitting points were obtained and plotted against the corresponding experimental g-KPFM measurements after normalization at the peak value at the 7V to 8V step, as shown in Fig. 3.

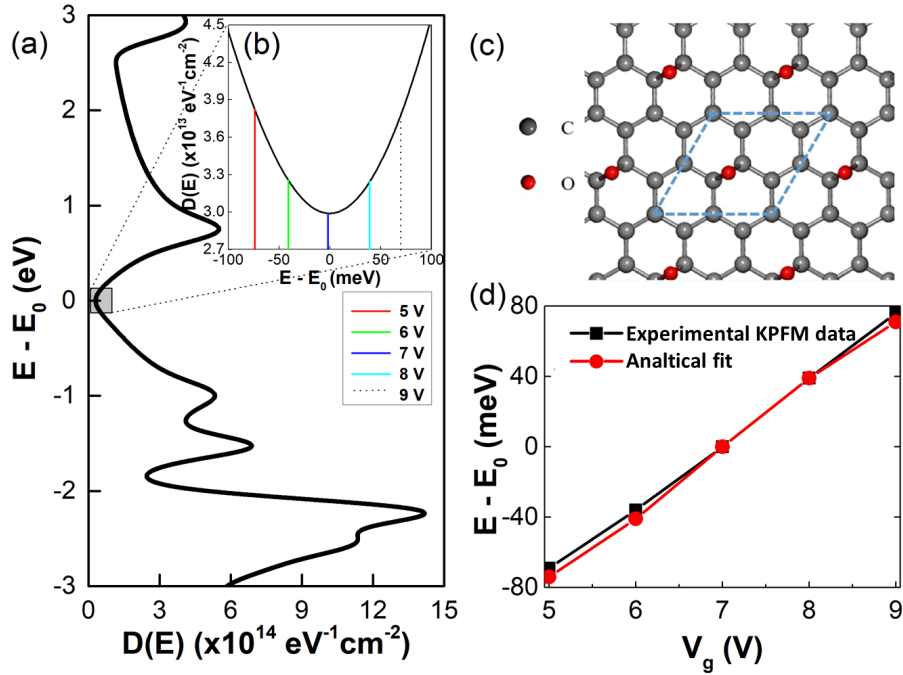


Fig. 1. DFT calculation and analytical modelling results: (a) density of states distribution obtained with DFT calculations ($E - E_0$ is the relative energy with respect the Dirac point (E_0) equivalent of graphene, E being the absolute energy); (b) enlarged image of the highlighted rectangular region in (a) along with the fitted points for $V_g = 5$ to 9 V sequentially represented by red, green, blue, cyan and dotted black vertical lines; (c) ball-stick representation of the rGO model used for DFT calculations; (d) relative work function at V_g from 5 V to 9 V obtained experimentally with FM-KPFM (black) and theoretically (red) with fitting shown in (b)

Further, a projection of these obtained energy values from reduced graphene oxide density of states onto that of graphene (obtained using DFT: Gr- DFT), i.e., the inverse of the slope corresponding to the steady-state energy was estimated and plotted in Fig. 4. In addition, tight binding approximations based on Huckel's model were calculated. Of the two tight-binding fittings shown in Fig. 4, one corresponds to Cerda Graphitic carbon (Gr - TB1), and the other being Hoffmann basis sets (Gr - TB2).

IV. RESULTS AND DISCUSSIONS

The transfer characteristics of the rGO channel-based FET is as shown in Fig. 2 under an applied drain to source voltage of $10V$. It shows a minimum current corresponding to the charge neutrality point (V_{CNP}) of the material between $7V$ and $8V$. The sweeping rates were set to $300mV/s$ and the starting voltage at $\pm 20V$ so that the forward and the reverse sweep almost match, implying that there was enough time to overcome the effect of hysteresis. Thus obtained hysteresis free transfer

characteristics can give the intrinsic doping of the rGO independent of the field-induced transient effects. A positive V_{CNP} at about $+7V$ indicates an intrinsic n-type doping of the rGO [7]–[9]. Further, the presence of hysteresis was observed in the transistor devices by repeating the measurement at faster rates which shows a clear shift in the V_{CNP} point in opposite directions during the forward and backward sweeps away from the former V_{CNP} value. These shifts indicates the inability of the states to respond to the fast-changing applied gate voltages, which accumulate over the previous voltage stresses [16] and leads to the hysteresis effects. The source of this might be the traps or defects present in the system. Hence a lower voltage rate was used to extract the intrinsic V_{CNP} signature from the transfer characteristics of the film, which signifies the ambipolar switching behavior in these materials.

In the g-KPFM analysis, the instance of the time transient right after the applied step voltage change ($t = 0^+$) was considered to remove the effect of slow responding

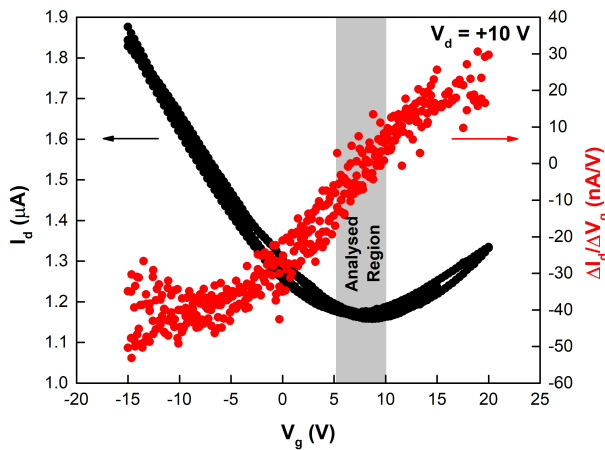


Fig. 2. The black curve represents the I_d vs. V_g double sweep / transfer characteristics on rGO FET measured over a range of ± 20 V and the red curve corresponds to the $\Delta I_d / \Delta V_g$. The highlighted area shows the range of interest for this study

trap/defect states and probe only the intrinsic states of rGO film. The transient g-KPFM on rGO film was done with switching the voltage in steps of $+1V$ transitions and recording the V_{CPD} continuously as a function of time. Fig. 3 shows the change in work function as a time transient for the applied unit step voltage change for different initial voltages. The applied gate voltage switching was very steep of values less than 1 V/ms. The transient shows a dependence on the voltage switching regime, which might be a reflection of the difference in the density of the states to be filled/emptied and to arrive in equilibrium with the change in applied gate voltage. The slope near the edge of the switching point will reflect the nature of the voltage stress and the effects of nature of the distribution of intrinsic states in rGO film. The initial slopes at $t = 0^+$ were plotted against the initial equilibrium gate voltage values and are shown in Fig. 4.

The comparison of the experimental and the theoretical values in Fig. 3(a) can be explained as follows. The experimentally obtained normalized change in work function (Normalized $\Delta\phi_f$ in Fig. 1b) corresponds to $\Delta\phi_f / \Delta t$ which for a small Δt at $t = 0^+$, where the effect of the applied step voltage change will be dominant. Hence, the rate of change of measured work function is proportional to the number density of states at that instant, which needs to be filled/emptied assuming the amount of induced charge is fixed for all unit step voltage independent of the initial gate voltage. The DFT fitting data was obtained by taking the inverse of the slope at the fitted points on the density of states curve, which equals $\Delta\phi_f / \Delta D(E)$ where $D(E)$ is the absolute density

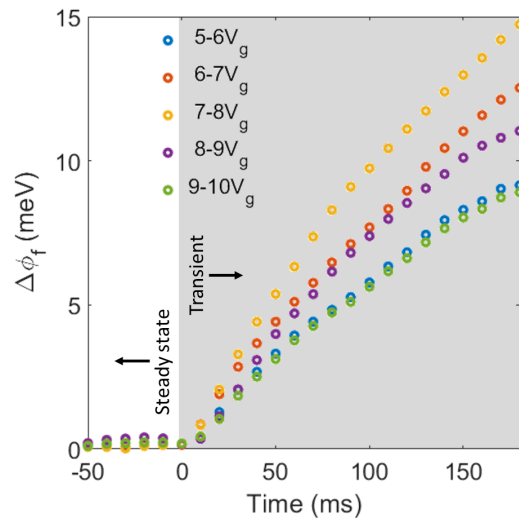


Fig. 3. Time transient of work function change on application of unit voltage steps on the rGO thin film surface

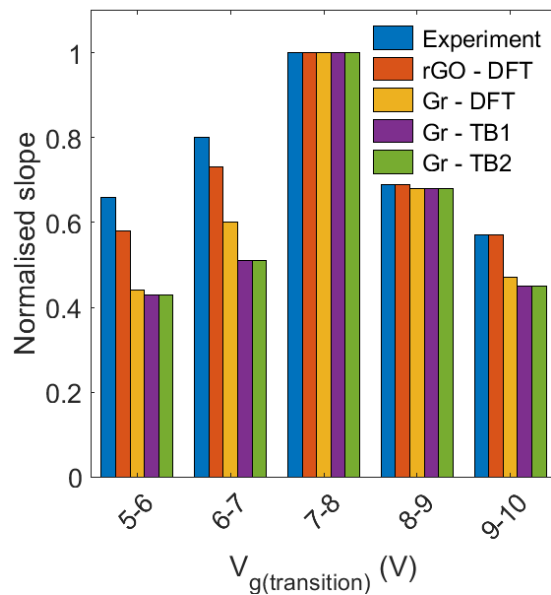


Fig. 4. A comparison of the fitted data obtained with first principle calculations represented against the experimental data; TB1 and TB2 represents the tight binding models corresponding to Cerda Graphitic carbon and Hoffmann based calculations respectively

TABLE I
LIST OF THE EXTRACTED DENSITY OF STATES AND THE FERMI POSITION AT THE VARIOUS APPLIED GATE VOLTAGES

Parameters	Model	Gate Voltage Step (V)				
		5-6	6-7	7-8	8-9	9-10
$\phi_f - V_{CNP}$ (meV)	Experiment	-57	-25	0 ^a	34	76
	DFT	-69.1	-37.2	0.5	34.5	66.5
	TB1	-68.3	-48.3	-1.0	33.0	68.0
	TB2	-69.1	-48.8	-1.0	33.0	68.2
Density of States ($\times 10^{-3} eV^{-1} atom^{-1}$)	rGO	14.2	12.7	12.1	12.9	14.7
	Gr	11.0	9.7	9.1	9.5	10.6
	TB1	11.0	10.1	9.1	9.4	10.6
	TB2	11.0	10.1	9.1	9.4	10.6

^aReference value set to zero; TB1 - Cerda (Graphite) tight binding model; TB2 - Hoffmann tight binding models

of states. $D(E)$ can be related to the absolute charge as $Q = qN(E)$ where q is the charge of an electron equals 1.6×10^{19} eV and $N(E)$ is the number of states obtained by integrating the area under $D(E)$. Thus for a fixed $\Delta D(E)$ one can estimate the $\Delta\phi_f$ corresponding to the fixed change in induced charges which is then normalized against the experimental $\Delta\phi_f$ at the peak point of 7V - 8V step.

Table I lists the absolute values of the voltage, energy, and density of states obtained by curve fitting onto the DFT and tight-binding model-based data by projecting the experimental data as stated in the previous section. These data points correspond to the five vertical lines in Fig. 1c and those encircled points in Fig. 1d.

The amplitude of the normalized $\Delta\phi_f$ increases till the 7V - 8V and then decreases, implying that the number of states decreases and then increases clearly reflecting the V_{CNP} point similar to the data obtained with the transfer characteristics of the fabricated rGO transistor devices. The estimates corresponding to the pristine Graphene sheet also show similar trends. The tight-binding models (Gr - TB1 & TB2) show almost symmetric results with a slight deviation from those obtained with first principle calculations (Gr - DFT). However, the tight-binding approximation showed more deviations from the experimental measure as expected due to their pronounced approximation for simplicity, unlike the DFT-based calculations. This implies that the states of the oxygenated bulk of the rGO film do not contribute to the slow responding trap states; rather, it should be from the surface, interface, or gate oxide.

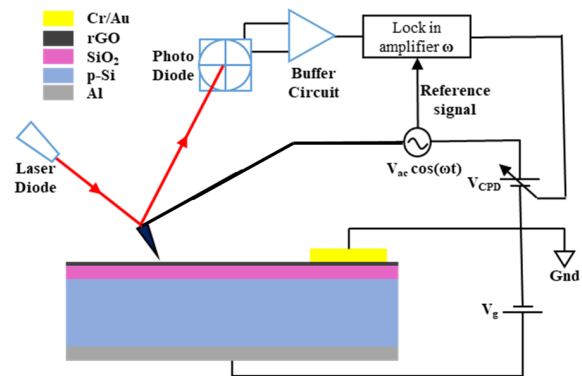
V. CONCLUSION

With the help of the g-KPFM measurements, we were able to capture the time transients of the surface potential of rGO film with the applied change in gate voltages. These measurements clearly show a varying response

that depends on the difference between the applied gate voltage and the voltage corresponding to the intrinsic charge neutrality point of the rGO film over the general operating range of these devices. We have performed density functional theory based calculations with analytical fitting to support the observed experimental data on the states around charge neutrality point in reduced graphene oxide thin films, and they are in good agreement with each other. The fabricated transistor devices on these films showed a transfer characteristic behavior with the charge neutrality point in agreement with that obtained from the transient g-KPFM measurements.

APPENDIX

The general schematic of the circuitry and instrumentation employed for g-KPFM measurements is shown in appendix Fig. A1. A step voltage setup was used in the transient measurements at the source V_g .



Appendix Fig. A1. Schematic representation of g-KPFM measurement setup employing frequency modulation technique for non-contact measurements

Appendix Fig. A2 shows the Energy Dispersion analysis with peaks at Atomic numbers of carbon and oxygen,

which forms the composition in the rGO film understudy. The signal from the SiO₂/Si substrate is corrected by post-processing with the modified form of the model described by Pinos et al [22]. The modification imposed to this random walk simulation of backscattered electron (BSE) assumes no significant contribution from the rGO, which might attribute to 3-5% error for 5 nm thick rGO film. Further, a discrete boundary between SiO₂ and Si was taken at 90 nm (represented by the dashed vertical black line in appendix Fig. A2b-inset) to account for the few nanometer thick SiO_x resulting from the thermal growth process. In the case of a solid-solid interface, it can be assumed that there is no secondary electron emission from silicon. Hence, it has been estimated that the recesses of the total energy for 90 nm thick oxide from that of infinite thickness contribute to BSE in Si. In the case of SiO₂, the effective atomic number was taken as the weighted average of atomic numbers of oxygen and silicon corresponding to their stoichiometry. From the results of this analysis, a concoction of carbon and oxygen in the ratio of 8:1 as represented by the unit cell in Fig. 1b to account for the 10-15% of oxygen-based functional group in them, for the DFT calculations of rGO films. Appendix Fig. A2c-inset shows the AFM height profile of the rGO film on SiO₂/Si substrate. The topographic image was obtained after oxygen plasma-assisted patterning through a metal masking layer deposited on the film followed by metal etching with wet chemicals method to get the rGO step.

ACKNOWLEDGMENT

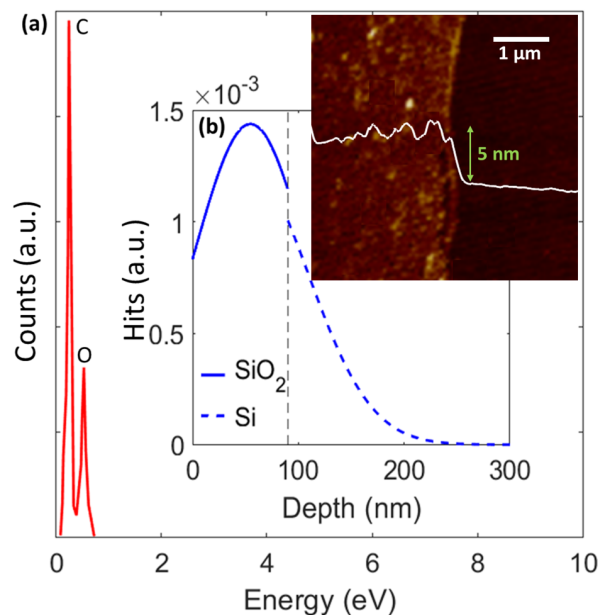
We thank the financial support from the Ministry of Human Resource and Development (MHRD), Ministry of New and Renewable Energy (MNRE), and TATA Steel Pvt. Ltd., India. We also acknowledge the fabrication and characterization facility provided by the Center for NEMS and Nano-Photonics (CNNP) and Microelectronics Laboratory at IIT Madras. Ragul S thanks the Innovation in Science Pursuit for Inspired Research (INSPIRE) fellowship scheme by the Department of Science and Technology (DST), India for the financial support. We also thank the P. G. Senapathy Center for Computing Resource, IIT Madras for the super-computing facility.

REFERENCES

[1] K. S. Novoselov, A. K. Geim, S. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. Grigorieva, A. A. Firsov, "Electric Field Effect in Atomically Thin Carbon Films," *Nat. Mater.*, vol. 6, pp. 666–669, January 2004.

[2] A. M. Dimiev, J. M. Tour, "Mechanism of Graphene Oxide Formation," *ACS Nano*, vol. 8(3), pp. 3060–3068, 2014.

[3] M. Y. Song, Y. S. Yun, N. R. Kim, H. J. Jin, "Dispersion stability of chemically reduced graphene oxide nanoribbons in organic solvents," *RSC Adv.*, vol. 6(23), pp. 19389–19393, 2016.



Appendix Fig. A2. (a) Energy Dispersion Spectrum of the rGO thin film; (b) generated analytical profile of BSE emission as a function of depth in Si/SiO₂ substrate (used for correcting the EDS signal from substrate); (c) AFM topography of the rGO thin film step (5 nm) on Si/SiO₂ substrate

[4] N. J. Lee, J. W. Yoo, Y. J. Choi, C. J. Kang, D. Y. Jeon, D. C. Kim, S. Seo, H. J. Chung, "The interlayer screening effect of graphene sheets investigated by Kelvin probe force microscopy," *Applied Physics Letters*, vol. 95(22), pp. 222107, 2009.

[5] A. Nourbakhsh, M. Cantoro, T. Vosch, G. Pourtois, F. Clemente, M. Veen, J. Hofkens, M. Heyns, S. D. Gendt, B. Sels, "Bandgap opening in oxygen plasma-treated graphene," *Nanotechnology*, vol. 21, pp. 435203, October 2010.

[6] J. Polfus, O. Løvvik, P. Rørvik, R. Bredesen, "Nanocomposites of few-layer graphene oxide and alumina by density functional theory calculations," *Journal of the European Ceramic Society*, vol. 36, November 2015.

[7] T. J. Ha, "Hybrid Graphene/Fluoropolymer Field-Effect Transistors With Improved Device Performance," *IEEE Electron Device Letters*, vol. 62(12), pp. 4340–4344, 2015.

[8] B. S. Park, T. J. Ha, "Charge Transport Properties of Improved Reduced-Graphene-Oxide-Based Field-Effect Transistors," *IEEE Electron Device Letters*, vol. 37(6), pp. 789–792, 2016.

[9] M. Jin, H. Jeong, W. J. Yu, D. J. Bae, B. R. Kang, Y. H. Lee, "Graphene oxide thin film field effect transistors without reduction," *Journal of Physics D: Applied Physics*, vol. 42, pp. 135109, June 2009.

[10] W. S. Hummers, R. E. Offeman, "Preparation of Graphitic Oxide," *Journal of the American Chemical Society*, vol. 80(6), pp. 1339–1339, 1958.

[11] R. Negishi, Y. Matsui, Y. Kobayashi, "Improving sensor response using reduced graphene oxide film transistor biosensor by controlling the adsorption of pyrene as an anchor molecule," *Japanese Journal of Applied Physics*, vol. 56(651), pp. 06GE04, April 2017.

[12] S. J. R. Neale, E. P. Randviir, A. S. A. Dena, C. E. Banks, "An overview of recent applications of reduced graphene oxide as a

- basis of electroanalytical sensing platforms," *Applied Materials Today*, vol. 10, pp. 218–226, 2018.
- [13] S. Masoumi, H. Hajghasem, A. Erfanian, A. Molaei, M. Rajipour, "Design and manufacture of field-effect transistor based on reduced graphene oxide (RGO-FETs)," *Iranian Conference on Electrical Engineering (ICEE)*, pp. 445–450, 2017.
- [14] T. K. Truong, T. N. T. Nguyen, T. Q. Trung, I. Y. Sohn, D. J. Kim, J. H. Jung, N. E. Lee, "Reduced graphene oxide field-effect transistor with indium tin oxide extended gate for proton sensing," *Current Applied Physics*, vol. 14(5), pp. 738–743, 2014.
- [15] E. Carrion, A. Malik, A. Behnam, S. Islam, F. Xiong, E. Pop, "Pulsed nanosecond characterization of graphene transistors," *70th Device Research Conference*, pp. 183–184, June 2012.
- [16] H. Wang, Y. Wu, C. Cong, J. Shang, T. Yu, "Hysteresis of Electronic Transport in Graphene Transistors," *ACS Nano*, vol. 4(12), pp. 7221–7228, 2010.
- [17] J. T. Robinson, M. Zalalutdinov, J. W. Baldwin, E. S. Snow, S. Wei, P. Zhongqing, B. H. Houston, "Wafer-scale Reduced Graphene Oxide Films for Nanomechanical Devices," *Nano Letters*, vol. 8(10), pp. 3441–3445, 2008.
- [18] R. Fletcher, *Practical methods of optimization*, 2nd ed., New York: John Wiley and Sons, 1987.
- [19] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, et al., "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials," *Journal of Physics: Condensed Matter*, vol. 21(39), pp. 395502, September 2009.
- [20] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, et al., "Advanced capabilities for materials modelling with Quantum ESPRESSO," *Journal of Physics: Condensed Matter*, vol. 29(46), pp. 465901, October 2017.
- [21] H. J. Monkhorst, J. D. Pack, "Special points for Brillouin-zone integrations," *Phys. Rev. B*, vol. 13(12), pp. 5188–5192, 1976.
- [22] J. Pinos, S. Mikmekova, L. Frank, "About the information depth of backscattered electron imaging," *Journal of Microscopy*, vol. 266(3), pp. 335–342, 2017.
- [23] H. Wang, Y. Wu, C. Cong, J. Shang, T. Yu, "Hysteresis of Electronic Transport in Graphene Transistors," *ACS Nano*, vol. 4(12), pp. 7221–7228, 2010.
- [24] S. Ragul, S. Dutta and D. Ray, "Defect Mediated Small Molecular Doping of Graphene," *Advanced Optical Materials*, vol. 9(14), pp. 2002046, 2021.

Broadband Printed Dipole Antennas

¹R.P.Ghosh,
¹Dept. of Electronics,
 Vidyasagar University,
 Midnapore-721102
 India
 rajendra.ghosh@gmail.com

²B.Gupta
²Dept. of Electronics and Telecommunication Engineering,
 Jadavpur University,
 Kol-32.
 India
 gupta_bh@gmail.com

Abstract- Two arms of Microstrip Dipole antenna are printed on the same side of a substrate. But two arms of the Printed Dipole Antenna (PDA) are printed on the opposite sides and hence contain no ground plane. The Return Loss (RL) Bandwidth (BW) of PDA is much higher compared to Microstrip Dipole as it does not contain ground plane like the Microstrip Dipole and therefore Q value is low which results in large RL bandwidth. The PDA is fed by printed version of two wire transmission line. The cross polar discrimination is high as width to length ratio of the arm is very small. In general the width is kept below $0.05\lambda_0$. A novel design of PDA is reported in which the antenna arms have been widened to make it a Flag shaped to achieve large RL bandwidth. The -10dB RL bandwidth as high as 76% is achieved. Due to widening of the arms the fundamental resonant frequency and its first harmonic is stagger tuned to result a high RL. Three antennas are designed and simulated in MOM based simulator IE3D. One of them has been fabricated for experimental verification. The experimental result is in good agreement with the simulated results.

Keywords- printed dipole antenna (PDA), return loss bandwidth, broadband, flag shaped antenna.

I. INTRODUCTION

Two arms of PDA are printed on the two sides of a dielectric substrate and also known as Double Sided Printed Dipole Antenna (DSPDA) [1]. The DSPDA is fed by the printed version of two wire transmission line and is extensively analysed by the Wheeler [2].The structure of the DSPDA along with its feed structure is shown in Fig.-1[1].The closed form expression for the resonant frequency of the antenna is investigated by R.P. Ghosh, et. al. and is given in Equation-1[3].

$$f_r = \frac{c}{4\sqrt{\epsilon_r} L_0} * \frac{1}{(0.6463 e^{-0.4792 \epsilon_r} + 0.5453 e^{-0.01283 \epsilon_r})} \quad (1)$$

Where L_0 is the length of the antenna and ϵ_r is the relative dielectric constant of the medium.

The arm width of the antenna is typically kept less than $0.05\lambda_0$ [1]. This results in low transverse current component and hence cross polar discrimination is too high. The antenna has no ground plane and hence

no field confinement takes place, so the Q value is low compared to other form of Microstrip Antennas. The patch antenna in its basic and simple form provides 2-5% of -10dB RL bandwidth [1], whereas the RL bandwidth of DSPDA in its basic structure is nearly 15-18% [3].This has created immense interest among the researchers to use the PDA for wideband applications like medical imaging, high speed data applications, real time navigation etc. The research has been carried out to increase the RL bandwidth further using various techniques like integrated BALUN, flaring arms, parasitic elements, introducing slots on arms for gradual impedance transformation [4-10]. But these techniques of boarding the RL bandwidth have increased the complexity of the structure of the DSPDA.

In this communication the broad band DSPDA is reported with large impedance bandwidth (RL). The broadband is achieved by widening the arm's width making it a Flag shaped. The widening of arms stagger tunes the fundamental frequency and the first harmonic to provide large RL bandwidth. No BALUN, parasitic element or any other impedance matching technique is used to design the DSPDA. It is also found that cross polar discrimination is relatively high (more than 20dB) in spite of widening the arm width. Thus antennas are simply structured and fabrication of these is easy. Three antennas are designed and reported with highest 10dB RL bandwidth of 76%. All antennas are designed and simulated in MoM based simulator IE3D [11]. A DSPDA is experimentally verified and is in good agreement with the simulated data.

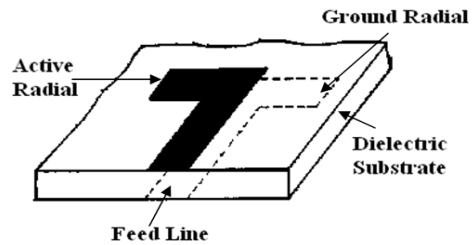


Fig.-1: Double Sided Printed Dipole Antenna (DSPDA) along with its feed structure [1].

II. PARAMETRIC STUDY TO DESIGN BROADBAND DSPDA

The DSPDAs consist of two arms known as active radial and ground radial which are printed on opposite faces of dielectric material. The length of the DSPDA for the resonant frequency is determined by the Equation-1.

$$f_r = \frac{c}{4\sqrt{\epsilon_r} L_0} * \frac{1}{(0.6463 e^{-0.4792 \epsilon_r} + 0.5453 e^{-0.01283 \epsilon_r})} \quad (1)$$

Where L_0 is the length of the antenna and ϵ_r is the relative dielectric constant of the medium. In case of DSPDA no field confinement takes place and hence there exists no fringing field. So the length correction is not required. The arm width of the DSPDA is usually kept less than $0.05\lambda_0$. Initially antenna is designed with the arm width much less than the prescribed value. But it is gradually increased and is found to increase the RL bandwidth. A parametric study is carried out to find the increase of bandwidth with increasing arm width. The common area as described by R.P.Ghosh et. al. [3] is also adjusted to get optimum RL bandwidth. A parametric study is also carried out to get the optimum RL bandwidth with the variation of common area.

An antenna is designed initially with the following specifications to carry out the parametric study. Dielectric Constant (ϵ_r) = 2.4, Length of each arm (L_0) = 130mm, Height of the substrate (h) = 1.524mm, Arm width = 6mm.

According to Equation-1, the DSPDA resonates at 0.5GHz. The arm width is kept well below the $0.05\lambda_0$. The variation of RL band width with the variation of common area is shown in Fig.-2. And the variation of bandwidth with the widening arm width is shown in Fig.-3.

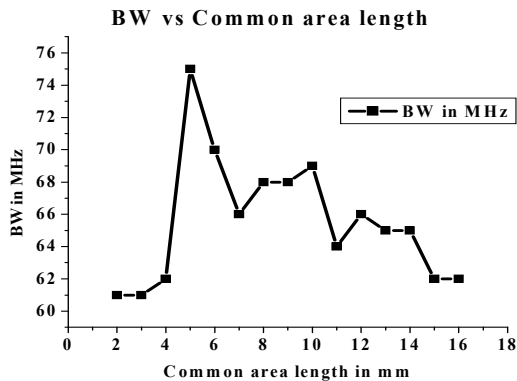


Fig.-2: Variation of RL bandwidth with common area of the DSPDA.

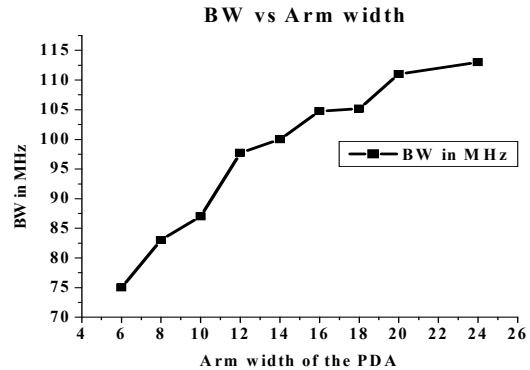


Fig.-3: The variation of BW with arm width.

So from the above parametric study it is concluded that the common area and the arm width may be simultaneously adjusted to get broad RL bandwidth of DSPDA. It is also found that the cross polar discrimination is well above the acceptable limit even after the widening the arm's width.

III. DESIGN OF BROADBAND DSPDA

Following the observations of the above parametric studies, three broad band DSPDAs has been designed. The DSPDAs resonate at 0.5GHz, 1.0GHz and 1.5GHz. All three antennas are designed and simulated in MOM based simulator IE3D. In all three DSPDAs it is found that the maximum RL bandwidth is obtained at the arm width of $0.1\lambda_0$ and with no common area. In each case, the fundamental frequency and its first harmonic is staggered tuned to produce large RL bandwidth. The cross polar discrimination is above 20 dB which is well above the acceptable limit. The DSPDA whose resonant frequency is 1.0GHz is fabricated for experimental verification. The experimental values agree with the simulated results.

IV. RESULTS

A. DSPDA-1

The DSPDA-1 operates at the frequency 0.5GHz. The antenna arms are flattened to $0.1\lambda_0$ to stagger tune the fundamental mode and first harmonic. It provides RL bandwidth of 62% with cross polar discrimination more than 20dB. The top and back view of the antenna are shown in Fig.-4(a,b). The antenna dimensions are given by Dielectric Constant (ϵ_r) = 2.4, Length of each arm (L_0) = 130mm,

Height of the substrate (h) = 1.524mm,
 Arm width = 60mm.
 Common area between arms = 0 mm.

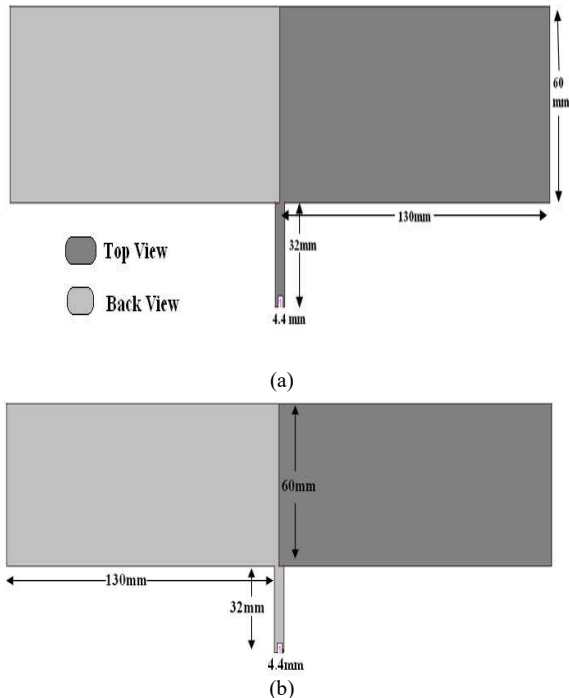


Fig.-4: Top and Back views of DSPDA-1
 (a) Top View (b) Back View.

The simulated RL plot, radiation patterns for E and H plane at three different frequencies and gain vs frequency plots are shown in Fig. - 5, 6, 7, 8 and 9.

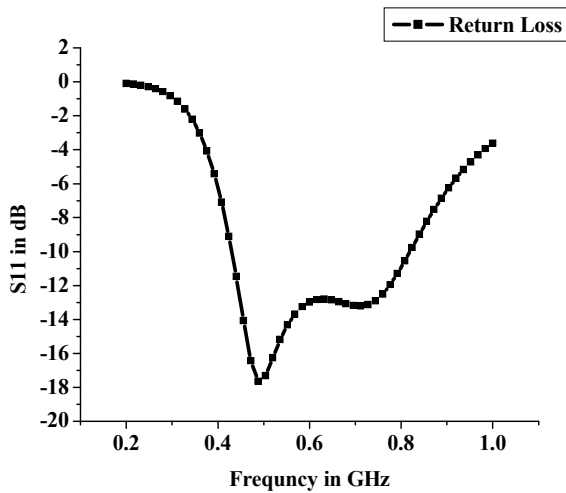
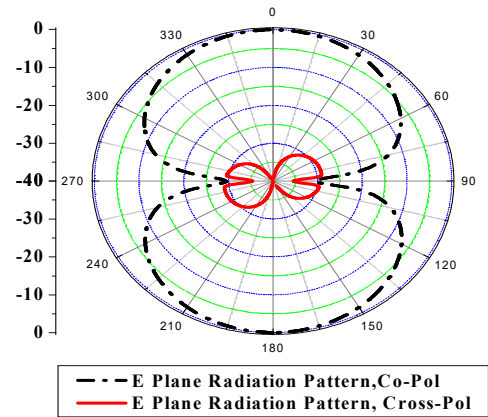
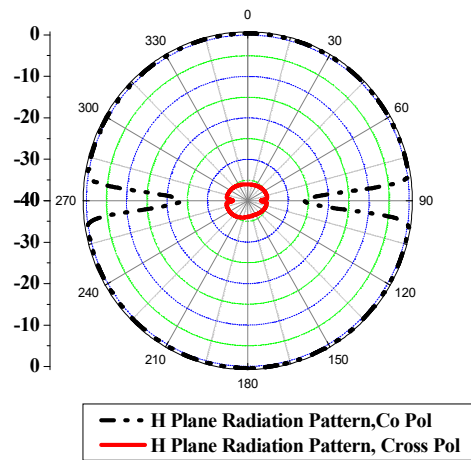


Fig.-5: Return Loss plot of DSPDA-1.

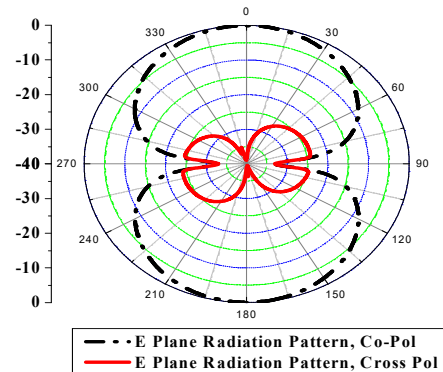


(a)

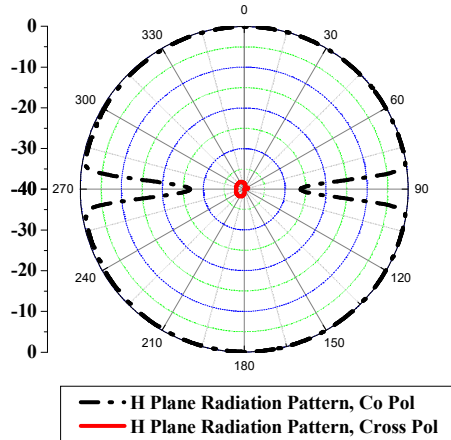


(b)

Fig.-6: Radiation Patterns at 0.43GHz of DSPDA-1.
 (a) E-Plane (b) H-Plane

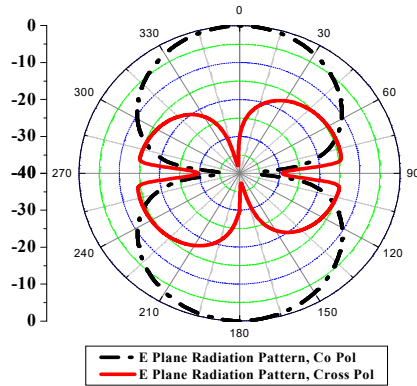


(a)

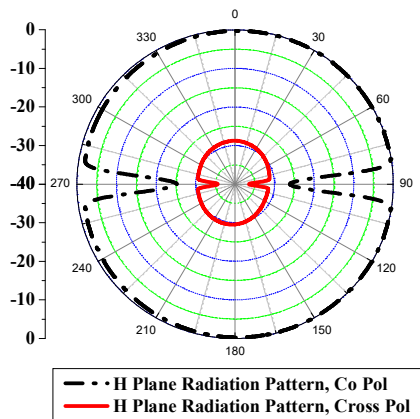


(b)
Fig. -7: Radiation Patterns at 0.63GHz of DSPDA-1.

(a) E-Plane (b) H-Plane



(a)



(b)

Fig.- 8: Radiation Patterns at 0.81GHz of DSPDA-1.

(a) E-Plane (b) H-Plane

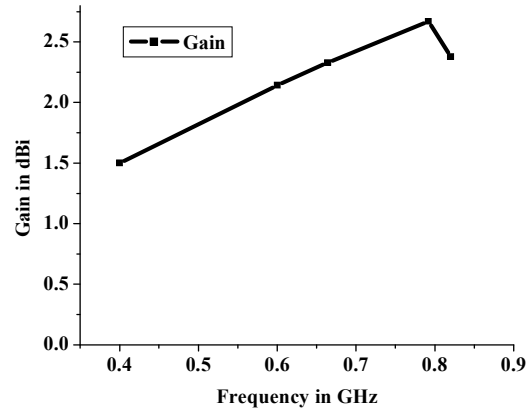


Fig.-9: Gain versus Frequency plot of DSPDA-1.

B. DSPDA-2:

The DSPDA-2 operates at the frequency 1.0 GHz and arms are flattened to $0.1\lambda_0$. It provides RL bandwidth of 69% with cross polar discrimination more than 20dB. The top and back view of the antenna are shown in Fig.-10 (a,b). The antenna is fabricated and experimentally verified. The antenna dimensions are given by

- Dielectric Constant (ϵ_r) = 2.4,
- Length of each arm (L_0) = 66 mm,
- Height of the substrate (h) = 1.524mm,
- Arm width = 30mm.

Common area between arms = 0 mm.

The top and back views of the designed antenna along with all dimensions are shown in Fig.- 10. The photograph of the fabricated antenna is shown in Fig.-11.

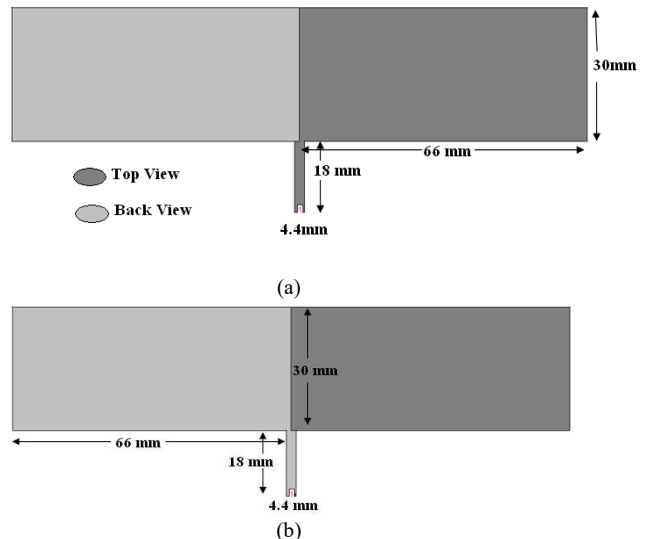


Fig.-10: Front and Back views of DSPDA-2

(a) Top View (b) Back View.

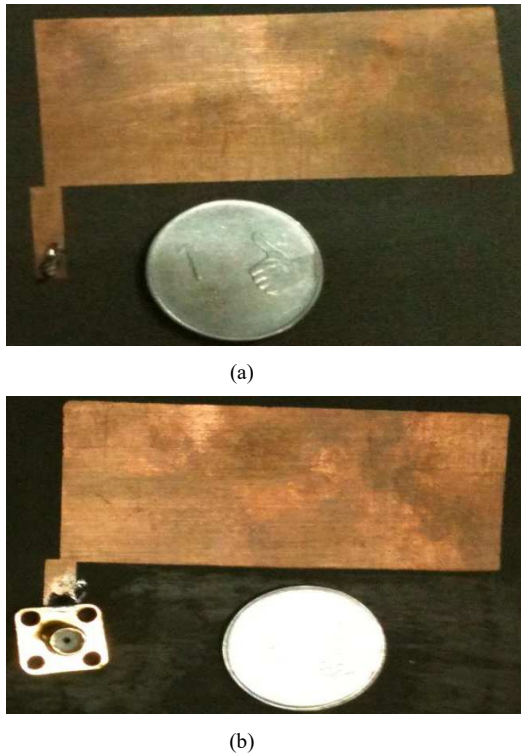


Fig.-11: Photographs of fabricated DSPDA-2
(a) Top View (b) Back View

The simulated and measured RL plot, radiation patterns for E and H plane at three different frequencies and gain vs. frequency plots are shown in Fig. - 12, 13, 14, 15 & 16.

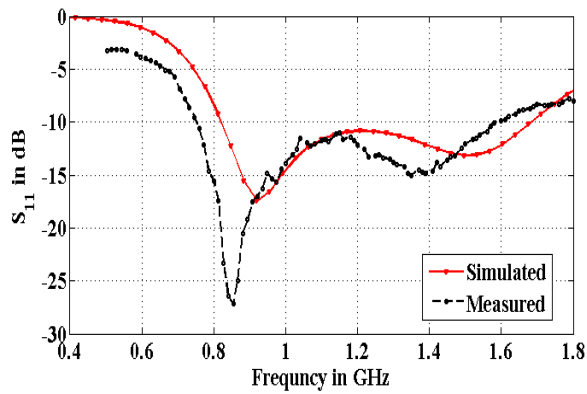
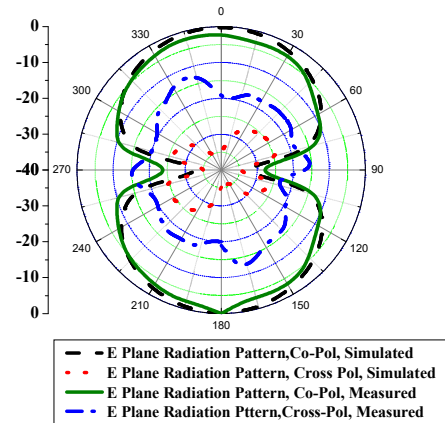
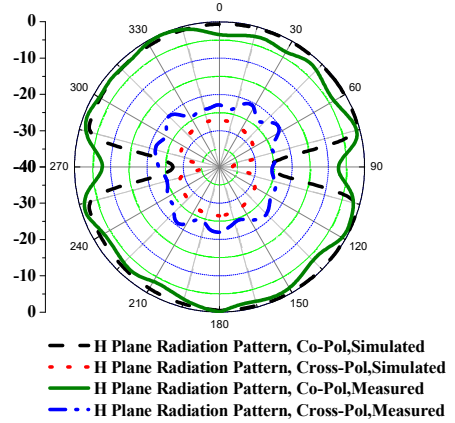


Fig.-12: Simulated and Measured Return Loss Plots of DSPDA-2.

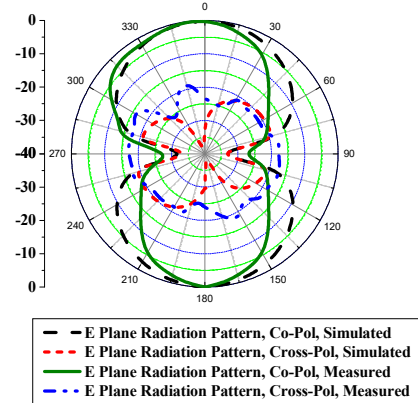


(a)



(b)

Fig.-13: Radiation Patterns at 0.80GHz of DSPDA-2
(a) E-Plane (b) H-Plane



(a)

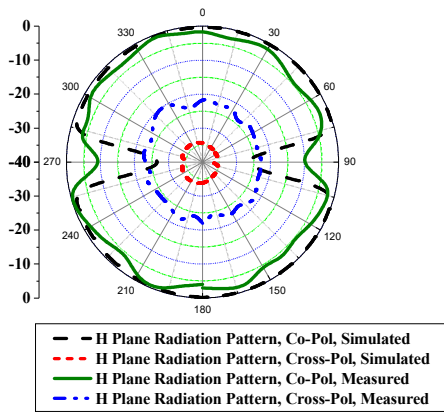
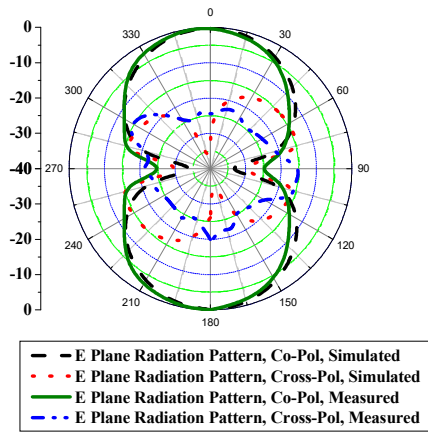
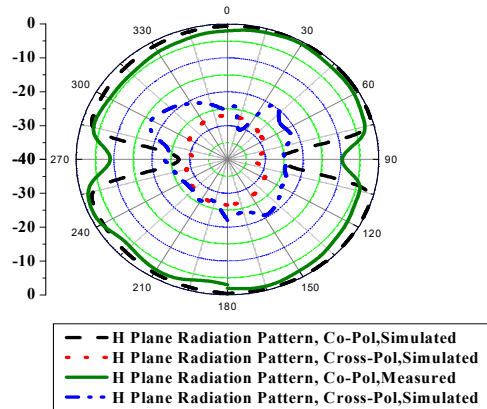


Fig.-14: Radiation Patterns at 1.20 GHz of DSPDA-2

(a) E-Plane (b) H-Plane



(a)



(b)

Fig.-15: Radiation Patterns at 1.5 GHz of DSPDA-2

(a) E-Plane (b) H-Plane

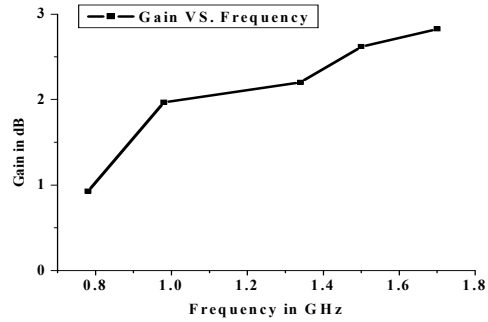


Fig.-16: Gain vs frequency of DSPDA-2.

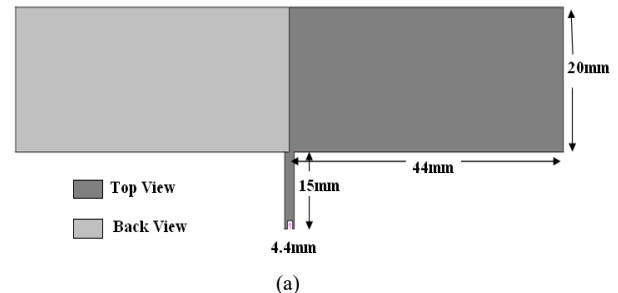
C. DSPDA-3

DSPDA-3 is also designed to operate at 1.5GHz with the following dimensions. The antenna provides impedance bandwidth of 76%.

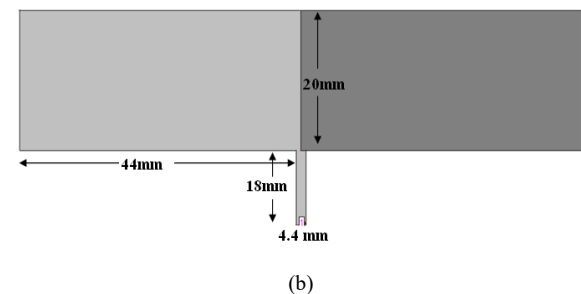
Dielectric Constant (ϵ_r) = 2.4,
 Length of each arm (L_0) = 44 mm,
 Height of the substrate (h) = 1.524mm,
 Arm width = 20mm.

Common area between arms = 0 mm.

The top and back views of the designed antenna along with all dimensions are shown in Fig. - 17.



(a)



(b)

Fig.-17: Top and Back views of DSPDA-3

(a) Top View (b) Back View.

The simulated RL plot, radiation patterns for E and H plane at three different frequencies and gain vs. frequency plots are shown in Fig. - 18, 19, 20, 21 and 22.

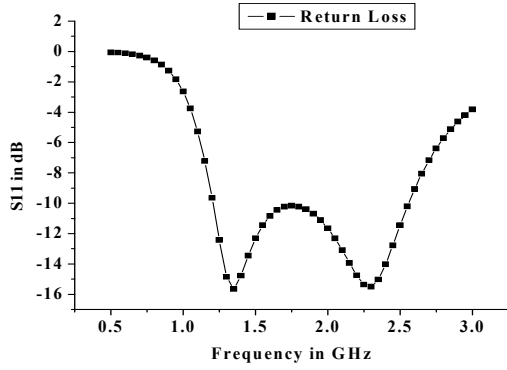
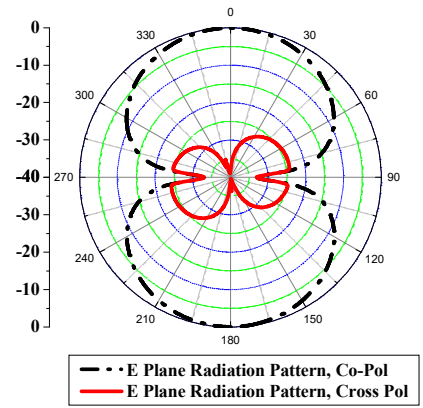
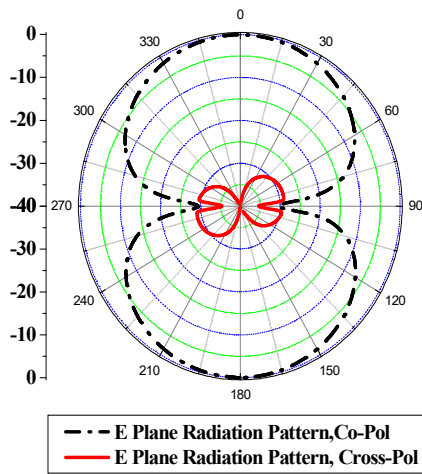


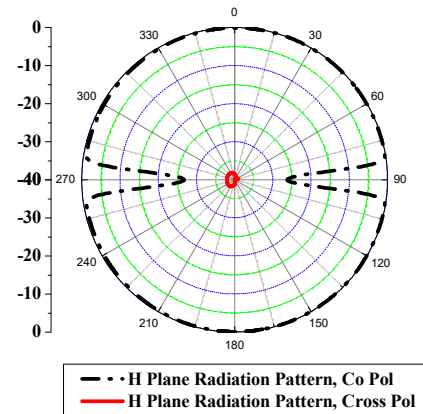
Fig.-18: Return Loss Plot of DSPDA-3.



(a)



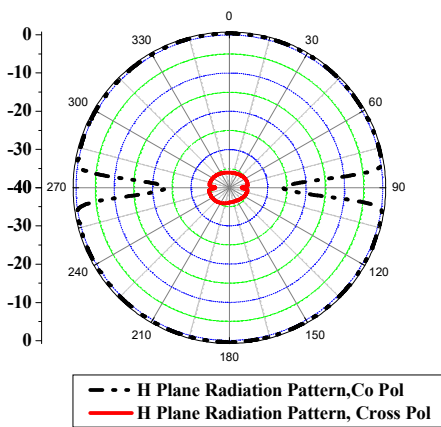
(a)



(b)

Fig.-20: Radiation Patterns at 1.85 GHz of DSPDA-3.

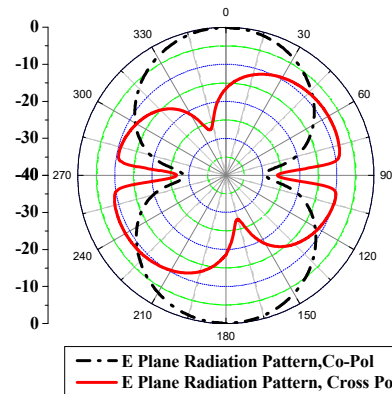
(a) E-Plane (b) H-Plane



(b)

Fig.-19: Radiation Patterns at 1.2GHz of DSPDA-3.

(a) E-Plane (b) H-Plane



(a)

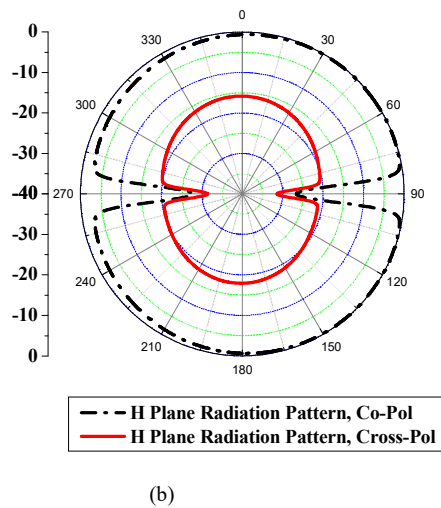


Fig.-21: Radiation Pattern at 2.54 GHz of DSPDA-3.

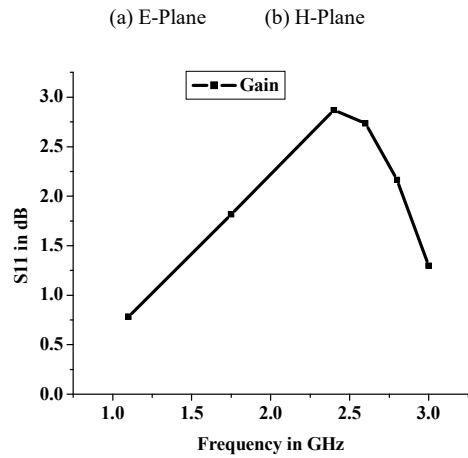


Fig.-22: Gain versus Frequency plot for DSPDA-3.

V. CONCLUSIONS

Three broadband DSPDAs are developed with the fundamental frequency of 0.5GHz, 1.0GHz and 1.5GHz. In each case, the fundamental and first harmonic is staggered tuned when the antenna arms are widened to $0.1\lambda_0$. It makes the antenna Flag shaped. The Q value of the antenna is lowered at both the frequency due to the increase of arm’s width and results in high RL bandwidth. The DSPDAs provide RL bandwidth of 62%, 69% and 76%. The antennas are simply designed. No complicated antenna design techniques are used to achieve large RL bandwidth. Even no BALUN or complicated impedance matching techniques are used. The cross polar discrimination is well above 20dB in each case even after widening the arm width. A generalised design procedure may be adopted to design the broad band

DSPDA. At first, calculate the arm length for any resonating frequency (fundamental frequency) using the Equation-1 and then widen the antenna arm to $0.1\lambda_0$. In the process, the fundamental frequency and its first harmonic will be staggered tuned to provide large RL bandwidth. All the three antennas have been designed following this method.

REFERENCES

1. I.J.Bahl, P.Bhartia, R.Garg and A. Ittipiboon, “ Microstrip Antenna Design Handbook,” 2nd Edition, Artech House, Dedham, MA, pp. 399,2001.
2. Wheeler H.A., “Transmission Line Properties of Parallel Strips Separated by a Dielectric Sheet”, IEEE Trans. on Microwave Theory and Technique.Vol. MTT 13, pp. 172-185, Nov. 1965.
3. R.P.Ghosh, B.Gupta, Kaushik Patra, S.K.Chowdhury, “Accurate formula to determine resonant frequency of double sided printed dipole antenna,” IETE Journal of Research, 2017, <https://doi.org/10.1080/03772063.2017.1355749>.
4. R. P. Ghosh, B. Gupta and S. K. Chowdhury, "Broadband printed dipole antennas with shaped ground plane," TENCON 2010 - 2010 IEEE Region 10 Conference, 2010, pp. 416-421, doi: 10.1109/TENCON.2010.5686680.
5. S. Thirakoune, A. Petosa, A. Ittipiboon and K. Levis, "Broadband printed dipole antennas," IEEE Antennas and Propagation Society International Symposium (IEEE Cat. No.02CH37313), 2002, pp. 52-, doi: 10.1109/APS.2002.1018154.
6. Thirakoune s., Petosa a.,Ittipiboon a., Levis K., "Broadband Printed Dipole Antennas," IEEE Antennas and Propagation Society International Symposium, vol.3, pp.52, 16-21 June 2002.
7. K. E. Kedze, H. Wang, T. S. Xuat, and I. Park, “Wideband Low-Profile Printed Dipole Antenna Incorporated with Folded Strips and Corner-Cut Parasitic Patches above the Ground Plane,” IEEE Access, vol. 7, pp. 15537-15546, 2019.
8. Y.-D. Lin and S.-N. Tsai, "Analysis and Design of Broadside-Coupled Striplines-Fed Bow-Tie Antennas," IEEE Transactions on Antennas and Propagation, vol.46, no.3, pp.459- 460, Mar. 1998.
9. A. R. Behera and A.R. Harish, “A Novel Printed Wideband Dipole Antenna,” IEEE Transaction on Antennas and Propagation, vol. 60, No. 9, Sep. 2012.
10. Vasiliadis, T.G. and Sergiadis, G.D. (2006), “Wideband printed dipole antenna parasitically enhanced with over-an-octave bandwidth. Dual-band variant for WLAN applications”. Microw. Opt. Technol. Lett., 48: 444-449, 2006. <https://doi.org/10.1002/mop.21375>.
11. <https://www.rfglobalnet.com>.

SQL ChatBot – using Context Free Grammar

Rajvardhan Patil
School of Computing
Grand Valley State University,
Allendale, USA
patilr@gvsu.edu

Sorio Boit
School of Computing
Grand Valley State University,
Allendale, USA
boitj@gvsu.edu

Nathaniel Bowman
School of Computing
Grand Valley State University,
Allendale, USA
bowmnath@gvsu.edu

Abstract—In this work, we derive the semantics of a given English query and convert it into its equivalent SQL query. Instead of using neural networks for semantic analysis of English queries, we opt for a Context Free Grammar approach. Most neural-network-based systems can handle only one semantic at a time, whereas, because of the flexibility offered by our CFG approach, our system manages to handle simultaneous usage of conjunctive, disjunctive, and negative semantics. It also handles complex statements comprising of main as well as dependent clauses. In addition, the system also takes into account aggregate functions and constructs the required GROUP-BY and HAVING clauses. We describe how the system analyzes English queries by understanding the role that each part-of-speech has to play in constructing SQL queries. Numerous examples demonstrate the effectiveness of our approach where state-of-the-art techniques relying on deep learning algorithms fail to deliver.

Keywords—*chatbots, natural language processing, semantic analysis*

I. INTRODUCTION

Traditionally, relational databases have relied on a standard query language (SQL), a powerful tool for extracting data. But, often times this can be very challenging, specifically for non-technical users. Even for power users, expertise is still required to understand the specific schemas, and entity relationships further add to the complexity of using SQL. However, in the real-world, many people use natural languages such as English to ask questions and find solutions to their problems. Programs, for example, Chatbots, have been developed to answer queries or deliver personalized responses to various requests. Chatbots, a word derived from “chat robots”, have dramatically revolutionized how people interact with online information and services, a trend that requires a rethinking of user needs during the development of chatbots [1]. Chatbots are also known by other names, for example: virtual assistants, conversational agents, dialogue systems, personal assistants and conversational interfaces [2]. In recent times, Chatbot adoption in many applications is on the rise due to automation of labor-intensive tasks at low costs, improved customer experiences, and other benefits [3].

According to recent research [4], the global Chatbot market is predicted to grow at the rate of 31 percent by 2024 across different sectors such as healthcare, retail, travel and BFSI (Banking, financial services, and insurance services). Additionally, the retail sector is projected to spend \$142 billion worldwide by leveraging chatbots [5]. These trends are not driven only by technological advancements but largely by the desire to improve customer service and delivery experiences on ubiquitous devices, social media, customer relationship management and more [6]. With the increase in demand by

consumers, the Chatbot ecosystem is rapidly expanding, providing more opportunities to develop technological-focused platforms.

The primary goal of this paper is to develop a Chatbot system that provides non-technical users with an interface to interact with RDBMS using English-like, human-readable statements. In this paper, we propose an SQL-Chatbot System (SCB) consisting of mainly three steps. In the first step, the given English query is decomposed into granular components (individual keywords). In the second step, the equivalent SQL query is built. In the final step, the query is executed to return the results back to the user. To the best of our knowledge, our system is the first to use Context Free Grammar (CFG) for an SQL-based Chatbot platform. Our system leverages Context Free Grammar (CFG) to generate semantically equivalent SQL queries while contextualizing the role of parts-of-speech (POS).

II. RELATED WORK

One early attempt to connect casual users to relational databases using natural languages was the RENDEZVOUS project [7]. With the increase of both structured and unstructured data, a resurgence in this field has attracted research interest in understanding how to convert natural-language statements to machine readable form, such as semantic parsing. One approach is to use table-based semantic parsing on large databases. For example, Seq2sql used reinforcement learning [8], SPIDER leveraged multiple tables and contains multiple SQL queries [9]. Other works have investigated more challenging database structures such as relational databases. These approaches include RAT-SQL[10], which translates relational structure in the database schema, evaluation on multiple datasets using neural text-to-SQL systems [11], ToTTo for summarization [12], Table2Vec for retrieval tasks [13], and HybridQA for question-answering [14].

With the recent rise of the importance of neural networks in natural-language processing, some natural-language database interfaces have incorporated these networks into their systems in some way. For example, the authors of [15] created a speech-based interface based on semantic matching that used a convolutional neural network as part of the design. As another example, in [16], the authors used a word embedding based on a neural network in order to enable semantic queries to a database. However, we decided not to incorporate neural networks in this work because when it comes to semantic analysis, we found that our CFG approach offers more flexibility as compared to deep-learning approaches.

A recent survey of NLIDB with Deep learning evaluated 349 articles with a focus on text-to-SQL tasks across well-known

datasets such as Spider, ATIS, GeoQuery, Sparc and WIKISQL [17]. For example, the use of “Group By” and “Order by” SQL clauses remains a challenging task for deep learning applications due to limited availability of large datasets, varying schema information and cross-domain adoptability [17]. Another limitation for NLP-based SQL methods is the representation of queries into intermediate language [17]. In contrast, our CFG-based algorithm handles the challenging tasks of constructing clauses such as “GROUP BY” and “HAVING” and generating semantic queries while contextualizing the role of parts-of-speech in the use of SQL ChatBots.

Previous research in neural networks has attempted to map NLP to SQL queries to solve problems based on popular crowd-sourced datasets such as WikiSQL and Seq2SQL [8]. Another approach used autoregressive generative models trained on WikiSQL, ATIS, and GeoQuery datasets[18]. This method used execution guided-decoding of SQL leveraged on the semantics of SQL. Our work leverages Context Free Grammars (CFG) by applying a Framing Procedure (FP), a Grouping Algorithm (GA), and a Clause-Construction Algorithm to construct associative SQL clauses. The objective of this work, therefore, is to create a novel SCB system that allows even naïve users to interact with chatbots by specifying ad-hoc queries. We introduce a SQL-Chabot system built on CFG for semantic parsing to automatically map English queries to their corresponding SQL queries. Further details about the SCB system are discussed in the methodology section.

III. SYSTEM OVERVIEW

SCB system has mainly three built-in components used for query’s analysis purpose: thesaurus, lists of delimiters and connectors. Delimiters and connectors are used for English query’s decomposition purpose (to be discussed later); whereas inbuilt thesaurus is used for query expansion purpose, where if a user enters any word synonymous to any attribute or value, then with the help of thesaurus the system is able to map it back to the original keyword. In addition, after a particular database is connected to our system, the system will automatically generate an inverted index, based on the underlying schema and data graph.

Thereafter, when an English query (text or voice format) is submitted to the system, the parsing process takes place, and the English sentence is decomposed to construct an SQL query. The system first builds a sub-graph (candidate network) that represents a path between relations having all the query keywords. To locate such involved relations, our search engine uses depth first search technique for traversing the schema graph, which further results into the construction of FROM clause of SQL query. The system also uses inverted index for lookup purpose; i.e. for a given keyword it can retrieve the attributes and tables to which the keyword belongs to, in the underlying database. These table and attribute pairs (associated with the query-keywords) together forms the SELECT clause. Construction of WHERE clause involves a couple of components as shown in Fig.1, which depicts key steps

involved in execution or working of our search system. A brief overview of main components is provided below. They will be further explained in detail in methodology section.

The system processes English queries iteratively until all the sub-queries within a given query are not processed/parsed. Working of this system is mainly based on a Context Free Grammar (CFG).

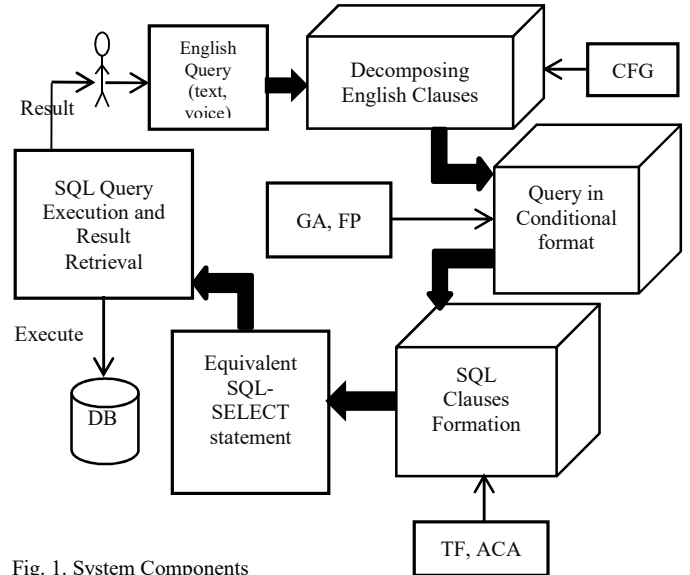


Fig. 1. System Components

Based on the list of delimiters, the query is first divided into sub-queries (main clause and subordinate clauses). Based on the list of connectors, each clause is then divided into subject and predicate, and is further analyzed from left-to-right to identify the attributes and values which in turn form the conditions. Also, every condition is associated with its neighboring condition by a logical operator. By applying the grouping algorithm (GA), conditions within each sub-query are grouped in such a way that the precedence and associativity of the involved logical operators is preserved. The precedence of such logical operator is determined by the number of parenthesis surrounding the condition of which it is being part of. GA works iteratively to produce parenthesized form for each clause. All the subordinate clauses, based on the Framing procedure (FP), are then eventually associated with the main clause so as to get the representation of the entire query in parenthesized or conditioned (attribute-value pair) format. This conditional format of the given English query is further converted to tabular format (TF). Tabular format makes it easier for search engine to deduce SQL’s SELECT, FROM and WHERE clauses.

Additionally, Clause-Construction Algorithm is used to detect whether or not the given English query has any aggregate functions involved or not; if in case it does, then it helps in identifying the attributes associated with aggregate functions, and the ones around which the grouping has to be done. The algorithm further constructs a data structure called Attribute-Clause-Association (ACA), which is used for the construction

of SQL’s GROUP-BY and HAVING clauses. All the SQL-clauses are then concatenated to get the entire SQL statement, which is further executed on the underlying database so as to fetch and display the results to the user. Also, to understand the semantics of the query, we have considered the roles of parts-of-speech that they play in breaking down the given English query and then in construction of its equivalent SQL query.

The strategy implied by the search system, therefore consists mainly of two phases: decomposition and construction phases. In the decomposition-phase, the English query is decomposed by using CFG to the smallest possible granularities. In the construction-phase, GA along with FP are used to parenthesize the parsed query, from the above phase, in a way so as get semantically equivalent parenthesized format of the original English query. Also, whenever needed, ACA algorithm is applied to take care of aggregate-functions, and then ensures the construction of GROUP-BY and HAVING clause. Throughout these phases, the given query is skimmed over in order to retain just the needed information, as per the CFG’s requirement; rest is discarded. Eventually, with the support of data-structures and algorithms, the parenthesized (conditional) format is converted into SQL’s SELECT statement for execution purpose.

IV. METHODOLOGY

A. Sentence Components

Fig. 2 below, illustrates the roles that various parts-of-speech have to offer in constructing SQL statement, for a given English query. It also helps in understanding how the components of English sentences comprise the WHERE clause (conditions) of SQL statement. In this diagram the dotted line represents that within a clause the conditions, if not stated explicitly, are implicitly connected through logical operator (AND).

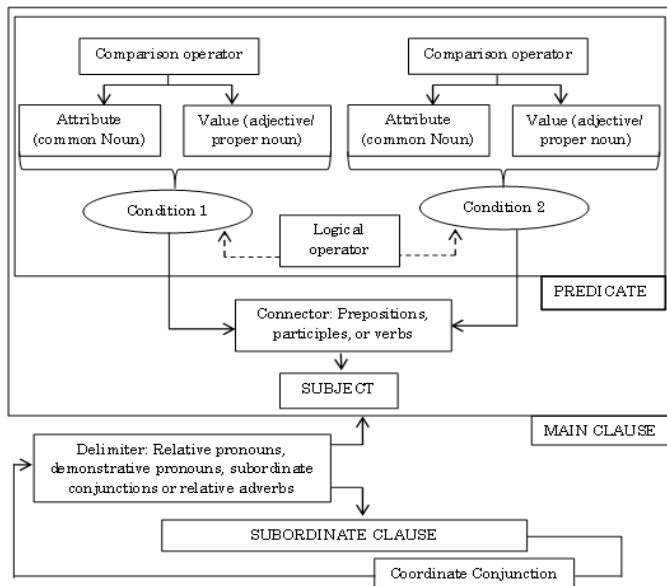


Fig. 2. Sentence Components

In addition, the conditions are individually associated with the subject in a clause through *connectors*, indicated by the solid line. Similar to the conditions, subordinate clauses introduced by delimiters, are also connected with each other by the help of coordinate conjunctions, and then are individually associated with main clauses.

B. Context Free Grammar (CFG)

Decomposition of the English query by Search engine is carried out through the Context Free Grammar (CFG) shown in Table-1.

Table 1. Context Free Grammar (CFG)

Grammar $G = (N, \Sigma, P, Q)$, where-
The non-terminal symbols are: $N = \{\text{Query, Main-Clause, Subordinate-Clause, Subject, Predicate, Condition}\};$
The terminal symbols are: $\Sigma = \{\text{attribute, value, delimiter, connector, logical-operator, coordinate-conjunction, comparison-operator, wrapper}\};$
‘Query’ = is the starting symbol, and
Query \rightarrow Main-Clause Main-Clause Subordinate-Clauses
Subordinate-Clauses \rightarrow Subordinate-Clause (coordinate-conjunction Subordinate-Clause)*
Main-Clause \rightarrow Subject connector Predicate
Subordinate-Clause \rightarrow delimiter connector Predicate
Predicate \rightarrow (Condition) (logical-operator Condition connector Condition logical-operator connector Condition)*
Subject \rightarrow (Condition) (logical-operator Condition)*
Condition \rightarrow (attribute comparison-operator value value comparison-operator attribute)*
Condition \rightarrow (attribute)* (value wrapper)* (value attribute)

From the CFG we can see that the delimiters are used to introduce subordinate clause into the sentence. Connectors are used to connect subject and predicate within a given clause (main or subordinate). Coordinating conjunctions like ‘or, and’ in the query, usually act as logical operators which connects the conditions within the subject or predicate of a clause, but additionally can also be used to connect subordinate clauses. Examples to be encountered in the paper, shall assert the fact of immense flexibility being offered by this CFG.

Wrappers are no different from logical operators, except that they operate on the values belonging to the same attribute domain. For example in the query, ‘car having red or green color’, both values (red, green) belong to the same attribute (color) and hence the operator ‘or’ is treated as a wrapper by the search system. The reason for such distinction is to give higher

preference for wrappers over other logical operators. The following example illustrates how the CFG decomposes the given query into smaller components:

Query-1: Look for a red color Toyota Camry that has price < \$8000 or which has been manufactured in year 2012 with mpg > 25. Here,

Main Clause (M): Look for a red color Toyota Camry
 Subordinate Clause (S1): that has price < \$8000 or
 Subordinate Clause (S2): which has been manufactured in year 2012 with mpg > 25.

C. Grouping Algorithm

After the query is decomposed by CFG, the next step is to group the conditions in a way so as to retain the precedence of involved operators. This job of retaining the semantics of query is done by GA. Parentheses, alongside the operands, are used to decide the preference of operators. The output of above CFG’s phase is given as an input to GA. GA processes each clause in an individual manner. Grouping Algorithm plays a crucial part as it helps in construction of WHERE clause, where the precedence of operators needs to be preserved. The grouping algorithm is summarized in Fig. 3.

Algorithm: Grouping Algorithm (GA)
 Input: un-parenthesized sub-query and its Attribute Domain Information (ADI)
 Output: parenthesized sub-query
 Procedure:

- (a) Wrappers are given higher priority than Logical operators.
- (b) As per the precedence rule, the logical operators are prioritized in the following order: NOT, AND, OR
- (c) If the same operator is repeated, then the priority is determined through Associativity rule: the left hand side operator is dealt first.

Grouping Algorithm satisfies the above priority rules, by building the following different cases. In each case, grouping of the operands takes place based on the attribute domain information (ADI) and the operator at hand:

- (1) AND, OR wrapper: If they act as wrappers, then group the values on which they operate
- (2) AND operator: If it operates on values belonging to different attribute domain then the values can be grouped, else not
- (3) OR operator: If it operates on the values belonging to the same attribute domain then the values can be grouped, else not
- (4) NOT operator: Either acts as a delimiter or as an unary operator

CASE A: As a delimiter

NOT (A operator B) --[Where “A” and “B” are values]

CASE B: As an unary operator

(NOT A) operator (NOT B)

Fig. 3. Grouping Algorithm

After the output of CFG is fed as input to GA, it generates the following query conditions:

Main Clause: ((color=red) and (manufacturer=Toyota) and (model=Camry))

Subordinate Clause-1: (price<\$8000)
 Subordinate Clause-2: ((production_year=2012) and (mpg >25))

After having the composed version of the English query from GA, Framing Procedure (FP) then applies propositional logic (particularly associative, commutative, and distributive rules) so that each subordinate clause can be separately associated with the main clause, and yet retaining the semantics of the original English-query. Along with GA, FP basically helps in constructing the parenthesized-format (intermediate representation) of the given English query. For the example in consideration, the conditions are grouped in the following way to construct the parenthesized format:

((color=red) and (manufacturer=Toyota) and (model=Camry))
and (price<\$8000))
OR (((color=red) and (manufacturer=Toyota) and (model=Camry))
and ((production_year=2012) and (mpg >25)))

The output of the above step represents the WHERE clause. For construction of SELECT and FROM clause, our system refers to the inbuilt inverted index; i.e., for every value involved in the query, the system derives its attribute and table name to which it belongs to. Based on this information and from parenthesized format of the query, we deduce the tabular format as shown in Table 2. Note, R.O. represents relational-operators.

Table 2: Tabular Format Of Parenthesized Query

open	table	attribute	R. O.	value	close	logical operator
((Table 1	Color	=	Red)	And
(Table 2	Manufacturer	=	Toyota)	And
(Table 4	Model	=	Camry))	And
(Table 1	Price	<	8000))	Or
((Table 6	Color	=	Red)	And
(Table 6	Manufacturer	=	Toyota)	And
(Table 5	Model	=	Camry))	And
((Table 5	production_year	=	2012)	And
(Table 3	Mpg	>	25)))	

Here, the name of tables in column2 will form the FROM clause, and the name of attributes in column3 will form the SELECT clause. In addition, each row represents a condition; so appending all the rows from above table will result in construction of WHERE clause. SCB also offers more flexibility

to users by allowing them to incorporate the aggregate functions into their query. The formation of GROUP-BY clause and HAVING clause is elucidated below.

D. Clause Construction Algorithm

Clause Construction Algorithm is used to deal with aggregate functions in the user’s English query, and hence responsible for construction of GROUP-BY and HAVING clauses in the background.

We now present the clause construction algorithm in Fig. 7, which parses the given English sentence to check for the presence of any aggregate function; if yes, then based on the following rules it constructs the data structure ACA, which is responsible for the construction of GROUP-BY and HAVING clause.

- (1) If any attribute in the English statement is preceded by aggregate function, our search engine infers that such attribute will be part of SELECT clause.
 - (a) Furthermore, such attributes together with aggregate functions, if in case do have any conditions imposed on them, then such condition shall be part of HAVING clause, and not of WHERE clause instead.
- (2) If any attribute is complemented by any value so as to form an attribute-value pair, then such attribute-value pair shall be part of WHERE clause, and not of the HAVING clause.
- (3) If any other attributes are mentioned in the *subject of a query* which neither have any associated value nor is preceded by any aggregate function (type-4) then such attributes shall be part of GROUP-BY and SELECT clause, and not of the WHERE and HAVING clauses instead.
- (4) Type-4 attributes that represent *objects in the predicate of a query* don’t make it to the SELECT and GROUP-BY clauses, unless explicitly stated; by default they can be only be the part of WHERE clause if accompanied by a value, or of HAVING clause if accompanied by both value and an aggregate function.

Fig. 4. Clause Construction Algorithm

After executing this algorithm, our system generates the *Attribute-Clause Association (ACA)* data structure. For illustration purpose, consider Query 2.

Query 2: Display the list of students in computer science branch whose/having average score greater than 75.

Here, the attribute ‘score’ has not only an aggregate function preceding it but also a condition that follows it. Therefore, such condition ‘avg(score)>75’ belongs to HAVING clause and not WHERE clause. ‘Students’ form the subject of the sentence, and as there is no aggregate function associated, it is part of SELECT as well as GROUP-BY clause. The ACA for this query is shown in Table 3

Table 3: ACA For Query-2

Attribute	Aggregate Function	Select Clause	Group-by Clause	Where Clause	Having Clause
Students	N/A	Present	Present	Absent	Absent
Branch	N/A	Absent	Absent	Present	Absent
Score	Average	Present	Absent	Absent	Present

The SQL query generated for Query-2 is:
 Select students, avg(score) from Table-t1
 where branch= ‘computer science’
 group by(students)
 having avg(score)>75.

Query 3: Find customers and their total/sum expenditure at Walmart organization, in the month of January 2012.

Here, the attribute preceded by aggregate function is ‘expenditure’. As it has no conditions imposed on it, it shall be just part of SELECT clause and not of the HAVING clause. Further, the attributes organization, month and year have been specified with the values of ‘Walmart’, ‘January’ and ‘2012’ respectively. As a result, those attributes along with their values will be part of the WHERE clause condition. Now for the attribute ‘customers’, based on the rules 3&4, search engine knows that ‘customers’ is subject and should be a part of the GROUP-BY clause. ACA for Query-3 is shown in Table 4.

Table 4: ACA For Query-3

attribute	aggregate function	select clause	group by clause	where clause	having clause
customers	n/a	present	present	absent	absent
expenditure	sum	present	absent	absent	absent
organization	n/a	absent	absent	present	absent
month	n/a	absent	absent	present	absent
year	n/a	absent	absent	present	absent

The SQL query generated for Query-3 is:
 Select customers, sum(expenditure)
 from Table-t2
 where (organization= ‘Walmart’ and month = ‘January’
 and year = ‘2012’)
 group by(customers)

Furthermore, there can be some cases where user might explicitly specify the attributes around which the grouping should be done. For instance, adjective words such as: each,

every, make it possible for users to be absolute or certain about the attributes that demands inclusion in the GROUP-BY clause. For example, consider Query 4.

Query 4: Count number of employees for each department in the company XYZ.

Here, the subject of sentence is ‘employees’ and hence is a part of the SELECT clause; whereas, when it comes to ‘department’ attribute, it even though being the object of the sentence still makes it to the GROUP-BY clause, since here the user is explicitly stating its inclusion through the use of adjective ‘each’. The ACA for Query 4 is shown in Table 5.

Table 5: ACA For Query-4

Attribute	Aggregate function	Select clause	grouby clause	Where clause	Having clause
Employees	Count	present	Absent	absent	absent
department	N/A	present	Present	absent	absent
Company	N/A	absent	Absent	present	absent

The SQL query generated for Query-4 is:

```
Select department, count(employees)
from Table-t3 where (company= 'XYZ')
group by(department)
```

Note that when it comes to English queries constructed by user, they are either of interrogative or imperative nature. As a result, most of the queries shall begin with a ‘verb’. Our search engine ignores such initial verbs (and prepositions) until the subject has being encountered. Also, the aggregate functions in the English query will only be taken into account, if the user is specifying them explicitly, and not implicitly otherwise.

V. EXPERIMENTS & RESULTS

Sample queries executed by SCB system are shown from Fig.6 to Fig.9. The schema graph of the underlying database (on which the queries were executed) is shown in Fig. 5.

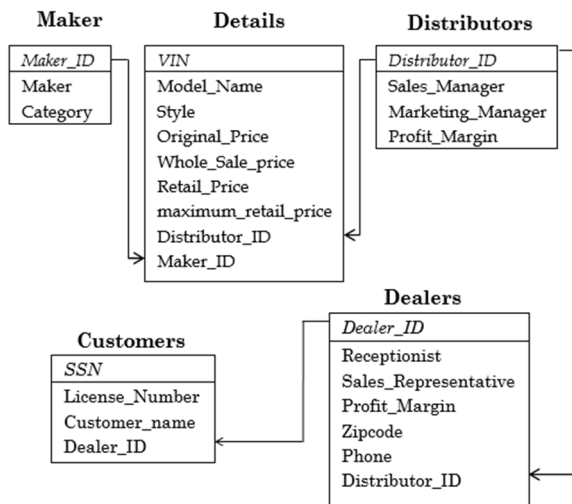


Fig. 5. Schema Diagram

Enter an English Statement Query:-
Count number of vehicles possessed by each dealer, located in Nebraska

----- OUTPUT -----

dealer_name	Count(vehicles)
Baxter	12350
Woodhouse	19180
Edmunds	21005

Fig. 6. Output for Query-5

Enter an English Statement Query:-
Find Honda model having the least/minimum price at Baxter

----- OUTPUT -----

Model	Min(Price)
Civic	\$10000

Fig. 7. Output for Query-6

Enter an English Statement Query:-
Find maximum mileage associated (5000 MILES ON THEM) with each car maker at Woodhouse

----- OUTPUT -----

Maker	Max(mileage)
Ford	23000
Honda	2800
Toyota	250000
Chrysler	20

Fig. 8. Output for Query-7

Enter an English Statement Query:-
Count number of coupe style cars sold in Omaha for year 2012

----- OUTPUT -----

Count(cars)
30560

Fig. 9. Output for Query-8

VI. CONCLUSION & FUTURE WORK

As discussed, the extended keyword search facility in our search engine initially deduces the structure of an English query with the help of CFG. Conditions within the original query are then grouped/parenthesized in a way so as to preserve the precedence of the involved operators, and hence the overall semantics of a query. The Framing Procedure goes a step ahead to deal with queries having more than one subordinate clause by informing the search engine of the way clauses of complex English queries can be associated. Roles played by parts of speech in composing English statements are understood and applied in the context or terminology of relational databases so as to construct their equivalent SQL query. The paper also illustrates the technique used to identify and further map the variety of attributes used in the English query statements to their SQL counterparts. The Clause-Construction algorithm is then

introduced to detect the presence of any aggregate functions in the query, and if needed, to construct the required GROUP-BY and HAVING clauses. By default, the search system allows the simultaneous use of disjunctive (OR), conjunctive (AND), negative (NOT) semantics, along with the possible combinations of NAND and NOR. Thus, the semantics derived by GA are used to construct the WHERE clause. The construction mechanism of SELECT and FROM clauses works in a way that is similar to traditional DBKWS. This paper also refines and exploits the list of delimiters and connectors used for decomposing the English queries. Overall, the paper demonstrates how the CFG approach enables extended keyword search and benefits DBKWS and NLIDB by capturing contextual information in English queries.

In upcoming work, we plan to explore and compare the performance in construction of queries of deep-learning approaches using sentiment-analysis against our SCB system.

REFERENCES

- [1] P. B. Brandtzaeg and A. Følstad, "Chatbots: changing user needs and motivations," *interactions*, vol. 25, no. 5, pp. 38–43, Aug. 2018, doi: 10.1145/3236669.
- [2] D. Altinok, "An Ontology-Based Dialogue Management System for Banking and Finance Dialogue Systems," p.9.
- [3] T.P. Nagarhalli, V. Vaze, and N. K. Rana, "A Review of Current Trends in the Development of Chatbot Systems," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Mar. 2020, pp. 706–710. doi: 10.1109/ICACCS48705.2020.9074420.
- [4] "Chatbot Market Size & Share | Growth Forecast Report 2024," *Global Market Insights Inc.* <https://www.gminsights.com/industry-analysis/chatbot-market> (accessed Mar. 18, 2022).
- [5] I. Intelligence, "Chatbot market in 2022: Stats, trends, and companies in the growing AI chatbot industry," *Business Insider.* <https://www.businessinsider.com/chatbot-market-stats-trends> (accessed Mar. 18, 2022).
- [6] Clickatell, "Chatbot Market to grow at 31 Percent CAGR from 2018 to 2024," *Supply and Demand Chain Executive*, Jul. 02, 2018. <https://www.sdcexec.com/software-technology/news/21011880/chatbot-market-to-grow-at-31-percent-cagr-from-2018-to-2024> (accessed Mar. 18, 2022).
- [7] M. J. Minock, "A STEP Towards Realizing Codd's Vision of Rendezvous with the Casual User," p. 4, 1977.
- [8] V. Zhong, C. Xiong, and R. Socher, "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning," *arXiv:1709.00103 [cs]*, Nov. 2017, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/1709.00103>
- [9] T. Yu *et al.*, "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task," *arXiv:1809.08887 [cs]*, Feb. 2019, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/1809.08887>
- [10] B. Wang, R. Shin, X. Liu, O. Polozov, and M. Richardson, "RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers," *arXiv:1911.04942 [cs]*, Aug. 2021, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/1911.04942>
- [11] C. Finegan-Dollak *et al.*, "Improving Text-to-SQL Evaluation Methodology," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 351–360, 2018, doi: 10.18653/v1/P18-1033.
- [12] A. P. Parikh *et al.*, "ToTTo: A Controlled Table-To-Text Generation Dataset," *arXiv:2004.14373 [cs]*, Oct. 2020, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/2004.14373>
- [13] L. Deng, S. Zhang, and K. Balog, "Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval," *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1029–1032, Jul. 2019, doi: 10.1145/3331184.3331333.
- [14] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Wang, "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data," *arXiv:2004.07347 [cs]*, May 2021, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/2004.07347>
- [15] J. Sangeetha and R. Hariprasad. An intelligent automatic query generation interface for relational databases using deep learning technique. *International Journal of Speech Technology*, 22(3):817–825, 2019.
- [16] R. Bordawekar and O. Shmueli. Using word embedding to enable semantic queries in relational databases. In *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning*, pages 1–4, 2017.
- [17] S. Abbas, M. U. Khan, S. U.-J. Lee, A. Abbas, and A. K. Bashir, "A Review of NLIDB With Deep Learning: Findings, Challenges and Open Issues," *IEEE Access*, vol. 10, pp. 14927–14945, 2022, doi: 10.1109/ACCESS.2022.3147586.
- [18] C. Wang *et al.*, "Robust Text-to-SQL Generation with Execution-Guided Decoding," *arXiv:1807.03100 [cs]*, Sep. 2018, Accessed: Mar. 18, 2022. [Online]. Available: <http://arxiv.org/abs/1807.03100>

Multiobjective Optimal Control of Power Electronic Loads in Small Scale Power Systems

Ramitha K. Dissanayake
*Department of Electrical and Electronic
 Engineering*
University of Peradeniya
 Peradeniya 20400, Sri Lanka
 e15084@eng.pdn.ac.lk

Amal R. Wimalaratna
School of Engineering
RMIT University
 Melbourne, Australia
 s3863302@student.rmit.edu.au

Anushka M. Dissanayake
Research and Development
Schweitzer Engineering Laboratories
 Pullman, USA
 anushka_dissanayake@selinc.com

Abstract—This paper introduces an optimal controller for the power electronics loads (PELs) based on the goal attainment multi objective optimization and linear quadratic regulator (LQR) architecture. Optimal set point of the PEL is derived to regulate the desired load demand while maximizing the input bus voltage of the PEL. Then a LQR controller is developed to drive the PEL state to the optimal equilibrium. Dynamic modeling of the PEL is done in the energy and admittance domain. Simulations were carried out to demonstrate the effectiveness and the applicability of the proposed method.

Keywords—DC microgrids, multi objective optimization, optimal control, power electronic converters

I. INTRODUCTION

To fulfil the ever increasing electricity demand due to the population growth and the technological advancement, energy authorities and policy planners are focusing on finding more reliable and flexible power systems. Over the years traditional power generation systems used to supply the electricity demands of the consumers [1], [2]. Due to the increased demand for electricity, high penetration of renewable energy sources into the existing grid and having a unidirectional passive distribution network [2], [3], these traditional power systems require significant re-modification in infrastructure in order to fulfil the future power system requirements which consist of bidirectional active distribution network [3], [4].

Contrary to large scale power systems, small scale power systems (SSPS) consist of localized generation, loads, storage systems and control systems [5]. Power systems of naval ships and communication centers and Micro-grids (MGs) are some of the examples for SSPS [6]. By implementing SSPSs such as MGs can offer many benefits to both electricity utility and the users. With the utilization of many distribution generations in MGs, transmission losses can be reduced as the power flow reduced in transmission and distribution lines. In addition to that improvement of network and power quality, reduction in emissions and cost are some other advantages provided by the MGs [7].

Typical MG consists of distributed generation sources, distribution systems, storage systems and communication and control systems [8]. Having low voltage operational networks

and capability of operating in different configurations such as grid connected mode and islanding mode are some of the special features in these SSPSs [7], [9]. MGs under SSPSs can be broadly categorized as AC, DC and hybrid MGs [10]. Compared to the AC MGs, DC MGs are becoming more popular due to the rapid growth in DC loads, DC based distributed generations such as solar PV, fuel cell and DC based energy storage technologies [11]. Efficiency improvement and reduction in losses due to the abatement of converters used for DC loads; smooth integration of distributed energy resources such as solar PV and energy storages are some of the benefits provided by the DCMGs [10]. Considering these factors, this paper focuses on the DCMGs.

Since the loads and sources in a DCMG have multiple objectives, MG control problem can be identified as a coupled multi objective optimization (MOO) problem. For example, sources try to maintain their load bus voltage around the desired voltage while loads try to regulate their input power at the desired value. Further, as an interconnected system, both loads and sources attempt to maintain the system stability while minimizing system losses. Game theoretic control and optimal control are two major control architectures which can be brought in to develop controllers that can solve MOO problems [12].

In game theoretic controllers, the active components in the system has to be modelled as players. Any constituent that influences the energy flow of the system such as loads, sources is considered as a player and objective function for each player should be defined [6], [13], [14]. On the contrary, optimal control is a tradeoff between requirements of the loads and the requirements of the sources [15]. Broadly speaking, in the optimal control, active components try to find the best possible transient path in the startup or subjected to a disturbance which minimizes a given cost function or individual cost functions by maneuvering their own control inputs [16].

In this paper, goal attainment MOO technique is employed to obtain the optimal set point of the PEL which minimizes set of multi objectives. Active power regulation and bus voltage maximization has been considered as the multi objectives in the modeling. Then, to maneuver the PEL state to this optimal set point, a LQR optimal controller is also proposed.

Rest of the paper is organized as follows. In the next section, PELs are modeled as an adaptive admittance. In section III, multi-objective problem for optimal PEL control is formulated. Section IV describes the solution for the multi-objective problem formulated. Then the PELs are modeled dynamically in the admittance domain in Section V. Section VI describes the optimal control sequence required. Simulation results are presented in Section VII and the paper is concluded in section VIII.

II. ADAPTIVE ADMITTANCE REPRESENTATION OF PELS

Power and impedance are the common parameters used to represent a power system load. In MG domain, End Loads (ELs) and power grid are connected through a Power Electronic Converter (PEC) as shown in Fig. 1 (a). PEC acts as an intermediate device between MG and EL. MG comprises of interconnected sources and loads while the ELs can be single or multiple loads with either constant impedance or current or power.

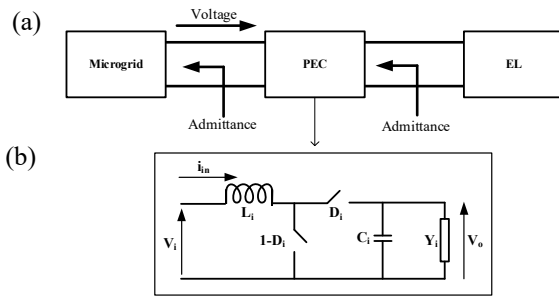


Fig.1 (a) Power electronic converter as an interface between the power system and end load (b) boost converter as a power electronic converter

Any DC-DC converter topology can be used as a PEC. Consider an average mode boost converter as the PEC as shown in Fig. 1 (b). EL properties can be varied by controlling the PEC input characteristics as the PEC works as a voltage and impedance translator between the MG and EL. In Fig. 1 (b), V_i , i_{in} , V_o and D_i represent the input voltage, input current, output voltage and duty cycle respectively. For any time, MG sees the input admittance of the PEC as,

$$y_i = \frac{i_{in}}{V_i} \quad (1)$$

If the PEC is assumed to be lossless, from the power balance,

$$V_i^2 y_i = V_o^2 Y_i \quad (2)$$

where Y_i represents the equivalent admittance of the EL. From (1) and (2), the input admittance of the boost converter based PEL can be shown as,

$$y_i = \frac{Y_i}{D_i^2}; \quad D_i \in [0,1] \quad (3)$$

This representation yields that the PEC based load can be modeled as an adaptive admittance and it acts as a member of the power system as a single quantity.

III. MULTIOBJECTIVES IN POWER SYSTEMS

A. Multiobjective formulation for Optimal PEL Control

There can be multiple different conflicting objectives in SSPSs control point of view. However, in DCMG perspective, active power and voltage regulation have paramount importance and will be the focus of this paper. Each PEL has the desire to regulate its active power input from the grid and it will be considered as the first objective in this paper as,

$$f_{1,i}(x_i) = (y_i V_i^2 - P_i)^2 \quad (4)$$

where, $x_i = [y_i \ V_i]$ is the decision variable vector, y_i , V_i and P_i are the input admittance, input voltage and active power demand of the i^{th} PEL. When the load demand increases, its terminal voltage decreases. Therefore, PEL would like to regulate its input voltage closer to the nominal value. However, regulation of both power and voltage could leads to unsatisfactory results since there is an inverse conflicting nature between the objectives. Hence, to overcome this issue and to satisfy voltage uplift, the second objective of the PEL is considered as maximization of its input voltage. Converting the maximization problem into a minimization problem yields,

$$f_{2,i}(x_i) = -V_i \quad (5)$$

In turn, minimization of this objective supports to increase the overall grid voltage which will be an extra advantage. The final MOO problem of the i^{th} PEL can be represented as follows.

$$\min F_i(x_i) = [f_{1,i}(x_i) \ f_{2,i}(x_i)] \quad (6)$$

subjected to: $\gamma_i^L \leq \gamma_i \leq \gamma_i^U$

where, γ_i contains all the constraints bounded between the lower and upper limits given by γ_i^L and γ_i^U respectively. This constraint vector includes the allowable maximum and minimum values of the decision variables and load demands.

Despite the fact that the input voltage is a function of PEL input admittance, and other power system parameters, it is treated as a variable in this implementation. Reason for this modification is to make the optimization problem decentralized and less complex. Even though input voltage is highly coupled and depends on other parameters which are not available to the local PEL, its value can be always measured locally. Hence, an additional constraint is included when solving the MOO problem to satisfy the equality of the measured input voltage and calculated optimal input voltage.

The MOO algorithm computes the optimal input admittance (y_i^*) and optimal input voltage (V_i^*) of the PEL as,

$$x_i^* = [y_i^* \ V_i^*] = \arg \min_{x_i} F_i(x_i) \quad (7)$$

subjected to: $\gamma_i^L \leq \gamma_i \leq \gamma_i^U$

Then the optimal control algorithm discussed in section V and VI utilizes the optimal input admittance and measure the corresponding input voltage in the steady state. This measured value must be the same as the optimal value computed from the MOO algorithm. This process continues until it finds a local optimal solution which satisfies both objectives and all the constraints. Solution to the defined MOO problem will be discussed in the next section.

IV. SOLUTION TO THE MULTIOBJECTIVE OPTIMIZATION PROBLEM

Optimization problem is a mathematical model with the main objective of maximizing desirable attributes or minimizing undesirable attributes [17]. One way of categorizing optimization methods is single objective optimization (SOO) and multi objective optimization (MOO). In SOO, solution is found considering a single objective function while in MOO, two or more objective functions are utilized to solve the problem [18]. Most of the realistic optimization problems have multi objective nature [19]. The best way to solve a MOO problem is obtaining the corresponding Pareto optimal front by simultaneous minimization of the objectives and then extracting a compromise solution [18]. Weighted sum approach [20], global criterion method [21], complex method [22] and goal attainment [23] are classical methods of MOO. Intelligent methods like simulated annealing [24], fuzzy logic approach [25] and genetic algorithms [26] are also used in MOO. In this paper, goal attainment is used to solve the multi objective problem.

Goal attainment method can be used in nonlinear programming problems. This method has a set of design goals $F_i^* = \{F_1^*, F_2^*, \dots, F_n^*\}$ associated with a set of objectives $F(x) = \{F_1(x), F_2(x), \dots, F_n(x)\}$. The solution is obtained by formulating the optimization problem such that the objective vector to be relatively imprecise about the design goal vector. The weighting coefficient vector $w = \{w_1, w_2, \dots, w_n\}$ is introduced to control the relative degree of over or underachievement of goals. This can be expressed as a standard optimization problem [23].

Minimize γ ; subject to,

$$F_i(x) - w_i\gamma \leq F_i^* ; i = 1, 2, \dots, n \quad (8)$$

The slackness term $w_i\gamma$ is introduced which otherwise goals to be rigidly met. In MATLAB fgoalattain multi objective solver is used to solve (7) in this paper.

V. DYNAMIC MODELING OF PELS IN ADMITTANCE DOMAIN

Dynamic modeling of the PELs in admittance and energy domain is discussed in this section. Derivation is given for the PEL based on the boost PEC. However, same procedure can be extended to any other DC-DC PEC. Consider the average mode boost PEC dynamic model given by,

$$L_i \frac{di_{in}}{dt} = V_i - D_i V_o \quad (9)$$

$$C_i \frac{dV_i}{dt} = D_i i_{in} - V_o Y_i \quad (10)$$

where L_i and C_i represent the inductance of the inductor and capacitance of the capacitor of the boost converter respectively. Define two new states, capacitor energy storage ($w_i = \frac{1}{2} C_i V_o^2$) and input admittance ($y_i = \frac{i_{in}}{V_i}$). In the context of these two states, the dynamic model can be reformulated as [12],

$$\frac{dw_i}{dt} = V_i^2 y_i - \frac{2}{R_i C_i} w_i \quad (11)$$

$$\frac{dy_i}{dt} = u_i \quad (12)$$

Here the control input u_i is given by,

$$u_i = \frac{1}{L_i} \left[1 - \frac{D_i}{V_i} \sqrt{\frac{2w_i}{C_i}} \right] \quad (13)$$

Once the optimal control input u_i^* is obtained, the optimal duty cycle can be extracted as,

$$D_i^* = (1 - L_i u_i^*) V_i^* \sqrt{\frac{C_i}{2w_i^*}} \quad (14)$$

In order to convert the control problem to an optimal control problem, error system dynamic model is obtained by defining the new state $\xi_i = [w_i - w_i^* \quad y_i - y_i^*]^T$. Here the optimal capacitor energy storage is derived from the desired EL voltage which is the output voltage of the converter. A boost PEC with the desired output voltage of v_d is considered in this paper and hence the optimal capacitor energy storage can be derived as,

$$w_i^* = \frac{1}{2} C_i v_d^2 \quad (15)$$

Under this state transformation, the obtained state space model of the PEL can be shown as,

$$\frac{d\xi_i}{dt} = \alpha_i \xi_i + \beta_i u_i \quad (16)$$

where, the system matrix and the control effectiveness matrix are given by,

$$\alpha_i = \begin{bmatrix} -\frac{2}{R_i C_i} & V_i^2 \\ 0 & 0 \end{bmatrix} \quad (17)$$

$$\beta_i = [0 \quad 1]^T \quad (18)$$

Since the system matrix contains the input PEL voltage, this dynamic is not linear. However, compared to the input admittance and capacitor energy storage state dynamics, input voltage variation is relatively slow. Hence, in the modeling input voltage is measured and assumed constant for a small time interval δt . This time interval is in the range of milliseconds and it's a reasonable assumption. In this period, the system matrix is constant, and dynamics can be treated as linear. The optimal control law is derived for this short δt period. After that, the input voltage is measured again, and the optimal control problem is solved for the next δt period with the updated input voltage. Optimal control problem and the solution is discussed in the next section.

VI. OPTIMAL MANEUVER OF PELS

Once the optimal states are computed as explained in the sections III and IV, the goal will be to drive the system states to those desired values. Finding out the best possible control sequence which can maneuver the system from a given initial state to the optimal values will be the focus of this section. The best control sequence for a given dynamic objective function is defined by the optimal control law. Since the simplified dynamics given by (16) in the linear form for a short δt period,

linear quadratic regulator (LQR) is employed to obtain the optimal dynamic behavior.

In the LQR control architecture, the optimal feedback control signal is obtained as,

$$u_i^* = -K_i \xi_i^* \quad (19)$$

which minimizes the performance index from the initial time t_0 to $t_0 + \delta t$,

$$J_i = \int_{t_0}^{t_0+\delta t} \xi_i^T Q_i \xi_i + u_i^T \rho_i u_i \, dt \quad (20)$$

Here, Q_i and ρ_i are positive definite diagonal gain matrices. In this paper, considered dynamical system is a single input system, ρ_i is a scalar constant. The optimal feedback gain matrix is given by K_i . The optimal cost of the performance index is a quadratic function of the optimal state as,

$$J_i^* = \xi_i^{*T} \psi_i \xi_i^* \quad (21)$$

where, ψ_i is the solution of the Lyapunov equation given by

$$(\alpha_i - \beta_i K_i)^T \psi_i + \psi_i (\alpha_i - \beta_i K_i) + Q_i + K^T \rho_i K_i = 0 \quad (22)$$

The solution of the LQR problem is associated with the algebraic Riccati equation (ARE) shown below.

$$\alpha_i^T \psi_i + \psi_i \alpha_i + Q_i - \psi_i \beta_i \rho_i^{-1} \beta_i^T \psi_i = 0 \quad (23)$$

The optimal feedback gain matrix is updated with the solution of the ARE as,

$$K_i = \rho_i^{-1} \beta_i^T \psi_i \quad (24)$$

There exist multiple ways of solving the ARE and this paper utilizes the Matlab LQR.

VII. SIMULATION RESULTS

Simulations were carried out considering the IEEE 9 bus test system [27]. A PEL was connected at bus 5 and two resistive loads with $20 \, \Omega$ were considered at buses 6 and 8. Nominal DC grid voltage was set to 120 V and the desired end load voltage (v_d) was considered as 230 V. The demand, inductance, capacitance and load resistance of the PEL were considered as 1 kW, 1 mH, 500 μF and 10 Ω respectively. In the goal attainment, equal unity weights were chosen for both the objectives with 0 and -120 goal values for objective 1 and 2 respectively. The lower and upper bounds for the input admittance were set to 0 S and 10 S while the corresponding values for the input voltage were 110 V and 130 V. First simulation started with no load initial conditions with zero input current and hence the input admittance is zero. Second simulation demonstrates a load change from 1 kW to 3 kW. The LQR controller parameters were considered as $\delta t = 1 \, \text{ms}$, $Q_i = 2 \times 2$ identity matrix, $\rho_i = 1$. Matlab LQR function has been used here to obtain the optimal feedback gains.

A. No Load Initial Transient

With the load demand of 1kW and under the constraints, the optimal input admittance and the input voltages of the PEL were found as 0.0720 S and 117.7654 V. The dynamic response of the LQR controller is shown in Fig. 2 to Fig.5.

Fig. 2 shows that the optimal states of input admittance and capacitor energy storage have reached the steady values of 0.07 S and 13.8 J respectively after 0.1 s. Starting values for both states were zero. According to Fig. 3, input voltage has reached 117.75 V within 0.1 s while the power of the PEL has reached the expected 1 kW value within the same time duration. Fig. 4 shows that the optimal control input has gone to zero after 0.1 s duration and the corresponding optimal duty ratio has reached 0.51 within the same time duration. Starting value for optimal control input was 2.5 and for duty ratio, it was 0.75. Then the optimal gains have reached 0.0399 and 33.3 as shown in Fig. 5. Time taken to reach the steady value is 0.1 s. Initial values for gains were 0.0409 and 34.2.

B. Load Change Scenario

In this simulation, the load demand was changed from 1 kW to 3 kW and corresponding optimal solution and variations in the states are shown below. The step load change was initiated after 300 ms from the startup. After the load change, the optimal input admittance and the input voltage of the PEL was found as 0.2283 S and 114.5890 V respectively. Due to the demand increase, the input admittance is increased to allow more current into the PEL. On the other hand, the terminal voltage drops to a lower value as a consequence of the load increment. The optimal dynamic response of the states, input voltage, power, control input and duty cycle are shown in Fig. 6 to Fig. 8.

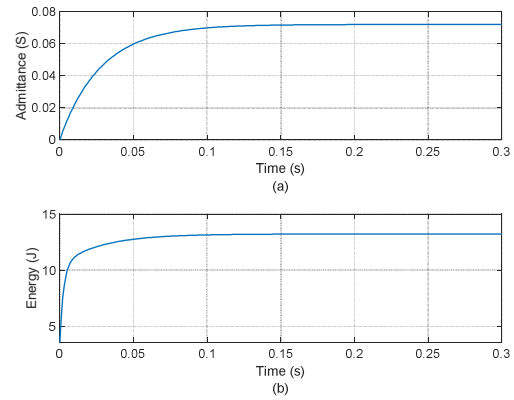


Fig. 2 Variation of the optimal states, (a) input admittance, (b) capacitor energy storage.

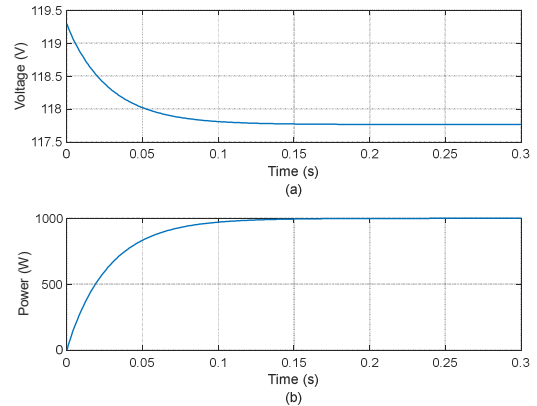


Fig. 3 Variation of the input voltage and power of the PEL, (a) input voltage, (b) input power.

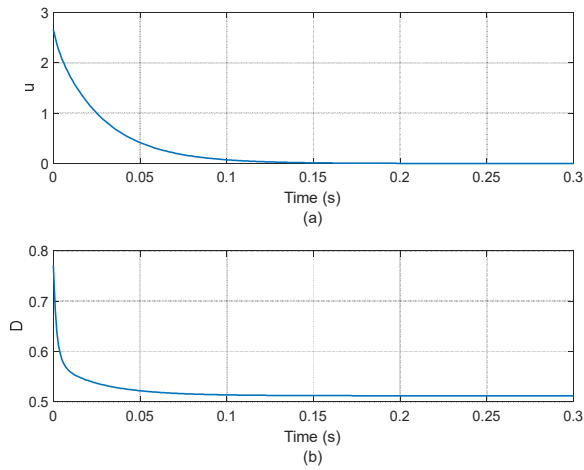


Fig. 4 variation of the optimal control input and the duty cycle, (a) optimal control input (u), (b) optimal duty cycle.

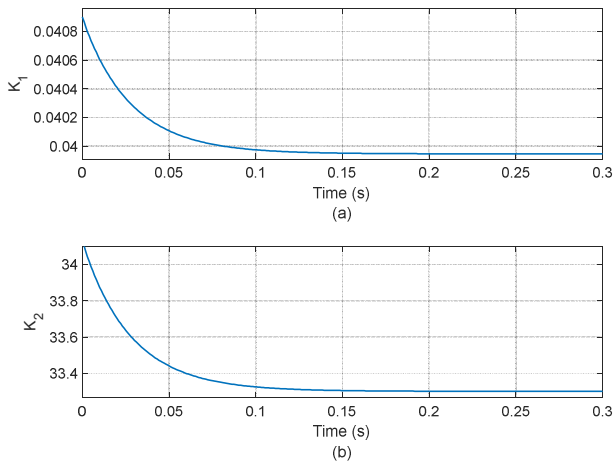


Fig. 5 Variation of the optimal feedback gains, (a) K_1 , (b) K_2 .

When considering Fig. 6, during the initial load demand of 1 kW, input admittance has reached 0.075 S and capacitor energy has reached 13 J. After changing the load demand to 3 kW, admittance has risen to 0.225 S within 0.1 s while the capacitor energy storage has shown a dip of 7 J minimum value for 0.1s and again reached the steady value of 13 J.

According to Fig. 7, input voltage has reached 117.8 V steady value from the starting value of 119 V within 0.1 s for 1 kW load demand. After increasing the load demand to 3 kW, input voltage has further reduced to 114.6 V steady value within 0.1 s. Furthermore, the power of PEL has reached 1 kW and 3 kW steady values according to the load demand changes.

Fig. 8 demonstrates that the optimal control input has reached zero value from the starting value of 3 within 0.1 s. After changing the load demand to 3 kW, the optimal control input has increased to 5 as a large control effort is required to get the voltage and power back to the desired values. After 0.1 s, again the optimal control input has reached zero steady value. The corresponding optimal duty ratio has started from 1 at 1 kW load starting transient and reached the steady value of 0.51 within 0.1

s. Then, at the moment of changing the load demand, there is a sudden increment up to 0.61 and after 0.1 s it has reached back the steady value of 0.4.

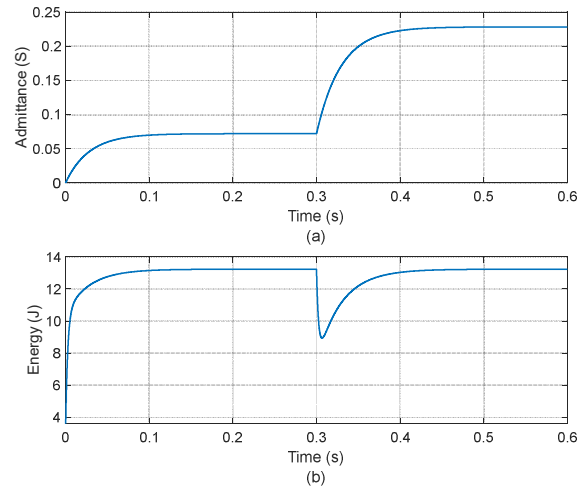


Fig. 6 Variation of the optimal states subjected to a load change, (a) input admittance, (b) capacitor energy storage.

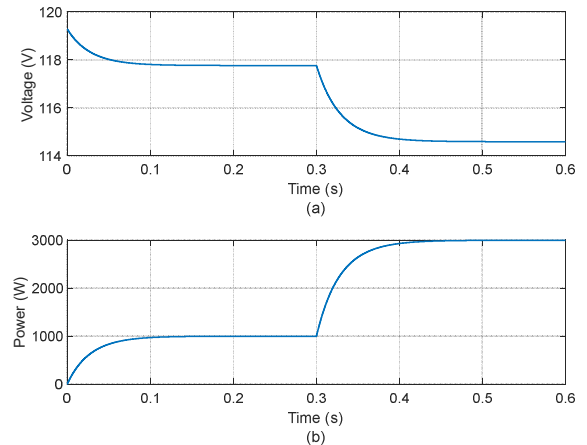


Fig. 7 Variation of the input voltage and power of the PEL subjected to a load change, (a) input voltage, (b) input power.

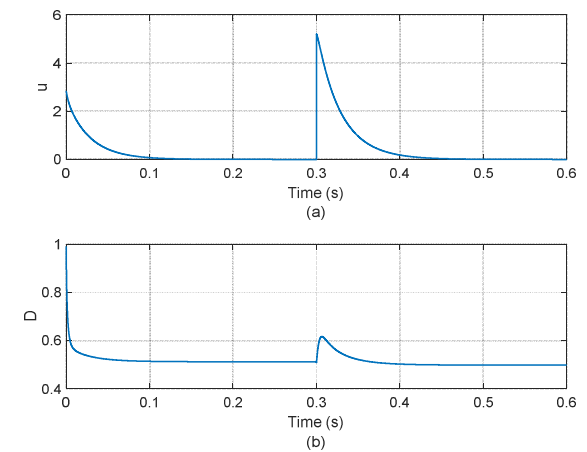


Fig. 8 variation of the optimal control input and the duty cycle subjected to a load change, (a) optimal control input (u), (b) optimal duty cycle.

VIII. CONCLUSION

This paper has presented a multi objective optimal control methodology to obtain the optimal control input and states of power electronic loads in small scale power systems. PEL in SSPS was modelled using the adaptive admittance representation of power electronic loads. Active power input regulation from the grid and maximization of input voltage of the PEL were considered as the multi objectives . Then, goal attainment method has been used to solve the MOO problem to obtain the optimal input admittance for each PEL. With the introduction of two new states, capacitor storage and input admittance, dynamic model of the PEL was obtained to develop an optimal controller to drive the system to the calculated optimal desired input admittance. Subsequently, LQR method was utilized to solve the developed optimal control problem. Simulations of the system were carried out considering the IEEE 9 bus test system and two scenarios were used to demonstrate the effectiveness of the proposed concept. According to the results it can be seen that the proposed methodology is capable of regulating the PEL demand while maximizing its input bus voltage. Experimental verification of the proposed concept would be the major future direction of this work.

REFERENCES

[1] F. Blaabjerg, Y. Yang, D. Yang and X. Wang, "Distributed Power-Generation Systems and Protection," *Proceedings of the IEEE*, vol.105, no. 7, pp. 1311-1331, July 2017, doi: 10.1109/JPROC.2017.2696878.

[2] H. J. Bhatti and M. Danilovic, "Making the World More Sustainable: Enabling Localized Energy Generation and Distribution on Decentralized Smart Grid Systems", *World Journal of Engineering and Technology*, vol. 6, no. 2, pp. 350–382, 2018.

[3] L. F. N. Delboni, D. Marujo, P. P. Balestrassi and D. Q. Oliveira, "Electrical power systems: Evolution from traditional configuration to distributed generation and microgrids," *Microgrids Design and Implementation*, Cham, Switzerland:Springer, pp. 1-23, 2019.

[4] M. Henderson, D. Novosel and M. Crow, "Electric power Grid modernization trends challenges and opportunities", *IEEE Adv. Technol. Humanity*, 2017.

[5] A. Bidram, F. L. Lewis, and A. Davoudi, "Distributed Control Systems for Small-Scale Power Networks: Using Multiagent Cooperative Control Theory", *IEEE Control Systems*, vol. 34, no. 6, pp. 56–77, 2014.

[6] W. W. Weaver and P. T. Krein, "Game-Theoretic Control of Small-Scale Power Systems," *IEEE Transactions on Power Delivery*, vol. 24, no. 3, pp. 1560-1567, July 2009, doi: 10.1109/TPWRD.2008.2007022.

[7] H. B. Santoso and B. Budiyo, "Microgrid Development Using A Grid Tie Inverter", *MAKARA of Technology Series*, vol. 17, no. 3, 2014.

[8] L. Mariam, M. Basu and Michael F. Conlon, "A Review of Existing Microgrid Architectures", *Journal of Engineering*, vol. 2013, ArticleID 937614, 8 pages, 2013.

[9] F. Khavari, A. Badri, A. Zangeneh and M. Shafiekhani, "A comparison of centralized and decentralized energy-management models of multi-microgrid systems," *2017 Smart Grid Conference (SGC)*, 2017, pp. 1-6, doi: 10.1109/SGC.2017.8308837.

[10] H. Lotfi and A. Khodaei, "AC Versus DC Microgrid Planning," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 296-304, Jan. 2017, doi: 10.1109/TSG.2015.2457910.

[11] R. E. Hebner et al., "Technical cross-fertilization between terrestrial microgrids and ship power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 2, pp. 161-179, April 2016, doi: 10.1007/s40565-015-0108-0.

[12] A. M. Dissanayake and N. C. Ekneligoda, "Online game theoretic feedback control of DC microgrids," *2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2018, pp. 1-5, doi: 10.1109/ISGT.2018.8403394.

[13] S. Mei, W. Wei, and F. Liu, "On engineering game theory with its application in power systems", *Control Theory and Technology*, vol. 15, no. 1, pp. 1–12, 2017.

[14] L.-L. Fan, V. Nasirian, H. Modares, F. L. Lewis, Y.-D. Song, and A. Davoudi, "Game-Theoretic Control of Active Loads in DC Microgrids", *IEEE Transactions on Energy Conversion*, vol. 31, no. 3, pp. 882–895, 2016.

[15] W. W. Weaver and P. T. Krein, "Optimal Geometric Control of Power Buffers," *IEEE Transactions on Power Electronics*, vol. 24, no. 5, pp. 1248-1258, May 2009, doi: 10.1109/TPEL.2008.2012260.

[16] A. M. Dissanayake and N. C. Ekneligoda, "Droop-Free Optimal Feedback Control of Distributed Generators in Islanded DC Microgrids," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 2, pp. 1624-1637, April 2021, doi: 10.1109/JESTPE.2019.2953869.

[17] S. Rao, 2009. *Engineering optimization*. 4th ed. New Jersey: John Wiley & Sons Inc, pp.1-4.

[18] A. M. Dissanayake and N. C. Ekneligoda, "Multiobjective Optimization of Droop-Controlled Distributed Generators in DC Microgrids," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2423-2435, April 2020, doi: 10.1109/TII.2019.2931837.

[19] L. Xiujian and S. Zhongke, "Overview of multi-objective optimization methods," *Journal of Systems Engineering and Electronics*, vol. 15, no. 2, pp. 142-146, June 2004.

[20] X. Cai, H. Sun, Q. Zhang and Y. Huang, "A Grid Weighted Sum Pareto Local Search for Combinatorial Multi and Many-Objective Optimization," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3586-3598, Sept. 2019, doi: 10.1109/TCYB.2018.2849403.

[21] J. H. Gomes, Aluizio Ramos Salgado Júnior, A. P. Paiva, J. Ferreira, S. C. Costa and P. Balestrassi. "Global Criterion Method Based on Principal Components to the Optimization of Manufacturing Processes with Multiple Responses," *Strojniski Vestnik-journal of Mechanical Engineering*, no. 58, pp. 345-353, 2012.

[22] A. Adelman, and W. F. Stevens, "Process optimization by the complex method", *AIChE J.*, vol. 18, no. 1, pp. 20-24, 1972, doi: 10.1002/aic.690180105

[23] N. Ravikumar, C. ShekarBesta and M. Chidambaram, "Multivariable control of unstable systems by goal attainment method," *2017 Trends in Industrial Measurement and Automation (TIMA)*, 2017, pp. 1-5, doi: 10.1109/TIMA.2017.8064806.

[24] H. Zhang, J. Wu, C. Sun, M. Zhong and R. Yang, "A Multi-objective Particle Swarm Optimizer Based on Simulated Annealing and Decomposition," *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2018, pp. 262-273, doi: 10.1109/CCIS.2018.8691225.

[25] Yanfang Shou and Jianmin Xu, "Multi-objective optimization of oversaturated signalized intersection based on fuzzy logic," *2010 8th World Congress on Intelligent Control and Automation*, 2010, pp. 5008-5013, doi: 10.1109/WCICA.2010.5554700.

[26] C. Liu, "New Multi-objective Genetic Algorithm for Nonlinear Constrained Optimization Problems," *IEEE International Conference on Automation and Logistics*, 2007, pp. 118-120, doi: 10.1109/ICAL.2007.4338541.

[27] S. Sharma, N. S. Velgapudi and K. Pandey, "Performance analysis of IEEE 9 Bus system using TCSC," *Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, 2017, pp. 251-256, doi: 10.1109/RDCAPE.2017.8358277.

Explicite model of a wheel-soil interaction over a rough terrain using terramechanics law

Rania Majdoubi

LCS Laboratory, faculty of sciences,
Mohammed V University in Rabat
rania_majdoubi@um5.ac.ma

Lhousasaine Masmoudi

LCS Laboratory, faculty of sciences,
Mohammed V University in Rabat
Lhmasmoudi@gmail.com

Abderrahmane Elharif

Mechanical laboratory, faculty of
sciences, Mohammed V University in
Rabat
elharifa@gmail.com

Abstract—Modeling the interaction of a mobile robot's wheels with the ground, is a key element in evaluating its performance. In this paper, to evaluate vehicle performance in a simulation environment, research studies were conducted to capture the physics in a mathematical framework. Vehicle performance can be directly correlated to forces and moments between the tire and the ground, and indirectly related to terrain variations such as sinking. The main objective of this paper is to propose a mathematical model describing the dynamic behavior of the robot wheel navigating a deformable soil. This model is validated using experimental data obtained from a test bench of a lunar wheel.

Keywords—mobile robot, the performance of the vehicle, tire, soil, ground.

I. INTRODUCTION

The modeling and control of the robot depends on the type of mission it must accomplish and the environment it must face [1–4]. Therefore, it is essential to know as much as possible what happens at the wheel-ground interaction in order to characterize and exploit useful information. The study of this interaction has been the subject of much research, with the objective of determining patterns and factors affecting the behavior of vehicles intended to travel over any terrain [5]. Tires are perhaps the most important component, but the most difficult to model. Not only do tires support the vehicle and dampen road irregularities, but they also provide the longitudinal and lateral forces necessary to change the speed and direction of the vehicle. These forces are produced by the deformation of the tire where it comes into contact with the road during acceleration, braking and navigation. In the literature we distinguish the Rigid Ground modeling, which presents the studies that include the effects of wheeled vehicles on undeformable ground. The main constraint is the study of the behavior of the vehicle on the road and the characterization of the wheel-ground interaction.

Over the years, a wide variety of models have been developed to formulate and simulate the wheel-soil interaction. The degree of complexity of these models is based on the application, accuracy and computational cost of development as elaborated in the Pacejka Model [6], the Bross

Model [7], the Gim Model [8], the Dugoff Model [9], the Kiench/Buchhardt Model [10] and the Linear Model [11]. The theory of deformable soil or Terra-mechanics which is defined as the science of studying the properties of the soil when interacting with both tracked and wheeled vehicles. As mentioned earlier, the main challenge in studying off-road vehicle behavior is the characterization of the wheel-ground interaction. This interaction is shown in literature as elaborated in the WES VCI model [12], the WES numerical mobility model [13], the Deere & Company Technical Center Models [14], the STIREMOD model [15], the VTIM Model (Vehicle Terrain Interaction Model) [16], the discrete element method (DEM) [17], the NWVPM model [18], the VDANL model [19], the SCM model [20] and the FTire model (Flexible Ring Tire Model) [16].

The main objective of this paper is to propose a mathematical model describing the dynamic behavior of the robot wheels navigating a deformable ground. This model is experimentally validated and gives good results. In this perspective, we will present in the second paragraph the dynamic analysis using the classical Bekker model which we will discuss the modeling hypothesis and the classical model which include the semi empirical equations. the third part. is dedicated give the explicit form of the laws of Bekker and the validation. At the end of this paper, we will summarize the results that we have reached, in order to reuse it in the next papers.

II. DYNAMIC ANALYSIS USING THE CLASSICAL BEKKER MODEL

A. Modeling hypothesis

To simplify the model, simplifying assumptions are adopted:

- The wheel is rigid and assimilated to a perfect smooth cylinder.
- The point contact is located on the plane of the trajectory.
- The ground is plastically deformed.
- The indentation is assumed to be negligible with respect to the wheel radius.

- The point contact is located on the plane of the trajectory.
- The wheels remain in contact with the road.
- The variation of the radius of the rolling wheel is neglected.
- In practice, the roll and pitch angles, as well as the tire angles are small, so we assume small angles.

B. Fundamental Basics

In this paragraph, we present some fundamental notions for the modeling of the wheel-ground contact using the classical Bekker model. These concepts are discussed as follows:

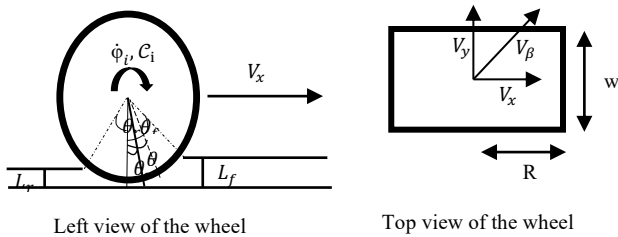


Figure 1. The geometric model adopted for the wheel-ground contact

- In the case where the robot moves on a trajectory :
The wheel can only produce thrust if $\dot{\phi}_i = \frac{V_x}{R}$
This phenomenon is called wheel slip rate, symbolized by s such that :

$$s = \begin{cases} \frac{R\dot{\phi}_i - V_x}{R\dot{\phi}_i} & \text{During acceleration} \\ \frac{R\dot{\phi}_i - V_x}{V_x} & \text{During braking} \end{cases} \quad (1)$$

- As in a curve, the vehicle has a lateral velocity can drive in relation to an angle called the slip angle defined by the expression above:

$$\beta = \text{atan} \left(\frac{V_y}{V_x} \right) \quad (2)$$

- The wheel motion causes the development of normal and shear stresses along the wheel-ground interface between θ_f and θ_r . That's why Bekker came to characterize the soil types, and to measure the properties of the soil using the bevametre technique. The latter allows for penetration tests (normal displacement stress) and shear tests (displacement shear stress).
- Using experimental tests and the bevameter technique, the researchers have developed functions to describe the stress distribution along the wheel-ground interface.

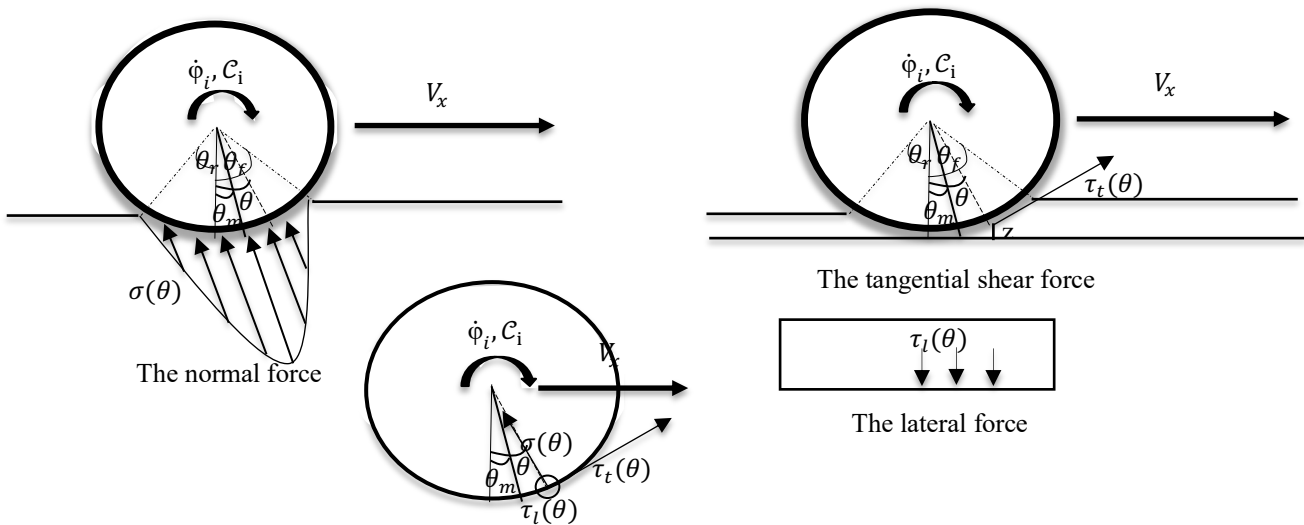


Figure 2. Geometric modeling of wheel-ground contact stresses in motion

Wong and Reece in [21], developed a form of the normal pressure at the wheel-soil interface using the results of plate penetration tests:

$$\sigma(\varphi_i) = \left(\frac{k_c}{w} + k_\phi \right) z^n(\varphi_i) \quad (3)$$

In which;

$$z(\varphi_i) = \begin{cases} R(\cos(\varphi_i) - \cos(\theta_f)) & \text{if } \theta_m < \varphi_i < \theta_f \\ R(\cos(\theta_e) - \cos(\theta_f)) & \text{if } \theta_r < \varphi_i < \theta_m \end{cases}$$

$$\theta_e = \theta_f - (\varphi_i - \theta_r)(\theta_f - \theta_m)/(\theta_m - \theta_r)$$

While;

- φ_i is the rotation angle of the wheel
- w the width of the wheel
- k_c, k_ϕ and n are respectively the modulus of cohesion, the modulus of friction and the exponent of sinking (bevametre)
- $z(\varphi_i)$ is the sinking of the soil along the wheel-soil interface

θ_f is the entry angle at which the wheel begins to touch the ground

θ_r is the exit angle at which the wheel loses contact with the ground, which is defined by :

$$\theta_r = (b_0 + b_1 s)\theta_f$$

θ_m is the maximum stress angle, which is defined as:

$$\theta_m = (a_0 + a_1 s)\theta_f$$

In which;

a_0, a_1, b_0 and b_1 Are soil constants.

The lateral shear rate is defined by :

$$V_{JL}(\varphi_i) = V_y \quad (4)$$

The tangential shear rate is defined by :

$$V_{Jt}(\theta) = R\dot{\varphi}_i - V_x \cos(\varphi_i) \quad (5)$$

The compression speed is defined by :

$$V_{Jn}(\varphi_i) = V_x \sin(\varphi_i) \quad (6)$$

The shear deformation of the soil can be defined as :

$$j_k(\varphi_i) = \int_{\theta}^{\theta_f} V_{Jk}(\varphi_i) \frac{1}{\dot{\varphi}_i} d\varphi_i \quad \text{in} \quad (7)$$

which ; $k=t,L$

Hence;

$$j_t(\varphi_i) = R(\theta_f - \varphi_i) - (1 - s)(\sin(\theta_f) - \sin(\varphi_i)) \quad (8)$$

And;

$$j_L(\varphi_i) = R(1 - s)(\theta_f - \varphi_i) \tan(\beta) \quad (9)$$

In the case of a plastic soil, Janosi, Hanamoto and Bekker developed an experimental behavior law that determines the shear stress:

$$\tau_k(\varphi_i) = \tau_{max} \tau_{base1} \quad k=t,L \quad (10)$$

In which;

$$\tau_{max} = C + \sigma(\varphi_i) \tan(\phi)$$

$$\tau_{base1} = 1 - \exp\left(-\frac{j_k}{K}\right) \quad k=t,L$$

While;

C : Soil cohesion

ϕ : Soil internal friction angle

K : Shear strain parameter (measures the amount of shear displacement required for the development of maximum shear stress). These values are given in the following table:

Table 1 The values of the shear modulus K(cm)

Soil type	K (cm)
Sandy loam	2.5 - 2.54
Snow	4
Clay	0.6 - 2.54
Sand	1 - 2.5

By linearizing, L_f et L_r are small in front of R. we obtain

The normal constraint is shown as follows:

$$\sigma(\varphi_i) = \begin{cases} \sigma_m \frac{\theta_f - \varphi_i}{\theta_f - \theta_m} & \text{if } \theta_m < \varphi_i < \theta_f \\ \sigma_m \frac{\varphi_i - \theta_r}{\theta_m - \theta_r} & \text{if } \theta_r < \varphi_i < \theta_m \end{cases} \quad (11)$$

While;

$$\sigma_m = \left(\frac{k_c}{b} + k_\phi\right) R^n (\cos(\theta_m) - \cos(\theta_f))^n$$

And the shear stress is :

$$\tau_k(\varphi_i) = \begin{cases} \tau_{m,k} \frac{\theta_f - \varphi_i}{\theta_f - \theta_m} & \text{if } \theta_m < \varphi_i < \theta_f \\ \tau_{m,k} \frac{\varphi_i - \theta_r}{\theta_m - \theta_r} & \text{if } \theta_r < \varphi_i < \theta_m \end{cases} \quad (12)$$

$k=L,t$

While;

$$\tau_{m,k} = (C + \sigma_m \tan(\phi))(1 - \exp\left(-\frac{j_k(\theta_m)}{K}\right)) \quad k=L,t$$

The reaction between the tire and the roadway is described by three forces and three moments, as modeled by the diagram defined below:

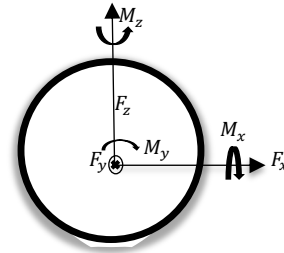


Figure 3 Modeling of wheel-ground contact forces

Net pulling force

(Drawbar pull) F_x : This force corresponds to the push force generating the wheel motion, which is equal to the difference between the drawbar pull and the resultant of the wheel resistances. Using the normal force $\sigma(\varphi_i)$ and the shear force $\tau_t(\varphi_i)$, the net tensile force (drawbar pull) acting in the direction from the ground to the wheel is calculated by integrating the input angle θ_f to the output angle θ_r , and is given by

$$F_x = R w \int_{\theta_r}^{\theta_f} (-\sigma(\varphi_i) \sin(\varphi_i) + \tau_t(\varphi_i) \cos(\varphi_i)) d\varphi_i \quad (13)$$

The lateral force

F_y : The authors modeled the force that acts along the lateral direction of the wheel when the vehicle makes a steering maneuver; as follows:

$$F_y = F_{sy} + F_{by} \quad (14)$$

While;

- F_{sy} is the force produced by $\tau_L(\varphi_i)$ under the wheel, is defined by :

$$F_{sy} = R w \int_{\theta_r}^{\theta_f} \tau_L(\varphi_i) d\varphi_i \quad (15)$$

- F_{by} Is the bulldozing resistance that is developed when a considerable mass of soil is moved by the wheel. This type of resistance is common when the wheel compresses the surface layers of the soil and pushes the compressed soil in front of and behind the tire. The phenomenon of bulldozing has appeared in the case of a

wide wheel (approx. 34 cm wheel width), crossing very loose soils. This phenomenon is estimated to cause a significant increase in the overall resistance to movement for values of depression greater than 6% of the wheel diameter. This force is defined as follows:

$$F_{by} = \int_{\theta_r}^{\theta_f} dF_b(\varphi_i) d\varphi_i \quad (16)$$

The vertical force

F_z : is obtained by the same method as for the longitudinal force F_x . The necessary condition for this force is that it must be equal to the normal load on the wheel. It is given by the following equation:

$$F_z = R w \int_{\theta_r}^{\theta_f} (\sigma(\varphi_i) \cos(\varphi_i) + \tau_t(\varphi_i) \sin(\varphi_i)) d\varphi_i \quad (17)$$

The moment of overturning

M_x : is present when the reaction of the ground on the tire F_{zr} (oriented upwards) is not aligned with the vertical load F_z (oriented downwards), It is given by the following equation:

$$M_x = -R^2 w \int_{\theta_r}^{\theta_f} \tau_L(\varphi_i) \cos(\varphi_i) d\varphi_i \quad (18)$$

Le moment de résistance au roulement

M_y : is the resistance to the forward motion of a free rolling tire with almost zero longitudinal slip. The rolling resistance is present as soon as the tires start to roll forward. This moment is defined as:

$$M_y = R^2 w \int_{\theta_r}^{\theta_f} \tau_t(\varphi_i) d\varphi_i \quad (19)$$

The moment of self-alignment

M_z : tends to rotate the tire about its z axis and align the x axis with the direction of the wheel speed vector. This moment always tends to reduce the drift angle α . It is due to the displacement of the lateral force behind the center of the contact area, It is given by the following equation:

$$M_z = -R^2 w \int_{\theta_r}^{\theta_f} \tau_L(\varphi_i) \sin(\varphi_i) d\varphi_i \quad (20)$$

III. ELABRATION OF THE EXTENDED BEKKER MODEL

A. Simplification of forces

Taking the condition L_f et L_r are \ll in front of R. And by linearizing .we find the following equations :

$$F_z = R w \tau_{m,L} \left(\frac{\cos(\theta_m) - \cos(\theta_r)}{\theta_m - \theta_r} + \frac{\cos(\theta_m) - \cos(\theta_f)}{\theta_f - \theta_m} \right) - R w \sigma_m \left(\frac{\sin(\theta_m) - \sin(\theta_r)}{\theta_m - \theta_r} + \frac{\sin(\theta_m) - \sin(\theta_f)}{\theta_f - \theta_m} \right) \quad (21)$$

$$F_y = R w \tau_{m,L} \frac{\theta_f - \theta_r}{2} + 2k_p c \left(\frac{\theta_f - \theta_r}{2} + \frac{\sin(2\theta_f) - \sin(2\theta_r)}{4} + \sin(\theta_r) - \sin(\theta_f) \right) + \frac{1}{2} \rho g k_p R^3 \left(\theta_f - \theta_r + \frac{\sin(2\theta_f) - \sin(2\theta_r)}{2} + \frac{\sin(3\theta_f) - \sin(3\theta_r)}{12} + \frac{7}{4} (\sin(\theta_f) - \sin(\theta_r)) \right) \quad (22)$$

$$F_z = R w \sigma_m \left(\frac{\cos(\theta_m) - \cos(\theta_r)}{\theta_m - \theta_r} + \frac{\cos(\theta_m) - \cos(\theta_f)}{\theta_f - \theta_m} \right) + R w \tau_{m,L} \left(\frac{\sin(\theta_m) - \sin(\theta_r)}{\theta_m - \theta_r} + \frac{\sin(\theta_m) - \sin(\theta_f)}{\theta_f - \theta_m} \right) \quad (23)$$

$$M_x = -R^2 w \tau_{m,L} \left(\frac{\cos(\theta_m) - \cos(\theta_r)}{\theta_m - \theta_r} + \frac{\cos(\theta_m) - \cos(\theta_f)}{\theta_f - \theta_m} \right) \quad (24)$$

$$M_y = R^2 w \tau_{m,L} \frac{\theta_f - \theta_r}{2} \quad (25)$$

$$M_z = -R^2 w \tau_{m,L} \left(\frac{\sin(\theta_m) - \sin(\theta_r)}{\theta_m - \theta_r} + \frac{\sin(\theta_m) - \sin(\theta_f)}{\theta_f - \theta_m} \right) \quad (26)$$

To simplify the model, we take: $\theta_r = -\theta_f$ and $\theta_m = 0$

Hence, we find;

$$F_z = 2R w \sigma_m \left(\frac{1 - \cos(\theta_f)}{\theta_f} \right) \approx R w \sigma_m \theta_f \quad (27)$$

While;

$$\sigma_m = \left(\frac{k_c}{w} + k_\phi \right) R^n (1 - \cos(\theta_f))^n \approx \left(\frac{k_c}{w} + k_\phi \right) R^n \left(\frac{\theta_f^2}{2} \right)^n \quad (28)$$

It is the expression of a non-linear spring because it is of the form $F_z = K_0 z^n$

While ;

$$z = R^{(1+\frac{1}{n})} \theta_f^{(2+\frac{1}{n})} \quad \text{Et} \quad K_0 = w \left(\frac{k_c}{w} + k_\phi \right)$$

This model is not valid in the transient regime.

To approach the real model Crahn [22] has developed a model that takes into account the sinking speed of particles in a granular soil after several penetration measurements.

Then, we obtain:

$$\sigma(\varphi_i) = \left(\frac{k_c}{w} + k_\phi \right) z^n(\varphi_i) K_q \dot{z}^q(\varphi_i) \quad (29)$$

Therefore, we added the term $\dot{z}^q(\varphi_i)$ a Where q is a new parameter characterizing the soil, which is called the soil penetration speed exponent. K_q is a positive coefficient, whose values are given in Table as follows:

Table 2 Soil penetration speed exponent

Soil type	q	K _q
compact Soil	0.30	2.86
Soft sand	0.30	2.86
Sandy loam	0.12	1

The problem occurs when $\dot{z} = 0$

Hence;

$$\sigma(\varphi_i) = \left(\frac{k_c}{w} + k_\phi\right) z^n(\varphi_i)(1 + K_q \dot{z}^q(\varphi_i)) \quad (30)$$

With this model, when the wheel is rotating, without normal component of the hub speed. We have:

$$\dot{z}(\varphi_i) = R\dot{\varphi}_i \sin(\varphi_i) \approx R\dot{\varphi}_i \varphi_i \quad (31)$$

Hence;

$$\sigma(\varphi_i) = \left(\frac{k_c}{w} + k_\phi\right) z^n(\varphi_i)(1 + K_q (R\dot{\varphi}_i \varphi_i)^q) \quad (32)$$

$$\sigma_m = \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \quad (33)$$

- the normal force is given in the following expression:

$$F_z \approx R w \theta_f \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \quad (34)$$

- The lateral force is given by:

$$F_y \approx R w \tau_{m,t} \theta_f \quad (35)$$

While;

$$\tau_{m,t} = (C + \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \tan(\phi)) \left(1 - \exp\left(-\frac{R(1-s)\theta_f \tan(\beta)}{K}\right)\right)$$

- The net traction is given in :

$$F_x = 2R w \tau_{m,t} \left(\frac{1 - \cos(\theta_f)}{\theta_f}\right) \approx R w \tau_{m,t} \theta_f \quad (36)$$

While;

$$\tau_{m,t} = (C + \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \tan(\phi)) \left(1 - \exp\left(-\frac{R \theta_f s}{K}\right)\right)$$

- The overturning moment is given in:

$$M_x = -2R^2 w \tau_{m,t} \left(\frac{1 - \cos(\theta_f)}{\theta_f}\right) \approx -R^2 w \tau_{m,t} \theta_f \quad (37)$$

While;

$$\tau_{m,t} = (C + \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \tan(\phi)) \left(1 - \exp\left(-\frac{R(1-s)\theta_f \tan(\beta)}{K}\right)\right)$$

- The rolling resistance moment is given by:

$$M_y \approx R^2 w \tau_{m,t} \theta_f \quad (38)$$

In which:

$$\tau_{m,t} = (C + \left(\frac{k_c}{w} + k_\phi\right) R^n \left(\frac{\theta_f^2}{2}\right)^n (1 + K_q (R\dot{\varphi}_i \theta_f)^q) \tan(\phi)) \left(1 - \exp\left(-\frac{R \theta_f s}{K}\right)\right)$$

- The self-aligning moment is given by :

$$M_z \approx 0 \quad (39)$$

B. Model validation

The use of this model respects the algorithm as discussed below:

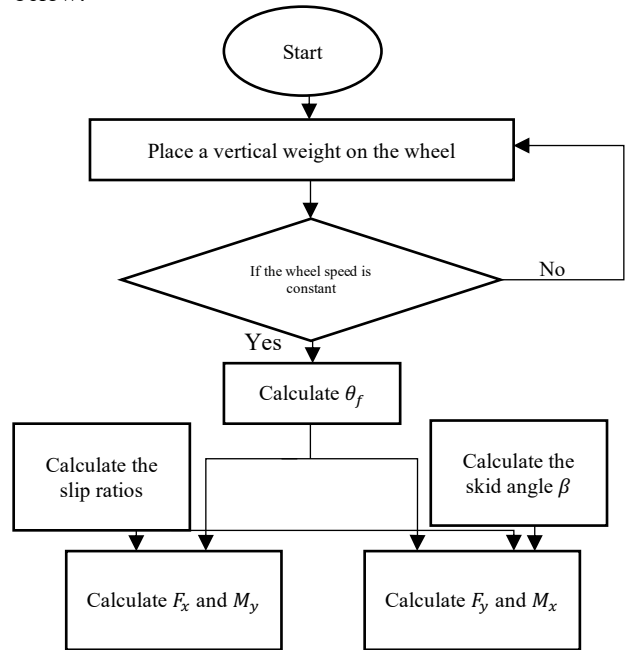


Figure 4. Algorithm for calculating forces and moments according to the extended Bekker's law

The numerical simulation procedure to obtain the forces and moments, introducing the normal load W of the wheel, the slip ratio s and the slip angle β. This Load allows us to calculate the input angle at which the wheel begins to touch the ground θ_f from the equilibrium relationship between W and F_z (equation (27)), to finally determine the traction force F_x and the lateral force F_y using equations (35) and (36).

To validate the wheel-ground contact model, we use the experimental data previously performed by Yoshida and Ishigami in [23]. This experiment was carried out on a single wheel test rig (wheel of a lunar vehicle) (Figure 5). The characteristic parameters of the model are defined in the table below:

Table 3 Parameters of the wheel of the lunar vehicle used as well as the parameters of the ground used during the experiment

Parameters	Value
C (KPa)	0.8
ϕ (deg)	37.2
k _c (N/m ⁿ⁺¹)	1.37 × 10 ³
k _ϕ (N/m ⁿ⁺²)	81.4 × 10 ⁴
N	1
K (m)	0.1
R (m)	0.09
w (m)	0.1
K _q	0
β (deg)	25

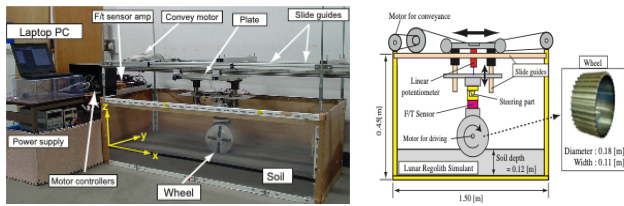


Figure 5 The Yoshida and Ishigami test rig [23]

The average variation of longitudinal and lateral force as a function of slip variation obtained experimentally after collecting about 100 force measurements at a given slip rate, are represented in concert with our model results as discussed in the following figures:

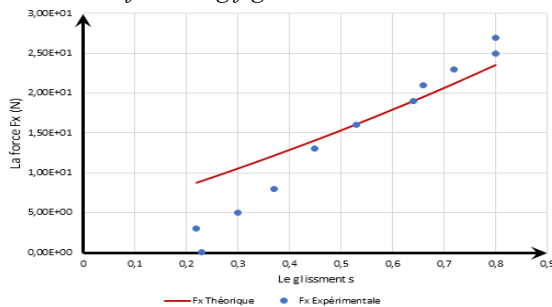


Figure 6. The traction force as a function of the slip

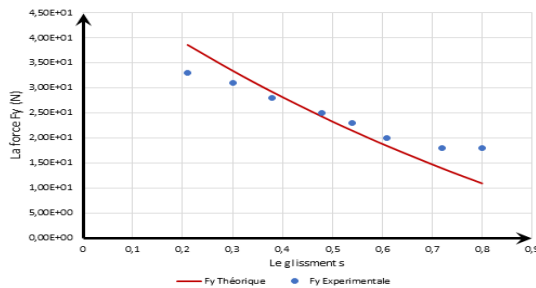


Figure 7. Lateral force as a function of slip

The Figure 6 shows that the tractive force increases with the slip ratio. This behavior is explained by the fact that the soil deformation (shear stress) in the longitudinal direction of the wheel increases with the slip ratio; this results in a large soil deformation which, in turn, generates a large tensile force on the drawbar. Also, the differences between the experimental and theoretical values are relatively small in the range of slip ratios above 0.4; however, relatively larger differences are observed for smaller slip ratios.

The Figure 7 shows that the lateral force decreases with the slip rate. In addition, it is observed that the lateral force whose maximum value is $s = 0$; this is due to the fact that the lateral velocity, proportionally to the longitudinal velocity. The theoretical curves show a relatively small difference, so we can deduce that the theory agrees well with the experimental results plotted.

These results confirm that the wheel-ground contact model proposed in this section is capable of representing the wheel motion behavior and contact/traction forces with appropriate accuracy.

IV. CONCLUSION

The main objective of the paper is to propose a mathematical model describing the dynamic behavior of the robot wheel navigating a deformable ground. This model is experimentally validated and gives good results.

The adaptation of the ground navigation model of the "Agri-Eco-Robot" robot will be the subject of future work.

ACKNOWLEDGMENT

The authors of this paper are thankful to the Ministry of Higher Education and Scientific Research of Morocco (MESRSFC), and the National Center for Scientific and Technical Research of Morocco (CNRST) for financing this project.

REFERENCES

- [1] R. Majdoubi, L. Masmoudi, and A. Elharif, "Torque control using metaheuristic optimization for optimal energy consumption of a BLDCM," *2021 IEEE Int. IOT, Electron. Mechatronics Conf. IEMTRONICS 2021 - Proc.*, 2021, doi: 10.1109/IEMTRONICS52119.2021.9422624.
- [2] Rania MAJDOUBI, Lhoussaine MASMOUDI, Mohammed BAKHTI, Abderrahmane ELHARIF, "Parameters estimation of bldc motor based on physical approach and weighted recursive least square algorithm," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 133–145, 2020, doi: 10.11591/ijece.v11i1.pp133-145.
- [3] R. Majdoubi, L. Masmoudi, M. Bakhti, and B. Jabri, "Torque Control Oriented Modeling of a Brushless Direct Current Motor (BLDCM) Based on the Extended Park's Transformation," *J. Eur. des Syst. Autom.*, vol. 54, no. 1, pp. 165–174, 2021.
- [4] R. Majdoubi and L. Masmoudi, "Eco-design of a mobile agriculture robot based on classical approach and FEM criteria," *Proc. - IEEE 2021 Int. Conf. Comput. Commun. Intell. Syst. ICCIS 2021*, pp. 978–982, 2021, doi: 10.1109/ICCIS51004.2021.9397234.
- [5] A. Maarif, W. Rahmani, M. A. M. Vera, A. A. Nuryono, R. Majdoubi, and A. Cakan, "Artificial Potential Field Algorithm for Obstacle Avoidance in UAV Quadrotor for Dynamic Environment," *10th IEEE Int. Conf. Commun. Networks Satell. Comnetsat 2021 - Proc.*, pp. 184–189, 2021, doi: 10.1109/COMNETSAT53002.2021.9530803.
- [6] H. B. Pacejka and I. J. M. Besselink, "Magic Formula tyre model with transient properties," *Veh. Syst. Dyn.*, vol. 27, no. Suppl, pp. 37–41, 1997, doi: 10.1080/00423119708969658.
- [7] J. Svendenius and B. Wittenmark, "Brush tire model with increased flexibility," *Eur. Control Conf. ECC 2003*, no. 1, pp. 1863–1868, 2003, doi: 10.23919/ecc.2003.7085237.
- [8] P. E. N. Gwanghun Gim, "Analytical model of pneumatic tyres for vehicle dynamic simulations. Part 2. Comprehensive slips," *Aerosp. Mech. Eng.*, vol. 12, no. 1, pp. 19–39, doi: 10.1007/s40435-013-0032-y.
- [9] H. Yokohama, "Investigation of Nonlinear Contact Problem in Pneumatic Tyres Interacting With Road Surface," 2010.
- [10] U. K. and L. Nielsen., *Automotive control systems*. 2000.
- [11] D. Lechner and D. Lechner, "Analyse du comportement dynamique des véhicules routiers légers : développement d'une méthodologie appliquée à la sécurité primaire To cite this version : HAL Id : tel-01730790," 2018.
- [12] A. A. Rula and C. J. Nuttall, "An Analysis of Ground Mobility Models (ANAMOB)," p. 333, 1971, [Online]. Available: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html>

- &identifier=AD0886513.
- [13] N. C. Rula AA, "An Analysis of Ground Mobility Models (ANAMOB).," *U.S. Army Eng. Waterw. Exp. Station.*, 1971.
- [14] R. D. Wismer and H. J. Luth, "Off-road traction prediction for wheeled vehicles," *J. Terramechanics*, vol. 10, no. 2, pp. 49–61, 1973, doi: 10.1016/0022-4898(73)90014-1.
- [15] R. W. Allen, J. P. Chrstos, and T. J. Rosenthal, "Tire model for use with vehicle dynamics simulations on pavement and off-road surfaces," *Veh. Syst. Dyn.*, vol. 27, no. Suppl, pp. 37–41, 1997, doi: 10.1080/00423119708969663.
- [16] J. Madsen *et al.*, "A Physics-Based Vehicle/Terrain Interaction Model for Soft Soil Off-Road Vehicle Simulations," *SAE Int. J. Commer. Veh.*, vol. 5, no. 1, pp. 280–290, 2012, doi: 10.4271/2012-01-0767.
- [17] H. Tanaka, M. Momozu, A. Oida, and M. Yamazaki, "Simulation of soil deformation and resistance at bar penetration by the Distinct Element Method," *J. Terramechanics*, vol. 37, no. 1, pp. 41–56, 2000, doi: 10.1016/S0022-4898(99)00013-0.
- [18] J. Y. Wong and V. M. Asnani, "Study of the correlation between the performances of lunar vehicle wheels predicted by the Nepean wheeled vehicle performance model and test data," *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.*, vol. 222, no. 11, pp. 1755–1770, 2008, doi: 10.1243/09544070JAUTO811.
- [19] C. Y. Liang, R. W. Allen, T. J. Rosenthal, J. P. Chrstos, and P. Nunez, "Tire modeling for off-road vehicle simulation," *SAE Tech. Pap.*, no. 724, 2004, doi: 10.4271/2004-01-2058.
- [20] M. G. Bekker, "Introduction to terrain-vehicle systems," *Univ. Michigan Press*, 1969, doi: 10.1016/j.bpj.2019.04.004.
- [21] J. Y. Wong and A. R. Reece, "Prediction of rigid wheel performance based on the analysis of soil-wheel stresses. Part II. Performance of towed rigid wheels," *J. Terramechanics*, vol. 4, no. 2, pp. 7–25, 1967, doi: 10.1016/0022-4898(67)90047-X.
- [22] M. Grahn, "Prediction of sinkage and rolling resistance for off-the-road vehicles considering penetration velocity," *J. Terramechanics*, vol. 28, no. 4, pp. 339–347, 1991, doi: 10.1016/0022-4898(91)90015-X.
- [23] G. Ishigami, K. Nagatani, A. Miwa, and K. Yoshida, "JFR_ishigami.pdf."

Optimal Inventory Policy in Oil Transportation: A Case Study

Rasha Kashef
Electrical, Computer, and Biomedical Engineering
Toronto Metropolitan University
Toronto, CANADA
rkashef@ryerson.ca

Shuo Xu
Azure Practice
Adastra North America
Toronto, CANADA
shuo.xu@adastragr.com

Abstract— The oil industry has become a significant industry that promotes the economy of North America. Finding the optimal inventory strategy and matching transportation methods have become a critical challenge in the oil industry to increase revenue while minimizing overhead costs. This paper focuses on the midstream and downstream sectors in the oil industry. The primary objective of this paper is to use discrete event simulation to find the optimal combination of transportation methods and the optimal inventory policy. Using a conceptual supply chain model and real data collection, experimentally, we can see that the best policy is using pipeline speed 1m/s with reorder point 500L and reorder quantity 35000L. Various pilot and production runs have been implemented.

Keywords— *Oil Transportation, Inventory Policy, Simulation, Conceptual Modeling, Optimization*

I. INTRODUCTION

The oil industry has become a significant industry that promotes the economy of North America. In 2020 Canada's three biggest oil companies (Imperial Oil, Husky Energy, and Suncor Energy) brought in record profits of \$11.75 billion. Most Canadian petroleum production, approximately 283,000 cubic meters per day (1,780,000 bbl/d), is exported, and Canada is the largest single source of oil imports into the United States [1]. The oil industry is divided into three components: Upstream, midstream, and downstream. The upstream is commonly known as the exploration and production sector, which includes searching for potential oil/gas fields, drilling wells, etc. The midstream sector involves transportation and wholesale marketing of refined petroleum products. The downstream involves refining, marketing, and distribution issues and more frequently targets end consumers through products such as gasoline, diesel, heating oil, etc. [2]-[5]. In this paper, we focus on the midstream and downstream sectors in the oil industry. The problem addressed in this paper includes the supply chain management of logistics and inventory. Two transportation schemes are discussed from the logistic perspective, pipeline and truck transportation. Different transportation methods will lead to different costs and time (lead time), affecting management decisions and demand satisfaction. From the inventory perspective, when a certain volume of oil product is ordered and stocked in a specific region to be sold, there will be a holding cost [6]-[9]. The high inventory will lead to high holding costs, and low inventory will risk the stock out. The proper trade-off between over-stocking and low-stocking levels

will reduce the cost and increase the profit. The best combination of transportation schemes and optimal inventory policy contributes to the least cost for corporation operations and improves efficiency in the supply chain sector. The primary objective of this paper is to find the optimal combination of transportation schemes and the optimal inventory policy using Discrete Event Simulation (DES) [10]-[12]. In this paper, the logistics methods (i.e., truck transportation, pipeline transportation), inventory level, reorder point, and transportation time affect the final cost result. We used different parameters for these criteria to find the optimal strategy to achieve a minimum cost. From the simulation results, inventory levels in different locations (distribution centers and local gas stations) will be collected to estimate the inventory cost. The number of times that the truck runs will also be collected to calculate the logistic cost. The best policy is using pipeline speed 1m/s with reorder point 500L and reorder quantity 35000L. The rest of this paper is organized as follows: Section 2 discusses the Canadian oil industry. The DES conceptual model is discussed in section 3. Various data requirements are discussed in section 4. Section 5 presents the proposed simulation model. The pilot and production runs of the model are discussed in sections 6 and 7, respectively. Finally, conclusions and future directions are summarized in Section 8..

II. THE CANADIAN OIL INDUSTRY

To fully present the oil industry's real supply chain model, research on the Canadian Oil industry is conducted from the supply, logistic, and end-user market perspective. Alberta is the largest producer of conventional crude oil, synthetic crude, and natural gas and gas products in the supply sector in Canada. Two of the largest producers of petrochemicals in North America are located in central and north-central Alberta. The total refined product produced in Alberta is about 20414 thousand cubic meters. From the logistics perspective, Pipeline transportation is an important sector of the Canadian economy. Canada is a significant energy producer globally, and pipeline transportation is a practical and economical way of moving large quantities of crude oil and natural gas. Transmission pipelines transport nearly 97% of Canadian natural gas and crude oil production. The network of pipelines in Canada is shown in Fig.1. We can see that the pipeline is applied in long-distance transportation. Hence in our model, from the Alberta to Toronto hub, the major transportation will be the pipeline.

Oil pipelines are made from steel or plastic tubes with an inner diameter of 4 to 48 inches (100 to 1,200 mm). Most pipelines are buried at about 3 to 6 feet (0.91 to 1.8 m). Various methods protect pipes from impact, abrasion, and corrosion, including wood lagging (wood slats), concrete coating, rock shield, high-density polyethylene, imported sand padding, and padding machines. The oil is kept in motion by pump stations along the pipeline and usually flows at about 1 to 6 meters per second (3.3 to 20 ft/s). The different flow rates lead to different volumes of oil products. The estimated unit cost for Oil Pipeline transportation [2] is illustrated in Table 1.

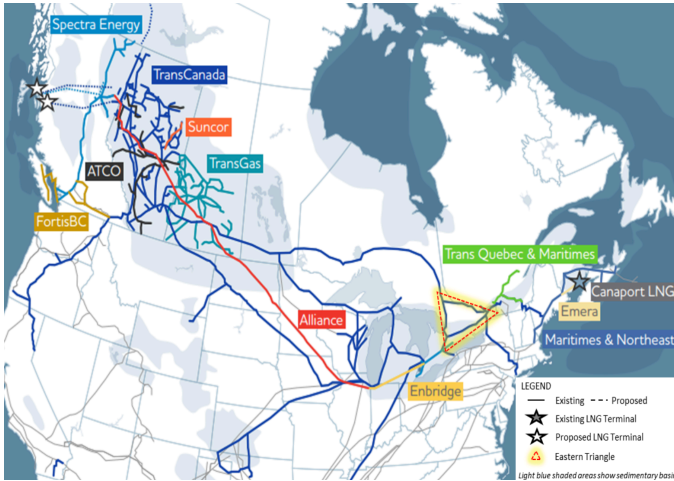


Fig. 1 Canadian Energy Pipeline Association, 2016.

Table 1: Estimated Unit Cost for Oil Pipeline Transportation (\$ per cubic meter-kilometre)

	Low	Base	High
Oil Unit Cost	\$0.011415	\$0.011968	\$0.012521

The oil tank truck will be applied for short-distance transportation due to its relatively low fixed cost and flexibility compared to the pipeline. Large trucks typically have capacities ranging from 20,800 to 43,900 Liters. Smaller tank trucks with less than 11,000 L are generally used to deal with light liquid cargo within a local community. Another common use is to deliver fuel with a capacity of 3,800 Liters. The truck operational cost in Canada from East to West Corridor LNG consists of the following cost components: Tire cost, Repair cost, Licensing Cleaning cost, driver cost, equipment ownership, administration, and interest costs. The collected cost data includes only the driver's wages, such that we can estimate the total truck cost [3].

III. THE CONCEPTUAL MODEL

The simulated supply chain system contains entities such as suppliers, distributors, retailers, and customers. To simplify the model, we assume only one manufacturer, distributor, and retailer in the supply chain. The process in this model can be described as follows: the supplier produces the products. The distributor delivers the products to the retailer when the retailer orders a specific quantity from the manufacturer. Each day,

customers come to the retailer with particular demands. When the inventory decreases to a certain point, the retailer will make the order. Fig.2 illustrates the material flow from left to right and the information flow (order) from right to left in the proposed conceptual model. We propose a conceptual supply chain model from the supplier to the end customer, as shown in Fig.3.



Fig. 2 Material Flow in the Supply Chain Model.

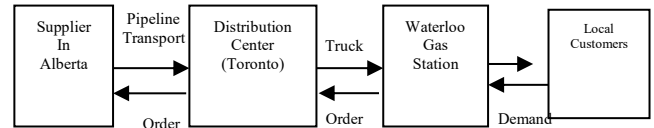


Fig. 3 The Supply Chain Conceptual Model.

We assume the supplier is located in Alberta, an oil production hub in Canada. The oil will be transferred from Alberta to Toronto via pipeline. A fuel tank truck will be used for short-distance transportation from Toronto to Waterloo local gas station. The Toronto distribution center and Waterloo local gas station will have a warehouse to stock a certain amount of gasoline. When the gasoline product's inventory level reduces to a specific threshold, which will be different for different locations, the staff will place an order for the upper level and wait for the order refill.

A. Assumptions

- We proposed one completed supply chain rather than a Multi-Echelon Model.
- We have focused on the perspective of inventory fluctuation rather than the queuing model; hence, we aggregate servers into one server.
- We have estimated the cost based on the data we collected from [2][3].
- We assumed the customer demand is based on the fuel capacity of the cars, and we estimated the average fuel capacity for three major types of vehicles in the market based on the market share statistic report published by the government [5].

IV. DATA REQUIREMENTS

In our model, four primary data sources are needed, including the inter-arrival time of cars entering a single gas station, the distribution of different types of vehicles entering the station, the average service time of each vehicle at the gas station, and the cost of transporting gas from Toronto to Waterloo. A real-data collection process took place at the ESSO gas station at King St. North and Weber St. in Waterloo, Ontario, each day from 11 am-12 pm and 5 pm to 6 pm and collected the inter-arrival time. We categorize vehicles into three types, small cars, SUVs, and Trucks. We received the sales data from the manufacturer's website, i.e., Ford, Chevy, Toyota, and Honda. We also calculated the service time for each car. Finally, we have collected an estimate for the truck transportation price and

pipeline transportation cost [2][3]. We have fit the inter-arrival data (Fig.4) to multiple distributions (Fig.5). We can see that the best fitting distribution of the data is exponential with rate=1/7.3197. A Kolmogorov-Smirnov test is performed to test this hypothesis. Based on the p-value, we did not reject the Null Hypothesis that the data follows the exponential distribution with a mean inter-arrival time of 7.3197 minutes. We have followed the same procedure in checking the distributions of other variables.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
IAT	Service	IAT	Service	IAT	Service	IAT	Service
2	4	17	18	2	22	22	
7	13	3	5	17	5	1	
0	3	11	1	14	0	1	
7	2	4	1	5	1	5	
12	2	0	3	18	1	1	
4	2	2	12	1	3	1	
4	2	1	2	10	12	6	
6	3	1	1	6	1	1	
1	5	9	3	4	1	4	
5	7	12	2	13	3	9	
5	2	8	16	1	2	0	
15	4	17	1	1	16	3	
22	6	4	3	2	1	2	
5	2	8	0	4	3	12	
2	1	8	2	0	0	12	
4	4	17	4	3	2	17	
13	2	2	3	36	1	1	
		3	0		4	2	
		1	1		9	1	
		4	9				
		9	12				
		0	4				
		5	8				
		1	4				
		1	1				

Fig. 4. Vehicles Inter-arrival time

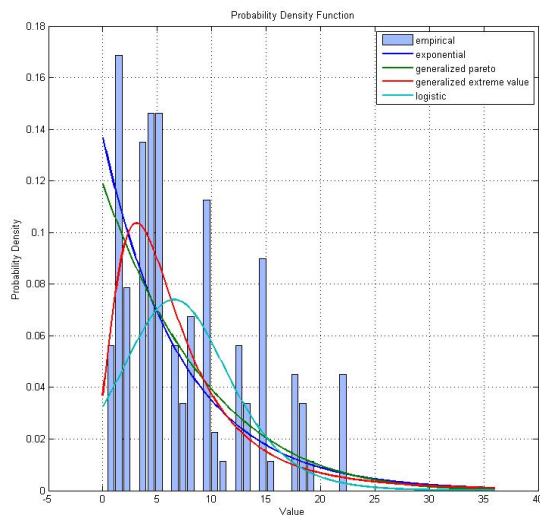


Fig. 5 Distribution Fitting

V. OIL TRANSPORTATION: A DISCRETE EVENT SIMULATION MODEL

Based on our conceptual model, we use two tanks to represent the main supplier of gas (Alberta) and the distributor (Toronto) (simplified version), as shown in Fig. 6. The simulation is conducted using SIMUL8 [6]. The flow rate between tanks is determined in the VL box. We determined the (s, S) policy inside the VL box to ensure that the flow is allowed when the lower tank level falls below a certain amount.

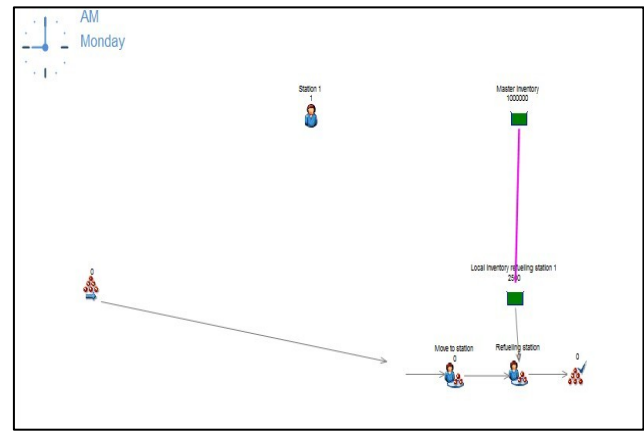


Fig. 6. The DES Model using Simul8

The tank object cannot clearly show the inventory level time by time. However, there should be a gap between the distributor, and the local gas station, which costs a delay since the transportation and the order must take some time. To create a more accurate model, we add an extra activity as a distributor and a new queue as the local station's inventory (Fig.7).

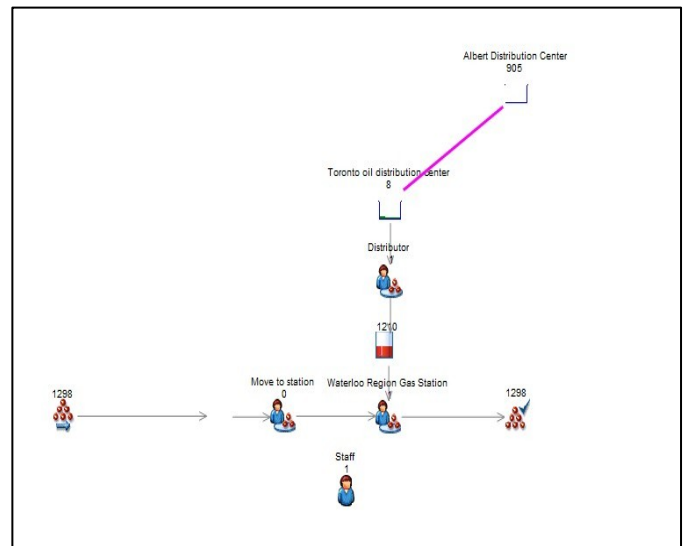
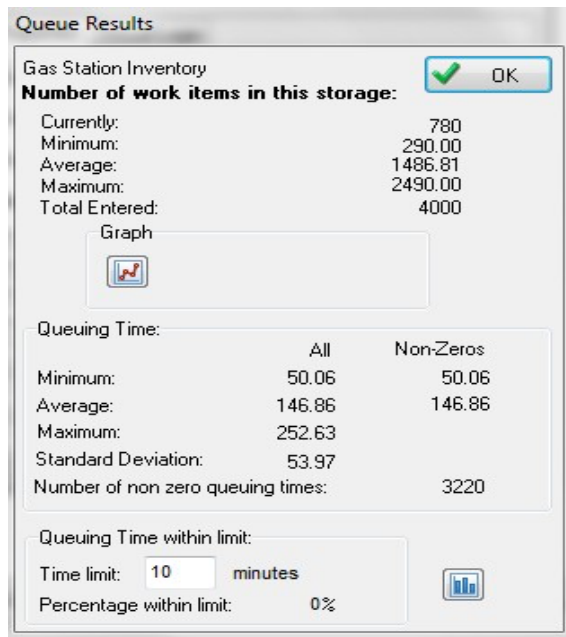


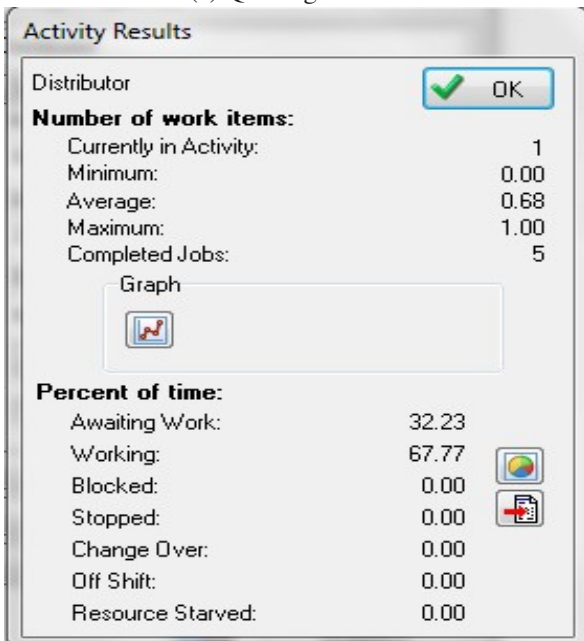
Fig. 7. The DES Model using Simul8 (Modified)

VI. PILOT RUNS

We input the distribution of the inter-arrival time of the car into the start point and the service time of the single machine into the Waterloo Region Gas Station WorkCentre. However, after a short time running, a serious problem happened: an unexpected long queue outside the gas station. We noticed at least eight servers at a single station in real life. Thus, the working rate is much lower in our model. We then refine the distribution based on the average servers each station has (8), adjust the parameters, and decide the distributor's batch size to be 1000 based on the normal capacity of a truck. To compare the cost between different policies based on different reorder levels, we calculated the average queuing size in the local tank queue object and the total times of order resulting in the Toronto distributor, as shown in Fig.8.



(a) Queuing Results



(b) Activity Results

Fig. 8. The Queuing and Activity Results

VII. PRODUCTION RUNS

We noticed that we cannot use one server to adequately represent the multiple servers' utilization since the working rate of the single equilibrium server is always lower or equal to the multiple sets. As the multiple servers' optimal policy should have a larger reorder quantity, the final model is designed, as illustrated in Fig.9. To calculate the total cost of a gas station with six servers, we used the following variables: pipeline speed, reorder point and reorder quantity. Thus, our goal is to compare the total cost based on different variable levels and find the lowest cost policy. We choose the decision variable levels. (R, Q, Speed) (Table 2).

Table 2: Decision Variable Levels

R	Q	Speed
500,1000 /L	10000,15000,20000,25000,30000,35000,40000 /L	1, 5 m/s

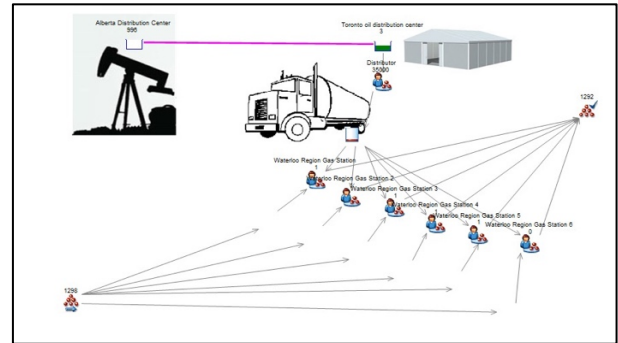


Fig. 9. The Complete Simulated Supply-Chain Model

We run a single month on the modified model to determine which policy level is optimal, i.e., has the lowest cost. For example, for reorder point 500L, reorder quantity 10000L with pipeline speed 1m/s, the inventory quantity graph is shown in Fig.10. Looks like a typical EOQ model. The statistics of the queue and the distributor are illustrated in Fig.11.

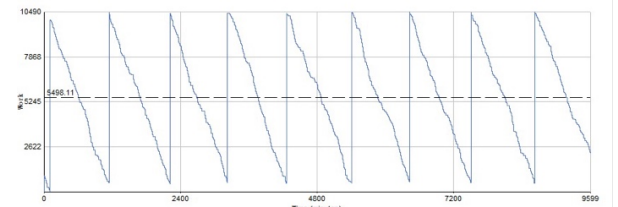
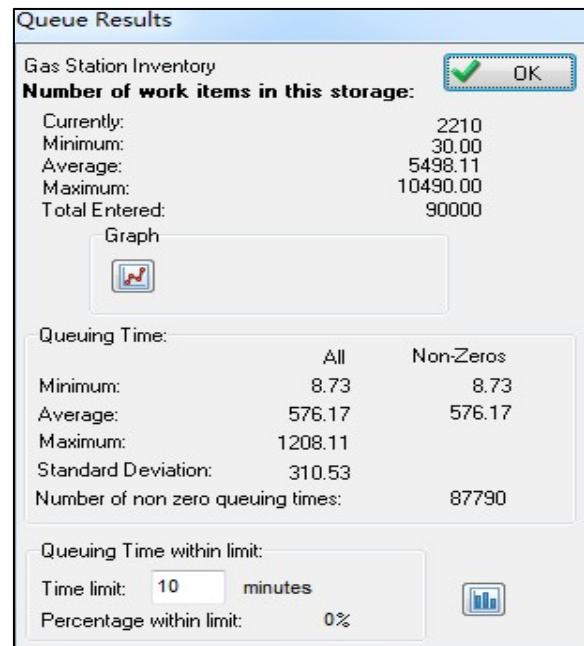
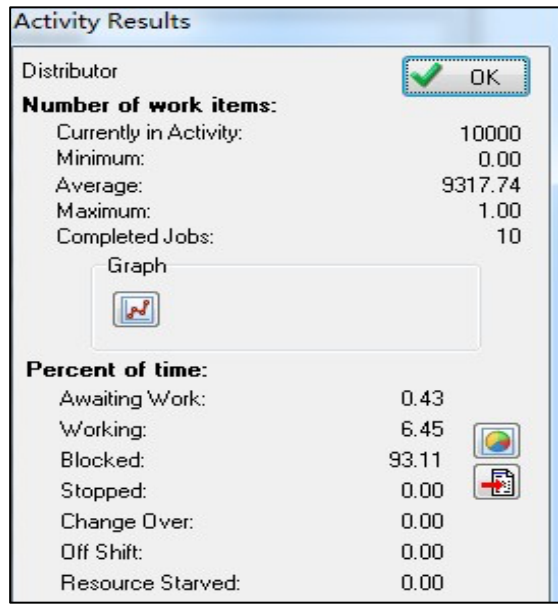


Fig. 10. The Inventory Quantity Graph



(a)



(b)

Fig. 11. The Queue and Activities (Distributor)

The status of each queue in the simulation period is shown in Fig.12. The green part of the distributor means the number of orders made during the whole period, and the green part of the waterloo gas station represents the fraction of stock out in the period. The tank volume during the simulation is illustrated in Fig.13. We can see that the minimum of the tank is never 0, which means we never faced the stock out of the simulation during the whole period.

Name	Completed Jobs	Current State	Utilization (%)
Distributor	4	Blocked	
Waterloo Region Gas Station	1298	Waiting	

Fig. 12. The Queue and Activities (Distributor)

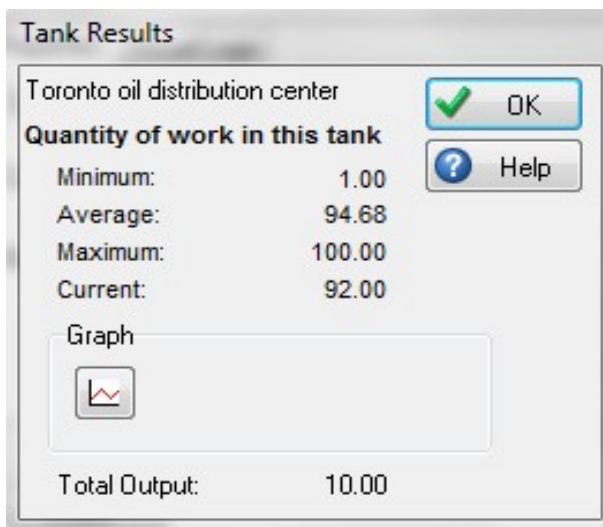


Fig. 13. The Tank Volume

Ten orders are made in a single month on simulation, so the cost is Pipeline cost+Holding rate*Average of inventory+number of orders*order cost. The cost of different policies is shown below: (for speed = 1m/s).

Table 3: Optimal Policy Levels

Quantity	Reorder	
	500	1000
10000	42656.73918	42662.93509
15000	42436.38504	42441.60887
20000	42388.02882	42393.73799
25000	42334.07632	42342.47215
30000	42273.40715	42277.89205
35000	42317.13049	42323.3408
40000	42342.6808	42349.28826

Table 4: Optimal Policy Levels

Quantity	Reorder	
	500	1000
10000	42738.15635	43329.86965
15000	42520.19496	42528.14132
20000	42384.58025	42390.78049
25000	42330.37035	42336.70882
30000	42356.11782	42365.76382
35000	42308.85685	42316.15549
40000	42346.06166	42352.15465

In the chart shown in Table 3, we see that the optimal policy for speed 1m/s is (500, 30000). Thus, the best order policy for our problem is (1m/s, 500L, 30000L). Then we run the full model to see if our prediction is correct. We use the same methodology as the simplified one, and after all the calculations, the final result is shown below in Table 4. We can see that the best reorder quantity now is 35000, greater than the simplified model and as we predicted initially. However, there is another issue if all the server is fully utilized at this current point. We checked the utilization graph of each working server, as shown in Fig 14. We can see that there is a substantial portion for each server to wait. In our model, we do not consider the maintenance fee of those servers, so more servers lead to a shorter waiting time for the cars but lots of idle time for the servers. Once the maintenance fee is considered, we should adjust current server numbers since it could cost a lot of loss because of its inefficiency. Also, in the simulation, we didn't see any stock-out situation happening in the Toronto oil distribution center. Thus, the pipeline speed doesn't need to increase since increasing the speed will increase the total cost but not improve the work rate. Overall, the best policy for the current problem is (1m/s, 500L, and 35000L). The result follows the classic EOQ assumption, a tradeoff between holding the asset and ordering goods. The best policy is using pipeline speed 1m/s with reorder point 500L and reorder quantity 35000L. The result differs slightly from what we have

proposed at the beginning. Some variables do not affect the result, like the pipeline speed and the tank capacity. In general, we can achieve a suboptimal solution as to the original goal of our project. If we could use a more extended time window, the model could be much more realistic as we can introduce more real-life factors such as idle cost and other cities. What is vital to the simulation is that the simulation should represent reality as close as possible.

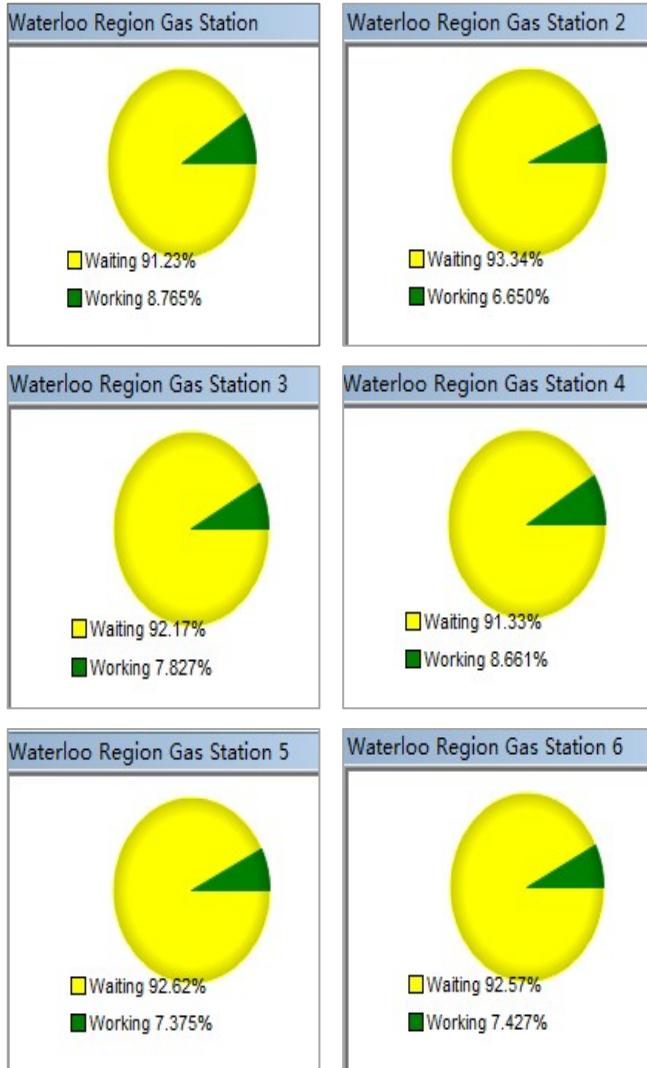


Fig. 14. The Utilization of working servers.

VIII. CONCLUSION AND FUTURE DIRECTIONS

The main objective of this paper is to use discrete event simulation to find the optimal combination of transportation methods and the optimal inventory policy. In this paper, we will focus on the midstream and downstream sectors in the oil industry. Various pilot and production runs have been implemented. Using a conceptual supply chain model and real data collection, experimentally, the best policy is using pipeline speed 1m/s with reorder point 500L and reorder quantity 35000L. Future direction involves the investigation of a longer time window and different assumptions about probability distributions. Combining machine learning [13]-[15] in

transportation systems, including supervised methods [16][17] and deep learning methods [18]-[22] with simulation to increase accuracy is a future extension of the current research.

REFERENCES

- [1] Canadian Energy Pipeline Association (CEPA), 2022. Liquids Pipelines Maps. Accessed March, 2022 <https://www.cepa.com/map/pdf/ng-cepa2022.pdf>.
- [2] <https://ctrf.ca/>. Accessed on Feb 2022.
- [3] <http://www.drivershortage.ca/wp-content/uploads/2017/01/Update-2016-%E2%80%93-Truck-driver-Supply-and-Demand-Gap-CPCS-Final-Report.pdf>
- [4] http://www.bctrucking.com/sites/default/files/tc_2008_operating_costs_of_trucks_in_canada_in_2007.pdf
- [5] <https://www150.statcan.gc.ca/n1/en/subjects/Transportation>
- [6] <https://www.simul8.com/>
- [7] Fengli Zhang, Dana M. Johnson, &Mark A. Johnson. (2012). Development of a simulation model of biomass supply chain for biofuel production, *Journal of Renewable Energy* 44 (2012) 380-391
- [8] Statistic Canada, (2021). Pipeline data on energy
- [9] Transport Canada, (2020). Operating Costs of Trucks in Canada.
- [10] Lai, J., Che, L., & Kashef, R. (2021, September). Bottleneck Analysis in JFK Using Discrete Event Simulation: An Airport Queuing Model. In *2021 IEEE International Smart Cities Conference (ISC2)* (pp. 1-7). IEEE.
- [11] Aldhubaib, H. A., & Kashef, R. (2020). Optimizing the utilization rate for electric power generation systems: A discrete-event simulation model. *IEEE Access*, 8, 82078-82084.
- [12] N. Attanayake, R. F. Kashef and T. Andrea, "A simulation model for a continuous review inventory policy for healthcare systems," *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2014, pp. 1-6, doi: 10.1109/CCECE.2014.6901005.
- [13] Karami, Z., & Kashef, R. (2020). Smart transportation planning: Data, models, and algorithms. *Transportation Engineering*, 2, 100013.
- [14] Hass, G., Simon, P., & Kashef, R. (2020, December). Business Applications for Current Developments in Big Data Clustering: An Overview. In *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 195-199). IEEE.
- [15] Gurnani, P., Hariani, D., Kalani, K., Mirchandani, P., & CS, L. (2022). Inventory Optimization Using Machine Learning Algorithms. In *Data Intelligence and Cognitive Informatics* (pp. 531-541). Springer, Singapore.
- [16] Kochak, A., & Sharma, S. (2015). Demand forecasting using neural network for supply chain management. *International journal of mechanical engineering and robotics research*, 4(1), 96-104.
- [17] Kashef, R. (2021). A boosted SVM classifier trained by incremental learning and decremental unlearning approach. *Expert Systems with Applications*, 167, 114154.
- [18] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11-26.
- [19] Zhang, Y., & Gao, J. (2017, November). Assessing the performance of deep learning algorithms for newsvendor problem. In *International Conference on Neural Information Processing* (pp. 912-921). Springer, Cham.
- [20] Ibrahim, A., & Hassanien, R. (2022). Homogenous and Heterogenous Parallel Clustering: An Overview. *arXiv preprint arXiv:2202.06478*.
- [21] Tobin, T., & Kashef, R. (2020, June). Efficient Prediction of Gold Prices Using Hybrid Deep Learning. In *International Conference on Image Analysis and Recognition* (pp. 118-129). Springer, Cham.
- [22] Ebrahimian, M., & Kashef, R. (2020, December). Efficient Detection of Shilling's Attacks in Collaborative Filtering Recommendation Systems Using Deep Learning Models. In *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 460-464). IEEE.

Simulating Software Support Delays in a 24/7 Environment Using Discrete Event Simulation

Rasha Kashef
 Electrical, Computer, and Biomedical Engineering
 Toronto Metropolitan University
 Toronto, CANADA
rkashef@ryerson.ca

Shabbir Mirza
 Channel Technology Solutions
 TD Canada Trust
 Waterloo, CANADA
shabbir.mirza@td.com

Abstract—Since there is a cost associated with providing 24/7 support, it is also common to divide user support into tiers where a group focused on usability and training of the software provides tier 1 support and another group, consisting of mainly developers, is focused on feature development, software bug fixing, and related issues work separately in a different tier. This paper focuses on software organizations where the developer tier operates from just one part of the world under a single time zone. This group is not available on a 24/7 basis, but the Tier 1 group works in different regions – and hence time zones - throughout the world to deal with user problems and questions. Users can be internal or external, depending on the purpose of the software. This paper discusses only internal users (developers, users and help desk employees who work for the same company) and the delays in support when dealing with these users who may be working in different regions and/or time zones. Tier 1 support is available around the clock due to hiring help desk resources in different locations under different time zones. If an issue cannot be resolved, the Tier 1 workers escalate the issue to the next tier (the developers), who work regular hours in one region under one time zone. This paper uses discrete event simulation modelling to provide insights for a company with a similar structure to better understand possible software support delays. These delays can be reduced, and better overall support can be provided in the future.

Keywords—Software Support, Simulation, DES, Performance.

I. INTRODUCTION

Customer support has become an integral part of software organizations. Companies tend to divide the support work into tiers to provide software support more effectively. It is common to place developers in a later tier so that tiers can handle user training and usability issues with non-developers. Anytime the support needs to be escalated to the developer tier and they are not available, a ticket is created so that the developers tier can handle the issue as soon as they are available at work [1]. With this concept, though it is ideal to have all types of resources available 24/7, for various reasons, that situation becomes challenging, especially when a tier of support is operational only under one time zone for a set number of hours. If the customer support or help desk (non-developer) tier is available around the clock to attend to the calls, the only concern would be the delay involved when issues are escalated to the development tier. In this paper, to learn more about service delays in software support, a model is created to simulate a company structure with the two-tier backing: Tier 1 – help desk tier, hired to provide 24/7 support at different branches around the world and Tier 2 - developer tier that operates only from North America under the

one-time zone. For simplicity, the model does not include overtime hours or on-call support options and excludes daylight savings time adjustments. Users of the software can be internal or external, depending on the purpose of the software. In most cases, external users do not interact directly with the developer tier; thus, we will focus on internal users alone. These users work for the same company as the help desk and developers, and they utilize the software to help run various processes or tasks within the company from different parts of the world. When they encounter an issue using the software, they may call the help desk available 24/7. If a software feature modification or alteration is required, the help desk forwards the requesting user to developers if they are available; otherwise, the user request goes to a queue waiting for the developers to be available to provide a resolution. This paper uses discrete event simulation (DES) modelling to provide insights for a company with a similar structure to better understand possible software support delays. These delays can be reduced, and better overall support can be provided in the future. We look at the latter queue to learn about the different parameters that influence the number of requests sitting in the queue and the overall time the requests spend in the system before the requests are resolved. These tasks and goals are achieved using discrete event simulation. This paper is structured as follows: related work is discussed in section 2. The conceptual model is presented in section 3. Section 4 discusses the validation and verification methods. Section 5 shows the experimental results along with pilot runs. The paper is concluded in section 6 with future directions.

II. RELATED WORK

As providing tech support on a 24/7 basis is becoming increasingly important to software companies, it makes sense that there is some related research on global support and collaboration and even examples of actual companies providing tiered service to businesses [1]. Most of the literature on this topic covers the challenges involved with support and collaboration in a global business environment. Authors in [2] examined the challenges of working with a virtual team versus working with a team in the same physical location. The challenges that affected relationships and communication involved working with a virtual team included different time zones, cultural differences, requirements creep, management leadership style, and the lack of clear responsibilities and roles. In [3], the authors studied issues that may arise when individuals work with each within a virtual team. This study interviewed sixty-five employees, including managers, senior executives,

team leaders, and team members. Some of the challenges discovered included building a sense of cohesiveness, trust, and specific team identity. Another issue that came up during this study is the issue of overcoming a sense of isolation that may arise from working in a virtual environment. A study of trends and practices for the Australian business sector [4] offers insight into the contact center offshore industry – a service often utilized by software companies. More specifically, they suggest that companies are experiencing cost savings and more business flexibility by choosing to employ offshore call center support. Some of the challenges faced with outsourcing these endeavors include employee recruitment and retention. In [5], they examined the additional benefits and challenges of outsourcing work to other countries. To arrive at their answers, a variety of companies were surveyed. With respect to the challenges, data security, intellectual property rights, and political risks were the top concerns. That said, 68% of the survey respondents were satisfied with their outsourcing arrangements, and just over 25% of the respondents were also unsatisfied or very unsatisfied with the arrangement. As most unsatisfied companies had been outsourcing work for less than a year, they suggested outsourcing is more of a long-term investment. For actual companies offering tiered tech support to other businesses, one example is a company called Harmonic [6]. Users with issues or questions contact Harmonic’s centralized Technical Assistance Center (TAC) for assistance. The staff employed at the TAC are located in multiple locations throughout the world, including the United States, Israel, Hong Kong, Singapore, and England. Available on a 24x7 basis, the TAC provides multiple-level tech support to users. Specifically, users first talk to a customer service agent and then are transferred to a technical support engineer. Suppose the initial issue cannot be resolved with the engineer. In that case, the problem is assigned to a Level 2 tech support engineer and specialists for further investigation (located throughout the world) and an eventual resolution to the problem. ExactTarget [7] is another example of a company that offers 24/7 global technical support to businesses. Once again, this company boasts of an international presence with offices located in Canada, the United States, England, Germany, France, Sweden, Singapore, Australia, and South America. With its standard support package, users call a Global Support Center and then relay their issues to one of the company’s technical support specialists. The issue is resolved or transferred to a Tier 2 advanced-level technician for resolution. ExactTarget also offers two other levels of technical support. The premium support package provides direct access to advanced-level technical support specialists. In contrast, the platinum support package offers specialized support from tech support experts who are extremely familiar with a business’ tech and business challenges [8]. Recently, discrete event simulation (DES) [9]-[11] shows a great protentional in providing insights for companies to better understand possible outcomes.

III. THE CONCEPTUAL MODEL

To design a conceptual model of the intended organization, the following four regions are considered: North America, Latin America, Asia and Europe. Four different cities are picked randomly under these four regions – Toronto for North America, London for Europe, Rio de Janeiro for Latin America and Singapore for Asia. A software company with branches in all

four regions hires a resource team for the developer tier in North America alone for the model in hand. Employees in North America are assumed to be available from 8 am – 6 pm EST. From the working hours in North America (Toronto), the associated times are highlighted in Fig.1 [1].

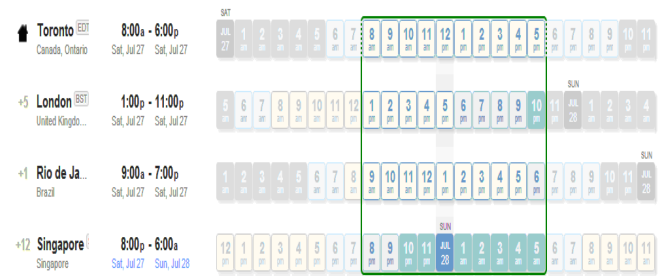


Fig. 1. Working hours in Toronto and how it overlaps with the rest [1]

North America will be the main reference point for this model when calculating hours covered around the clock for support or resource availability. This covers 8 am – 6 pm EST North America time, and both the help desk and developer tier are available to cover ten hours a day. The slider is moved towards the right to display the working hours in Europe (London), and the corresponding hours in North America (Toronto) are also highlighted in Fig.2.

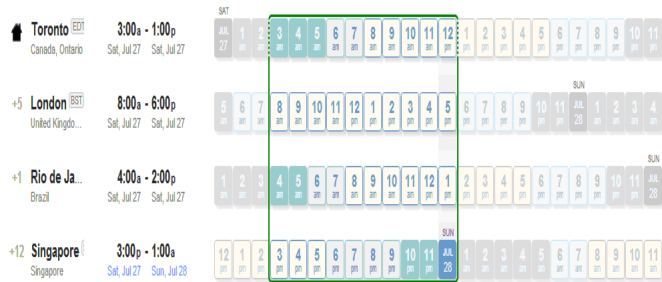


Fig. 2. Working hours in London and how it overlaps with the rest [1]

Here, it is visible that the European hours cover 3 am – 1 pm EST of North American time. During this time, the help desk team in Europe is available, and the help desk and development tier in North America. However, it must be noted that the North America support is only available for five hours – namely from 8 am – 1 pm EST. In other words, Europe can cover five hours of daily tier 1 support when North America’s support is not available. Similarly, the available working hours in Latin America (Rio de Janeiro) tell us that the region covers only one hour when North America is unavailable. From Fig.3, this region is available from 7 am – to 5 pm North American time. The only hour they can cover for North America overlaps with the Europe region (12 pm – 1 pm in London). Therefore, hiring resources in Latin America to provide 24/7 support is less efficient for this model when the cost is not included. Asia’s last region can cover a generous ten hours of the day for North America, as shown in Fig.4. Asia’s (Singapore) regular working hours fall under the 8 pm – 6 am EST time slot in North America, and Singapore’s help desk support is available from this region. As one may assume, this region is very important for any North American-based software company since this region

overlaps with the time when the North America support is unavailable.

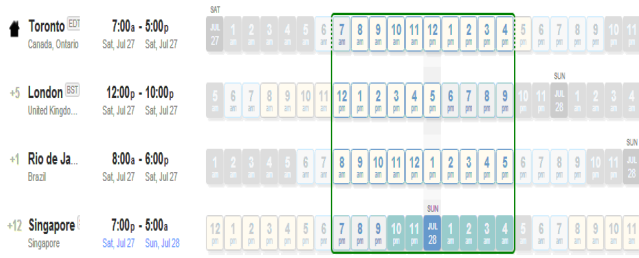


Fig. 3. Working hours in Rio de Janeiro and how it overlaps with the rest [1]

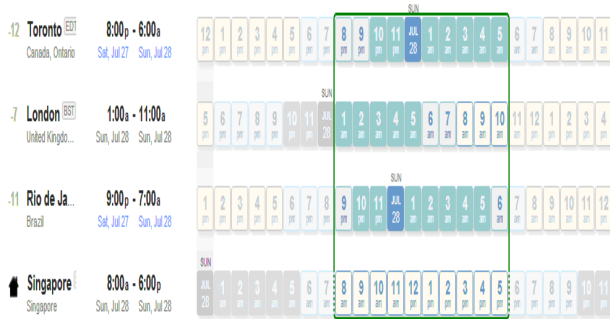


Fig. 4. Working hours in Singapore and how it overlaps with the rest [1]

There is an additional cost benefit to operate services from this region, but that topic is left out for future extension of this model. Another important observation is that the Asia (Singapore) region overlaps with three hours of work within the Europe region (London). After looking at the four regions with four different time zones, it is obvious that the main two regions that will cover twenty hours of support a day are North America (Toronto) and Asia (Singapore) in this model. The requirement for Latin America support is minimal in our model since its working hours are shared with at least one other region. In this example, Europe (London) can provide two additional hours of support or the 6th and 7th hour of the day. However, this situation leaves two hours of “no support” - the 18th and 19th hour of the day. For simplicity reasons, overtime working hours, on-call support and daylight saving time adjustments are not included in this model. Thus far, it is evident that a software organization should allocate a significant portion of their tier 1 support resources in North America and Asia. We randomly chose that 45% of the resource team should reside in North America, 45% in Asia, 8% in Europe, and 2% in Latin America for our model. The tier 2 support team (developer tier) belongs to the North American region and is available only from 8 am – 6 pm EST. In this paper, we created four work shifts for the model, namely a North American, Latin American, Asian and European work shift, as shown in Fig. 5. Specifically, the shifts occur between 8 am – 6pm, 7 am – 5pm, 8 pm – 6 am, and 3 am – 1 pm EST. With this structure, support is not available for two hours a day between 6 pm and 8 pm EST. Though support will not be available for those 2 hours, since users may try to contact the help desk during those hours, a fake shift called “No support shift” was created to cover those two hours. Software users are distributed among the four shifts and the “no support shift.”

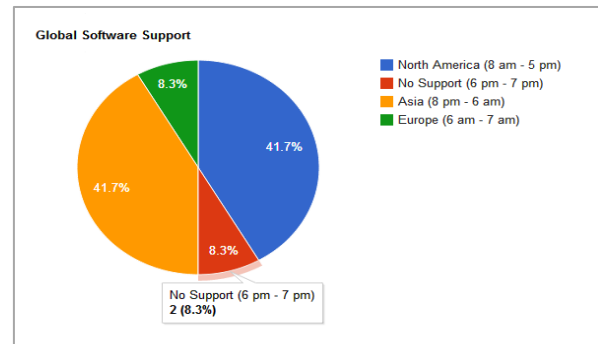


Fig. 5. Shifts used for the model and hours covered by each shift

At random, 15% of the company employees are placed in tier 2, consisting of software developers working the North American shift. Additionally, 20% of the employees sit at help desks for tier 1 support distributed among the available regions. The rest of the 65% of employees available in the company will be considered the end-users of the software. Tier 1 help desk helps these users with software usability, training and frequently asked questions. Unresolved issues are escalated to “Tier 2 Queue” for the developers to assist further, and these developers are available only during the North American shift. This can be depicted in Fig.6, where the requests are made by “users” at the “Tier 1 Queue” and served by the help desk at tier 1, the group of people working from four different regions on different shifts.

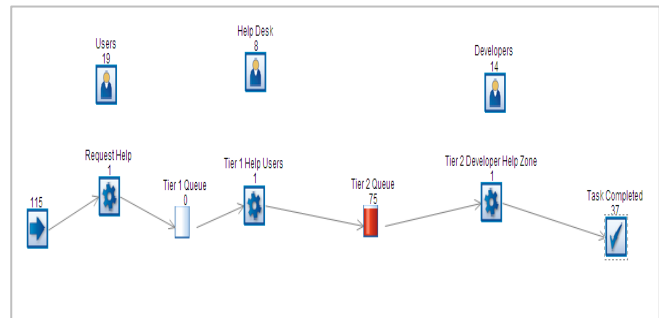


Fig. 6. Base model of software support in Simu8

If the tier 1 help desk can resolve most usability issues, tier 2 developers should encounter only software bugs or feature requests. Hence, only a comparatively smaller portion of the requests should go to tier 2 in an ideal world with stable software. Introducing some exceptions, such as routing only a percentage of the requests to the “Tier 2 queue”, makes this model more interesting. The model can be simulated with varying numbers for the percentages to test scenarios that closely match real-world situations, as shown in Fig.7.

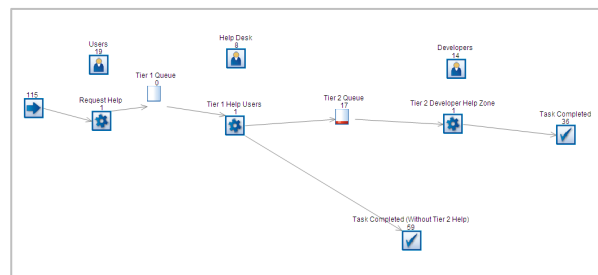


Fig. 7. Software support model in Simu8 with tier 1 and tier 2 layers

The layout of the rest of the company employees and software users is left out purposefully since the focus is on tier 1 and tier 2 support teams. Assumptions are made that they are equally distributed over the available regions. The complete process that this model follows is described in Fig.8. A more complex version of the model is shown in Fig.9, where two additional termination points are added should the users leave the queue in the case of a longer wait. Our base model can be extended to include the extra termination points where users leave the system after a certain amount of waiting time in the queue. A concept model is shown in Fig.10. To have a better representation of how resources are distributed and utilized over the four regions, extra nodes are created.

IV. VALIDATION AND VERIFICATION

Different percentages of requests completed without tier 2 help were implemented to validate this model. As expected, as more requests need “tier 2 support”, there is a longer waiting time as the tier 2 queue becomes busy. If the overall time to process each request at tier 2 is improved, fewer items will be waiting in the tier 2 queue. Service time at the help desk for each request is set to 20 minutes, and at the developer tier, it is set to 60 minutes - assuming it takes more time to resolve an escalated software issue. This processing time can have a significant impact on the overall model. It was also assumed that a user makes a request every 15 minutes on average. If users make requests more frequently, tier 1 and tier 2 queues start to experience hold-ups, as shown in Fig.11. Thus, if the service time at tier 1 is less than the inter-arrival time, then the tier 1 queue is less busy.

V. PRODUCTION RUNS WITH SIMUL8

First, the model was run on a company with 100 employees. Specifically, 65 users work in different time zones (different shifts for this case), 15 developers work in tier 2, and 20 people work at the help desk to provide tier 1 support. Tier 1 support is divided into four shifts to mimic working under different time zones. Nine resources are allocated to each North American and Asian shift, two resources working at a European shift and no one on the Latin American shift. A help desk call is made every 15 minutes on average, and each call servicing time is 20 minutes on average. 70% of the requests are resolved at this tier, and the rest of the 30% are escalated to the developer tier, where it takes 60 minutes on average to service each request. In this model, the clock is set to run for seven days a week around the clock and for an entire week’s duration. The result is shown in Fig.12. The developer tier queue is larger after running this model for a week (Fig.13). Further, although only 30% of the requests are going to tier 2, it takes significantly longer to resolve tier 2 issues since tier 2 is unavailable around the clock. On the other hand, requests resolved at tier 1 via the help desk show a significant improvement compared to the developer tier, as shown in Fig. 14. To see overall improvement, a software company can consider either extending developer help to a different region (so that they are available for more hours) or the company can consider reducing the service time at the tier 2 level. Plotting the queue size against a different percentage of requests routing to Tier 2 gives a clear picture of how it can impact an organization. Later the model is expanded to illustrate apparent differentiation between the regions by introducing

resource and task nodes corresponding to each region, as shown in Fig.15. Running this expanded model in Simul8 shows how the resources are utilized and along which paths. In reality, some users tend to leave the system after a certain time. Our current model can be extended to include these users who can leave when the expiration hits at the queues, as illustrated in Fig.16. After running both models for four weeks in Simul8, a significant improvement is visible in queuing time at the developer tier, and the average time a task spends in the system. Table 1 shows a performance comparison of these two models.

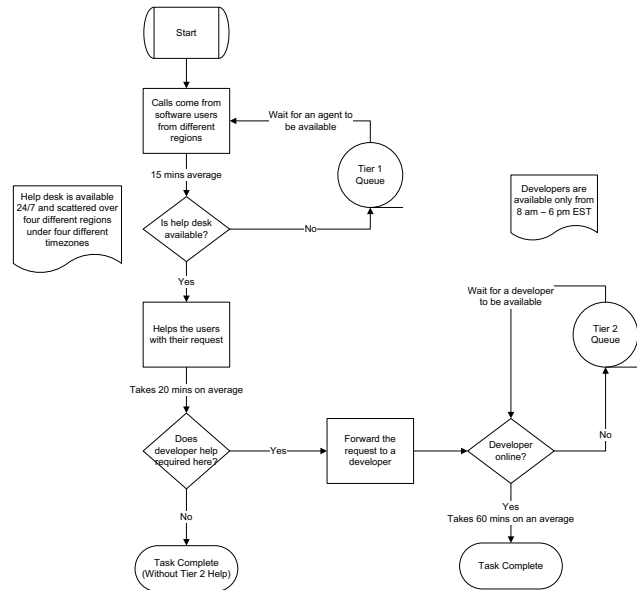


Fig. 8. Flow chart for the Software Support Model

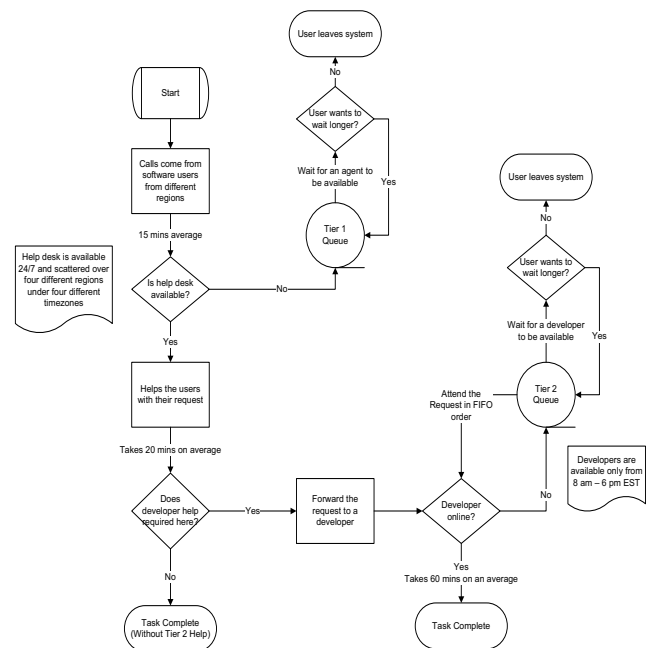


Fig. 9. Flow chart of a software support model where users can leave the queue

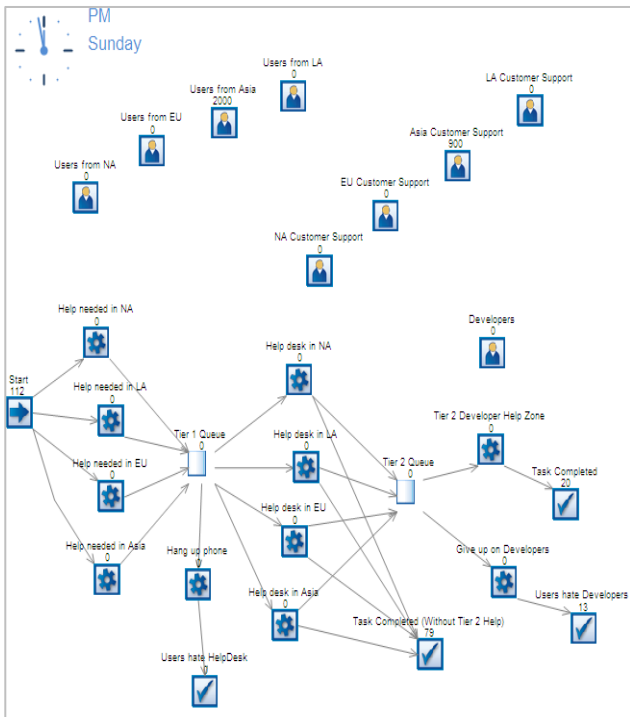


Fig. 10. Support Model with expiration at queues

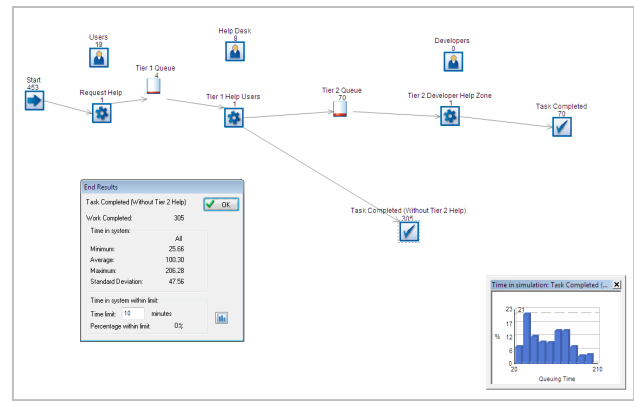


Fig. 13. Tier 1 Queue after a pilot run

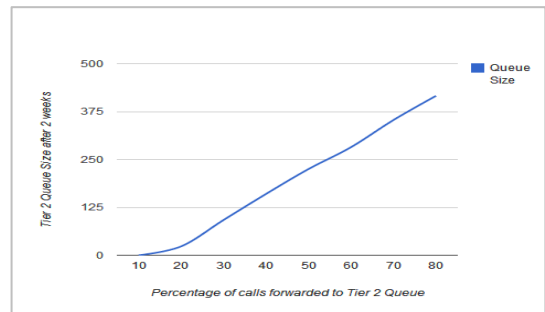


Fig. 14. Tier 2 queue is affected by load increase

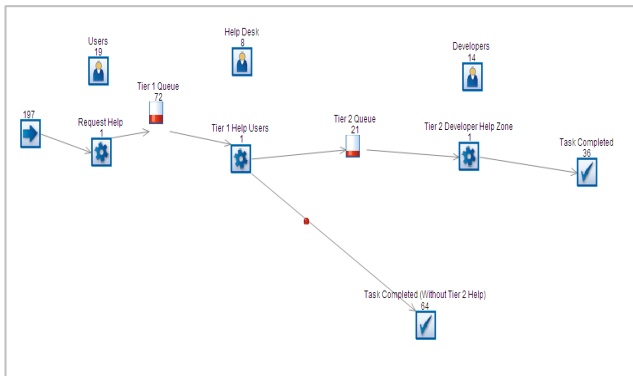


Fig. 11. Pilot run of the model in Simu8

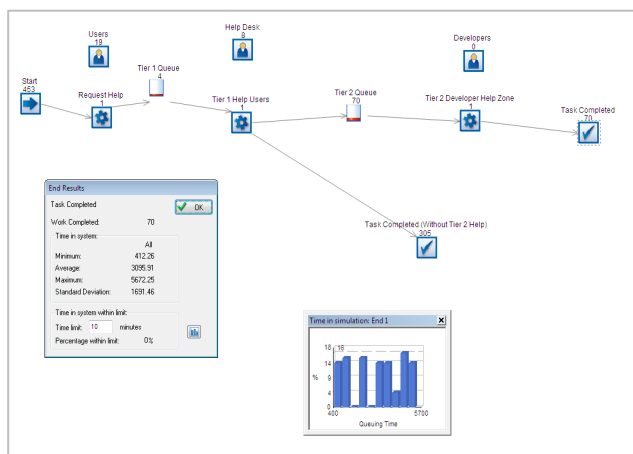


Fig. 12. Tier 2 Queue after Pilot Run

TABLE I. COMPARING THE TWO MODELS

	Model 1	Model 2
	A model with no expiration at queues	A model with expiration at queues
	Users wait until a developer is free to take care of their requests	Users may leave after a certain period of time
Average queuing time at the developer tier	260.63	64.45
Average time in the system at the developer tier	531.47	240.88
Task completed at developer tier	38	19
Average time in the system at the help desk tier	38.36	38.36
Task completed at developer tier	73	73
After running 100 trials average time in the system		
Low 95% range	455.29	222.16
Average result	465.22	223.94
High 95% range	475.22	225.71

VI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This paper develops a model that depicts a real-life software organization scenario where software users may require 24/7 support. This model can be further extended to associate cost with each resource and action performed; this can significantly affect how the company employs individuals in different regions and whether the tech support should work at tier 1 or tier 2 level. More software support tiers can be added to the model to add more complexity. The model will be useful as it will provide the

company with an advanced look at the overall efficiency and cost. If the tier 1 resource is more cost-effective than the tier 2 resource, a study can be performed to determine whether training the tier 1 resource team to perform tier 2 functions is a feasible plan. It is important to note that all the queues assumed an infinite capacity and shelf-life in this paper. This situation may be entirely different for many companies in the real world. To conclude, interacting with businesses through many channels can be challenging; however, the future of customer service can be well-planned and structured if a proper model is used to grasp a company's customer service needs ahead of time.

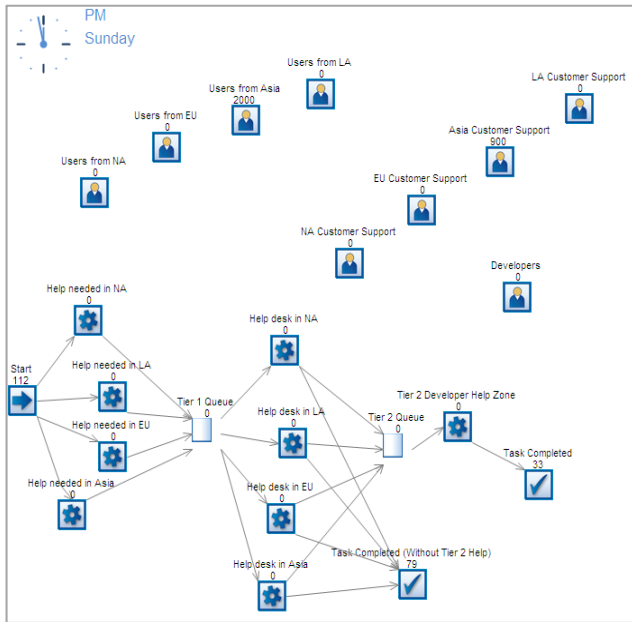


Fig. 15. The expanded model

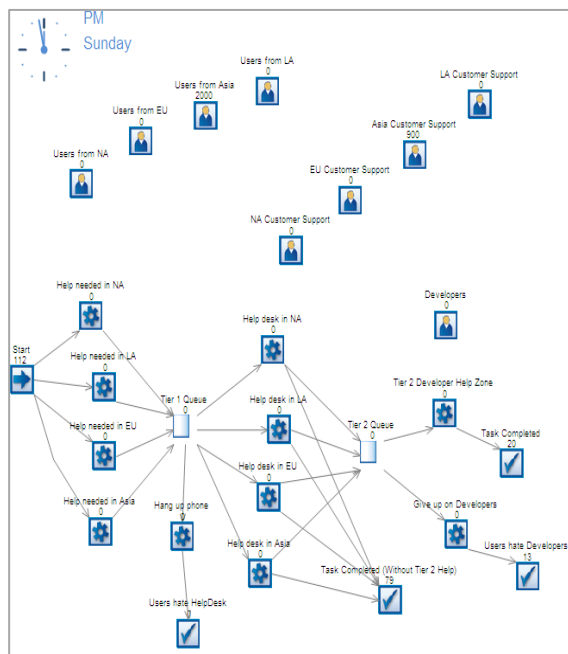


Fig. 16. Extended model with expiration at queues

Combining Ai-driven methods using machine learning-based [12]-[14], supervised learning [15][16], and deep learning [17]-[20] with DES to increase automation and accuracy is one of the main future extensions of this research.

REFERENCES

- [1] "World Time Buddy (WTB)," Retrieved on 28 July, 2020, <http://www.worldtimebuddy.com/?pl=1&lid=6167865,2643743,3451190,1880252&h=2643743>
- [2] Lee-Kelley, L., & Sankey, T. (2008), "Global virtual teams for value creation and project success: A case study", *International journal of project management*, 26(1), 51-62.
- [3] Thomas, J. O., Rankin, Y. A., & Boyette, N. (2009, November), "Self service technologies: eliminating pain points of traditional call centers", In *Proceedings of the Symposium on Computer-Human Interaction for the Management of Information Technology* (p. 9). ACM.
- [4] Alison R. Owens (2013, June), "Exploring the benefits of contact centre offshoring: a study of trends and practices for the Australian business sector", *The International Journal of Human Resource Management*.
- [5] Bajpai, N., Sachs, J. D., Arora, R., & Khurana, H. S. (2004), "Global services sourcing: Issues of cost and quality."
- [6] "Harmonic Inc", Retrieved on 28 July, 2013, from Web site: <http://www.harmonicinc.com/content/technical-support>
- [7] "ExactTarget", Retrieved on 28 July, 2013, from Web site: <http://www.exacttarget.com/services/global-support>
- [8] Kirkman, B. L., Rosen, B., Gibson, C. B., Tesluk, P. E., & McPherson, S. O. (2002). Five challenges to virtual team success: lessons from Sabre, Inc. *The Academy of Management Executive*, 16(3), 67-79.
- [9] Lai, J., Che, L., & Kashef, R. (2021, September). Bottleneck Analysis in JFK Using Discrete Event Simulation: An Airport Queuing Model. In *2021 IEEE International Smart Cities Conference (ISC2)* (pp. 1-7). IEEE.
- [10] Aldhubaib, H. A., & Kashef, R. (2020). Optimizing the utilization rate for electric power generation systems: A discrete-event simulation model. *IEEE Access*, 8, 82078-82084.
- [11] N. Attanayake, R. F. Kashef and T. Andrea, "A simulation model for a continuous review inventory policy for healthcare systems," *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2014, pp. 1-6, doi: 10.1109/CCECE.2014.6901005.
- [12] Karami, Z., & Kashef, R. (2020). Smart transportation planning: Data, models, and algorithms. *Transportation Engineering*, 2, 100013.
- [13] Hass, G., Simon, P., & Kashef, R. (2020, December). Business Applications for Current Developments in Big Data Clustering: An Overview. In *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 195-199). IEEE.
- [14] Gurnani, P., Hariani, D., Kalani, K., Mirchandani, P., & CS, L. (2022). Inventory Optimization Using Machine Learning Algorithms. In *Data Intelligence and Cognitive Informatics* (pp. 531-541). Springer.
- [15] Kochak, A., & Sharma, S. (2015). Demand forecasting using neural network for supply chain management. *International journal of mechanical engineering and robotics research*, 4(1), 96-104.
- [16] Kashef, R. (2021). A boosted SVM classifier trained by incremental learning and decremental unlearning approach. *Expert Systems with Applications*, 167, 114154.
- [17] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11-26.
- [18] Zhang, Y., & Gao, J. (2017, November). Assessing the performance of deep learning algorithms for newsvendor problem. In *International Conference on Neural Information Processing* (pp. 912-921). Springer, Cham.
- [19] Ibrahim, A., & Hassanien, R. (2022). Homogenous and Heterogenous Parallel Clustering: An Overview. *arXiv preprint arXiv:2202.06478*.
- [20] Close, L., & Kashef, R. (2020). Combining artificial immune system and clustering analysis: A stock market anomaly detection model. *Journal of Intelligent Learning Systems and Applications*, 12(04), 83.

Intelligent Feature Selection on Multivariate Dataset using Advanced Data Profiling

¹ **Abstract**—The differential diagnosis of diseases which share similar clinical features is a real and difficult problem in medicine. This paper demonstrates the use of data mining (DM) techniques to augment standard data profiling methods and establishes an efficient approach for an intelligent feature selection method for disease that share similar features. The results from experiments returned show that by using DM techniques to select features as an additional layer on top of data profiling, there is considerable improvement in the performance of the prediction model built to predict a disease such as “Psoriasis”. A brief comparison between features selected by existing mining tools such as Weka and the proposed approach with respect to predictive accuracy is recorded in this paper. The proposed algorithm works as a promising tool for assisting diagnosis of disease like erythemato-squamous diseases, where the symptoms are overlapping. By combining data cleansing and knowledge discovery techniques, the algorithm aims to be “agnostic” and can be used on a wide variety of data standards with variable data quality.

I. INTRODUCTION

Machine learning models are built to capture relationships in an n-dataset (dataset of size n) between its n-1 features and the (nth) target feature in the dataset. Since such models are highly data-driven, their accuracy relies on how “good” the data is. “Goodness” of any dataset can be determined by understanding the data and preparing it so that it can be deemed “good”. Typically, repositories such as Data Lakes [15] store data in its raw form, which is eventually (pre)processed before it is used in a data mining model. To understand any given data in a dataset requires gathering information about that data, commonly known as metadata. Type of metadata depends on the application for which the data mining model is being designed. As an example, let us consider an application that requires educational data, such as student information, the courses they have completed and grades achieved in these courses. Metadata in this example can be as simple as the student’s name and age stored in its raw form, or it could require some calculation such as average age of students in this dataset. If described and managed well by big data repositories such as data lakes, such metadata can be exploited by analysts to discover underlying relationships between different features in a dataset, and thereby allow data mining models to be better informed about the dataset. Data profiling refers to this structured activity of creating small but informative summaries of a database [8]. Data profiling uses scientific methods to

explore, understand and collect statistics on raw data of a given dataset for detecting statistical distributions and structural patterns.

In a typical real-world scenario, data is integrated from various different sources, and this makes it challenging to provide consistent, clean and accurate data for machine learning models. Data profiling can be used as a pre-processing step in the process of data mining to evaluate datasets for consistency, uniqueness and quality of data. It is a well known fact that data mining models spend 70% of their time in the pre-processing step which clearly needs to be improved [15]. This paper proposes a data agnostic algorithm for data profiling. In this algorithm, a full understanding of the entire data set is not required. The algorithm uses data mining techniques to understand the data (we call it as intelligent data profiling) and use that to select features that could prove to have a higher impact on the accuracy of the final model that is being built. One may argue that the high cost of computational complexity of profiling large datasets may deter data enthusiasts from using it. But this can be overcome by leveraging massively parallel clusters and will not be addressed in this project. Our research aims to build an intelligent data profiling algorithm that evaluates and analyzes different attributes or features in the dataset, with an overall objective of selecting those features that are of good quality and that improve the accuracy of the predictive model built using these features.

This paper is organized in the following way: Section II explains the proposed methodology. Sections III illustrates the evaluation of the proposed algorithm. A literature review of the related work and their limitations is discussed in section IV. Section V presents the conclusions and proposes future work.

II. PROPOSED METHODOLOGY

This section presents the dataset used and proposed methodology for intelligent data profiling and feature selection.

A. The Dataset

The dataset used in this research is on dermatology and is taken from an open-source repository [4]. It includes 34 clinical and histopathological features that have been transformed in a variety of ways to suit this research.

¹978-1-6654-8684-2/22/\$31.00 ©2022 IEEE

1) *Rationale for using the dataset:* Diagnosing skin diseases such as 'Psoriasis' and 'lichen planus' is a real problem in dermatology [5]. Firstly, they all share the clinical features with very little differences. Secondly, there are too many features, some of them do not have enough values or have values that are not meaningful. Some diseases have overlapping features from another disease at the beginning stage and may have the characteristic features at the subsequent stages. In terms of data, the significance of the presence or absence of a feature in the dataset makes the analysis challenging. For example, given the features, it is difficult to differentiate between 'Psoriasis' or 'Lichen Planus' as they have overlapping features (for example, itching is a common feature in both). The proposed method intelligently selects only those features that have a significant impact on the accuracy of the predictive model that predicts the disease.

2) *Notation used :*

1) The following notation is used:

D = dataset

2) Each instance n of D is represented as a triplet:

$$\forall_1^n i, (ca_i, ha_i, ta_i)$$

where ca represents clinical attributes, ha represents histopathological attributes and ta represents the target attribute. $|x|$ indicates the cardinality or number of instances of x

3) D^T – dataset after transformation of attributes

4) $D_{profiled}^T$ - dataset after data profiling of the attributes

5) $D_{reduced}^T$ - optimal set of attributes used for prediction

B. *Transformation of selected features in dataset D*

In general, preprocessing is used to transform raw input data into appropriate formats for subsequent mining [14]. Real-world data is often inconsistent, has missing values and may have data that has errors. Data preprocessing is a technique that is used to resolve such issues.

The transformations on different attributes of D and the rationale for doing so are listed below.

1. The target attribute (ta) in the original dataset [4] has 6 classes. For simplicity, we transform ta to a binary class code problem, i.e. to 2 classes. We aim to predict if a person with several given symptoms (represented by a vector of clinical and histopathological attributes) has psoriasis (class code = 1) or not (class code = 2).

2. Attribute age is a clinical attribute (ca) that is continuous. In order to see the impact of age categories on the outcome (i.e. presence of psoriasis), we chose to discretize age into 5 different bins. This decision was a result of a few experiments with attribute age. Binning is a method to group a number of continuous values into a smaller number of "bins". This research creates bins using equal-width unsupervised discretization method (except for the last bin that is of a larger size than others) [10]. The bins created using this algorithm and their distribution with the target attribute are as follows:

Bin	Interval
1	[< 12.5]
2	[12.5 - 25)
3	[25 - 37.5)
4	[27.5 - 50)
5	[>= 50]

The third bin that caters to the age group 25 - 37.5 has the maximum count of patients among all bins, whereas age group less than 12.5 has the least. The graph in figure 1 shows the distribution of all age groups.

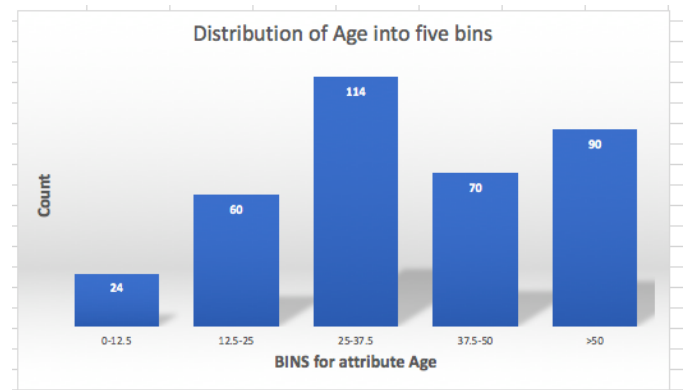


Figure 1. Discretization of Age into bins

3. All other clinical (ca) and histopathological (ha) attributes in the original dataset D can have any of the four values (0, 1, 2 or 3, where 0 indicates the absence of a symptom, 1, 2, 3 indicate the intensity of the symptom present, with 3 representing the highest intensity and 1 the lowest). All these attributes can be categorized as asymmetric attributes.

An asymmetric attribute is defined as an attribute in which the presence of one of the values (e.g. 1) is regarded as more significant than the other (e.g. 0), as opposed to a symmetric attribute in which the value 1 (or higher indicating presence of the attribute) is considered equally significant as its absence (0) [14]. For example, erythema in D is taken to be an asymmetric attribute in this research, where a value of 1 or higher indicates the presence of that symptom and 0 indicates its absence. In an asymmetric attribute, a 1-1 match of erythema in two rows is significant, whereas a 0-0 match has no significance (since 0 implies that the symptom is not present) and is ignored in the proposed method. On the contrary, an example of a symmetric attribute is gender, where 1 represents female and 0 represents male, then a 1-1 match (indicating a female-female match) is as significant as a 0-0 match (indicating a male-male match).

After the above transformation, our dataset now has:

- 11 clinical attributes: asymmetric
- 22 histopathological attributes: asymmetric
- 1 family income: binary symmetric attribute
- 4 binary asymmetric attributes: unique bin for age category

It is worth mentioning here that id numbers of patients are removed from the database in the original dataset D.

C. The proposed algorithm - Predicting Using Intelligent Feature Selection (PIFS)

Algorithm 1 lists the core steps of the proposed algorithm called PIFS (Predicting using Intelligent Feature Selection). Step 1 transforms the original dataset D to D^T . Section B defines the transformation on some of the attributes such as age and the target attribute (ta) and the rationale behind those transformations.

Step 2 performs the data profiling step to convert D^T to $D^T_{profiled}$. We did single-column profiling tasks in terms of the number of rows and uniqueness of the values [1]. Those features or columns that have 80% or more values as 0 (0 indicates absence of that symptom) were eliminated and not included in the new set $D^T_{profiled}$. At the end, the attributes selected for the next step were reduced to a new dataset ($D^T_{reduced}$).

Dataset $D^T_{profiled}$ is certainly more informed than D^T , however, many of its attributes still have a large number of zero values that indicate the absence of a symptom. In order to make our dataset more informed, this research runs an intelligent algorithm that is based on the distribution of different categorical values (0, 1, 2, 3) in each ca or ha attribute in $D^T_{profiled}$. Step 3 of PIFS (Algorithm 1) illustrates its details. Thresholds 1 and 2, selected by combining expert opinion with experiments, are given below. Here, \wedge represents AND; \vee represents the OR symbol.

$$\begin{aligned}
 &threshold_1 : |zeros| \geq 250 \\
 &threshold_2 : 150 < |zeros| < 250 \wedge \\
 &(|ones| \vee |twos| \vee |threes| > 50)
 \end{aligned}$$

These thresholds are determined by the attributes in the current dataset and by the different categorical values that these attributes can have. Figure 2 shows two attributes that do not meet these threshold values and therefore are not selected by PIFS. Figures 3 and 4 illustrate distribution for four attributes that were selected by the proposed algorithm. As can be seen, the number of zeros in both Erythema and Scaling (Figure 3) is negligible. Similarly, figure 3 shows two attributes that were filtered out by our algorithm since they do not meet the two thresholds.

Step 4 creates a predictive model using k-nearest neighbor algorithm to predict the target (psoriasis or not) given a vector from $D^T_{reduced}$. This research chose to use k-nearest-neighbor due to its simplicity. Its results were compared to decision tree model as well (as shown in the experimental section).

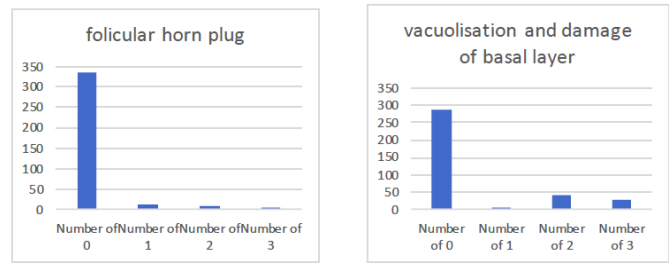


Figure 2. attributes not selected by PIFS

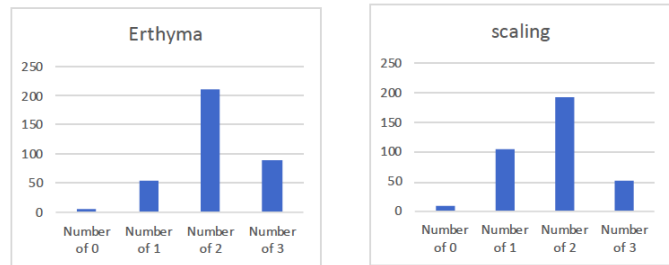


Figure 3. attributes selected by PIFS

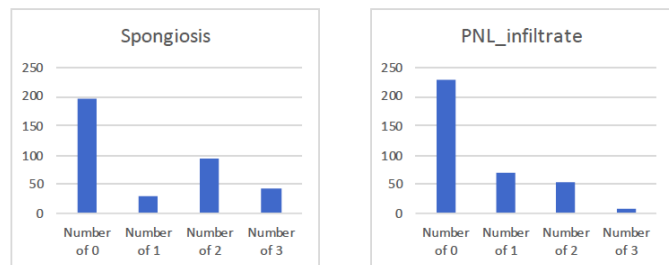


Figure 4. additional attributes selected by PIFS

K-nearest-neighbor (k-nn) algorithm [14] typically takes 4 inputs in order to predict a target attribute: an integer k (k=number of neighbors), a set of training samples whose target attribute y is known, a test vector t and a similarity or distance function. It then predicts test sample of class label t by performing the following steps :

- 1) calculate similarity between a test vector t and all training samples using the chosen similarity function (as explained below)
- 2) sort these similarity values and pick the top k samples - these are the k nearest neighbors of test sample t
- 3) Use the values of ta of each of the k nearest neighbors of t to predict a value for t. If the total number of neighbors that have a 'Yes' as its target attribute is greater than the total number of neighbors that have a 'No', assign 'Yes' as the class label of t; otherwise assign a 'No'.

Jaccard's Coefficient (JC) works best with asymmetric attributes [14] and therefore is more applicable to this research. Jaccard's Coefficient (JC) between two vectors x and y is

measured as

$$JC(x, y) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} \quad (1)$$

where f_{11} is the frequency of 1-1 match in vectors x and y , f_{01} is the frequency of occurrence of 0 and 1 in x and y (non-matching pair) and f_{10} is the frequency of occurrence of 1 and 0 in x and y (non-matching pair). For example, if $x = [1, 0, 0, 1]$ and $y = [1, 0, 1, 0]$, then $JC(x, y) = 1/3$. Since all the attributes in D are categorical and asymmetric, the similarity function that we choose to use in the proposed algorithm PIFS is Jaccard's coefficient of similarity.

Algorithm 1: The proposed algorithm PIFS

1. Preprocess and prepare attributes in the original dataset D (call it D^T) - as explained in section 2.2
2. Perform SCP (single-column profiling) data profiling algorithm on D^T to get a new set of attributes (call it $D_{profiled}^T$) - as explained in section 2.3
3. Use an intelligent feature selection method to eliminate those that are (lets call it $D_{reduced}^T$) - as explained in section 2.3

```

for each attribute  $i$  in  $D_{profiled}^T$  do
    find the distribution of its different categories
    if  $|zeros| < threshold_1$  then
        Add  $i$  to  $D_{reduced}^T$ 
    else
        if  $|two| > threshold_2$  OR  $|three| > threshold_2$ : then
            Add  $i$  to  $D_{reduced}^T$ 
        end
    end
end

```

4. Create a predictive model to predict the target attribute (presence or absence of psoriasis) - as explained in section 2.3
 - Divide data into 2 subsets using an 80-20 split: training (80% - $D_{reduced}^{Training}$) and test (20% - $D_{reduced}^{Test}$).
 - Training dataset is used to build the model, whereas test is used to test the model.
 - Build a model by applying a prediction algorithm such as k-nearest-neighbors on the training dataset $D_{reduced}^{Training}$.
 - Apply the model to each vector in the test dataset $D_{reduced}^{Test}$ to predict its ta values.

	Features selected by Weka	Number of zeros
1	Erythema	4
2	Itching	116
3	Scaling	8
4	Follicular-Papules	325
5	PNL-filtrate	229
6	Fibrosis of the papillary dermis	308
7	Clubbing of the rete ridges	248
8	Thinning of the supra-papillary epidermis	249
9	Follicular horn plug	336
10	Perifollicular parakeratosis	337

Table I
FEATURES SELECTED BY WEKA

	Features selected by PIFS	Number of zeros
1	Erythema	4
2	Itching	116
3	Scaling	8
4	Definite Borders	55
5	Exytosis	117
6	Acanthosis	9
7	Parakeratosis	85
8	Inflammatory mononuclear infiltrate	9
9	Spongiosis	195
10	Age	Categorical

Table II
FEATURES SELECTED BY PIFS

III. EVALUATION OF PIFS

The proposed algorithm is evaluated by the following criteria: (1) feature selection (steps 1, 2 and 3 of PIFS) (2) performance of the prediction algorithm (step 4 of PIFS) using measures such as accuracy, recall and fscore [11].

A. Feature selection

Intelligent feature selection is a core step of PIFS, as described in section 2. The selected features are compared with features selected by Weka [16]. Weka is a collection of machine learning algorithms for data mining tasks such as data preparation, feature selection, prediction and clustering.

The features (also referred to as attributes in this paper) selected by Weka using its feature selection algorithm (cfs-SubsetEval + BestFirst) and the number of zeros in each of the selected feature is shown in table I. Note that zeros in this dataset indicate the absence of a feature (or a symptom) of the target disease being predicted (e.g., Psoriasis). All features other than Erythema and Scaling have a very high number of zeros. This implies that Weka is not able to accurately distinguish between the features as they are extremely similar and overlapping – this finding concurs with the claim made by the creators of the dermatology dataset D[4]. Attributes selected by PIFS are shown in table II. The selected features in this table have less number of zeros (0 indicates absence of symptoms) than the features selected by Weka.

B. Performance measurement:

Accuracy measures the ability of the model to match the actual value of the target attribute with its predicted one (e.g.

"Yes" predicted as "Yes"). Accuracy alone is not always a deterministic measure, especially when dealing with target attribute values that are imbalanced or uncommon. For example, let's assume that in a dataset with 100 samples, there are 10 people with target attribute = "Yes" and 90 with a target attribute = "No". Also assume that the model predicts 1 out of 10 people correctly as "Yes", and 90 out of 90 as "No", then the accuracy of the model according to its definition (equation 2) will be calculated as high as 91%. However, this result is misleading as the model is not at all accurate in predicting one of the target attribute values (in this case "Yes"). Other measures that are used to evaluate predictive models are precision and recall. Most models achieve a trade-off between precision and recall, since it is very challenging to keep both the measures high. F-score is a combined measure that assesses this trade-off between precision and recall. Figures 6 and 7 show high values of accuracy and fscore for the proposed model.

$$accuracy = \frac{ActualTrueValuesPredictedCorrectly}{NumberOfPredictions} \quad (2)$$

$$precision = \frac{ActualTrueValuesPredictedCorrectly}{PredictedTrueValues} \quad (3)$$

$$recall = \frac{ActualTrueValuesPredictedCorrectly}{ActualTrueValues} \quad (4)$$

$$f\ score = \frac{2}{\frac{1}{r} + \frac{1}{p}} \quad (5)$$

Performance of PIFS is compared with the models created using Weka. Weka gives over 98% accuracy with classification models such as Naïve Bayes, iBK and J48 (decision tree) using the selected attributes shown in table I. Most likely, this is because weka is unable to distinguish between the presence or absence of a feature (symptom). We also compared the accuracy of PIFS using selected features and all 34 features in the original dataset D. PIFS accuracy with the selected features is as high as 92%, as compared to 60% when all features are used, as shown in figure 5, asserting that PIFS selects its features intelligently. Figure 5 also shows that the accuracy is highest at k = 7 with PIFS features.

IV. RELATED WORKS

In this section, existing work on data profiling and their limitations are discussed. We first describe the different applications of data profiling and thereafter describe the recent works done to cater to these applications. Commonly, data profiling is the set of actions required to determine metadata about a given dataset. Determining metadata requires computation on rows of the given dataset (e.g. finding rows that have missing values such as identifiers) or more commonly on columns in the dataset (single or multiple). In a survey on data

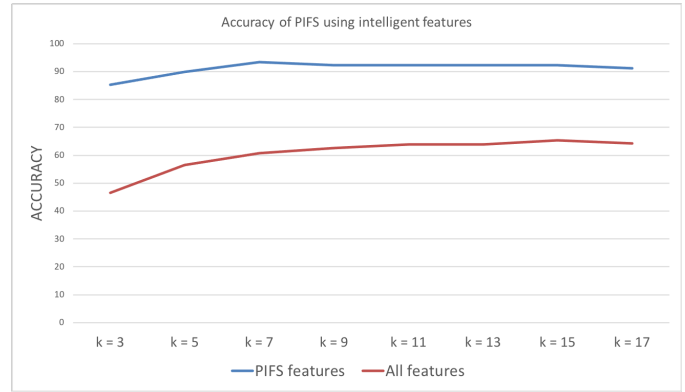


Figure 5. Accuracy of PIFS using intelligent features

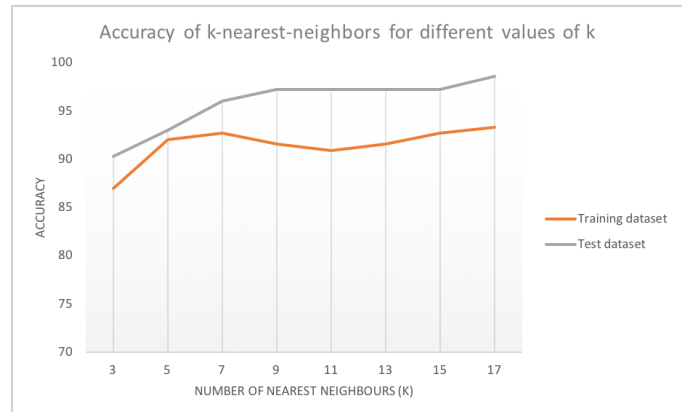


Figure 6. Accuracy of kNN model

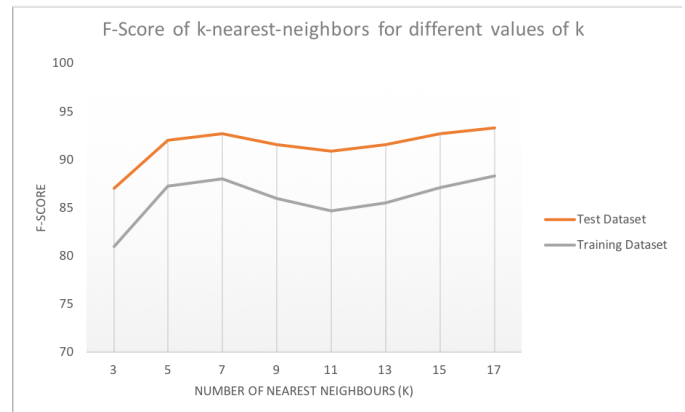


Figure 7. Fscore of kNN model

profiling [1], the authors classify data profiling tasks as single-column and multi-column profiling, depending on the use case of the task. A traditional application of profiling is in data exploration. Most times, database experts including researchers are faced with a collection of data with not much information about the data due to the proliferation of data generated by

machines. Two main challenges in data exploration are as follows: (1) to assume that the data to be explored is in a well-structured format [12] and (2) to use manual methods such as data gazing to explore data. Data gazing [3], involves experts manually skimming through the dataset to understand its characteristics. In order to do this in an automated way, experts need to know databases well, so that they can run SQL queries to get better insight of the data. But this poses an additional constraint that explorers always have to be database experts [7]. Gathering statistics about the data (typically called as metadata in DBMS) is a common task done by DBMS. Metadata could be on single columns (e.g. number of unique values) or on multiple columns (e.g. find pairs or groups of columns that can be used to optimize queries by DBMS). Such metadata is also useful in doing reverse database engineering (e.g. creating ER models from an existing dataset) [6][1]. The most common application of data profiling is data integration, where data from different sources and with possible different formats are integrated into a single source. When performing integration, columns or features of two schema from different sources are compared to find the matching ones [13]. Data cleansing is considered as an essential step in the process of knowledge discovery and data mining and the need to collect insight into data (using data profiling techniques) is inevitable [9]. Besides these applications, data profiling is extremely useful in Big Data analytics [2]. Operations such as storing and fetching big data are much more expensive than traditional structured systems. Since the data here is typically huge and heterogeneous, gaining an insight into the data and cleaning it before using it to store or query gives some hope to big data researchers [1]. Most of the methods of data profiling listed above follow the traditional cycle of extracting metadata and then using it towards an application.

The methods described above fail to address the volume and heterogeneity of datasets today. If traditional data profiling methods are followed, it would take days to achieve any significant information about the data. To keep up with these challenges, keeping in mind the computational cost of profiling is covered, this paper proposes a progressive algorithm that blends the traditional methods with state-of-the-art data mining methods (particularly supervised mining) to extract metadata. The proposed method is extensible and scalable and can be applied to very large datasets to generate results in a timely manner.

V. CONCLUSION AND FUTURE WORK

The importance of data profiling as a pre-processing step of machine learning models is well known. In this paper, we demonstrate that layering supervised data mining techniques to data profiling, thereby building intelligence in the feature selection process, results in higher accuracy of prediction of differential diagnosis of diseases where there are several overlapping features. The intelligent feature selection algorithm PIFS selects features from a dataset based on the data and its

distribution. No prior knowledge of the dataset is needed. The algorithm safely ignores “noisy data” and selects an optimal and reduced set of features that yields a highly accurate predictive model.

As future work, we plan to test and train the algorithm to achieve better accuracy for a variety of multivariate datasets where differential diagnosis of disease pose a challenge due to overlapping features. The proposed algorithm can also be extended to multiple classes in order to discriminate between other diseases.

REFERENCES

- [1] ABEDJAN, Z., GOLAB, L., AND NAUMANN, F. Profiling relational data: a survey. *The VLDB Journal* *The International Journal on Very Large Data Bases* 24, 4 (2015), 557–581.
- [2] AGRAWAL, D., BERNSTEIN, P., BERTINO, E., DAVIDSON, S., DAYAL, U., FRANKLIN, M., GEHRKE, J., HAAS, L., HALEVY, A., HAN, J., ET AL. Challenges and opportunities with big data 2011-1.
- [3] ARKADY, M. Data quality assessment.
- [4] DUA, D., AND KARRA TANISKIDOU, E. UCI machine learning repository, 2017.
- [5] GÜVENIR, H. A., DEMİRÖZ, G., AND ILTER, N. Learning differential diagnosis of erythematous-squamous diseases using voting feature intervals. *Artificial intelligence in medicine* 13, 3 (1998), 147–165.
- [6] HAINAUT, J.-L., HENRARD, J., ENGLEBERT, V., ROLAND, D., AND HICK, J.-M. Database reverse engineering. In *Encyclopedia of database systems*. Springer, 2009, pp. 723–728.
- [7] HANRAHAN, P. Analytic database technologies for a new kind of user: the data enthusiast. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (2012), ACM, pp. 577–578.
- [8] JOHNSON, T. Database reverse engineering. In *Encyclopedia of database systems* (2009), Springer.
- [9] KANDEL, S., PARIKH, R., PAEPCKE, A., HELLERSTEIN, J. M., AND HEER, J. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), ACM, pp. 547–554.
- [10] LIU, H., HUSSAIN, F., TAN, C. L., AND DASH, M. Discretization: An enabling technique. *Data mining and knowledge discovery* 6, 4 (2002), 393–423.
- [11] MARKOV, Z., AND LAROSE, D. T. *Data mining the Web: uncovering patterns in Web content, structure, and usage*. John Wiley & Sons, 2007.
- [12] MORTON, K., BALAZINSKA, M., GROSSMAN, D., AND MACKINLAY, J. Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment* 7, 6 (2014), 453–456.
- [13] NAUMANN, F. Data profiling revisited. *ACM SIGMOD Record* 42, 4 (2014), 40–49.
- [14] TAN, P.-N. *Introduction to data mining*. Pearson Education India, 2018.
- [15] TERRIZZANO, I. G., SCHWARZ, P. M., ROTH, M., AND COLINO, J. E. Data wrangling: The challenging journey from the wild to the lake. In *CIDR* (2015).
- [16] WITTEN, I. H., FRANK, E., HALL, M. A., PAL, C. J., AND DATA, M. Practical machine learning tools and techniques. In *DATA MINING* (2005), vol. 2, p. 4.

Optimization of Subsurface Imaging Antenna Capacitance through Geometry Modeling using Archimedes, Lichtenberg and Henry Gas Solubility Metaheuristics

Adrian Genevie Janairo^{1,*}, Jonah Jahara Baun¹, Ronnie Concepcion II², R-jay Relano², Kate Francisco², Mike Louie Enriquez², Argel Bandala¹, Ryan Rhay Vicerra², Melchizedek Alipio¹, Elmer P. Dadios²

¹Department of Electronics and Computer Engineering, De La Salle University, Manila, Philippines

²Department of Manufacturing Engineering and Management, De La Salle University, Manila, Philippines

{adrian_janairo*, jonah_baun, ronnie.concepcion, r-jay_relano, kate_g_francisco, mike.enriquez, argel.bandala, ryan.vicerra, melchizedek.alipio, elmer.dadios}@dlsu.edu.ph

Abstract—Capacitive resistivity subsurface imaging of roads operating at very low frequency is susceptible to antenna characteristic capacitance dynamics that may cause unwanted signal reflection, coupling, and unfavorable effect on reception sensitivity. Antennas are conventionally modeled using a complex and repetitive default mathematical method that is prone to human error and discrete results. To address this emerging challenge, this study has developed a new technique for plate-wire antenna capacitance optimization through equatorial dipole-dipole antenna geometry modeling using genetic programming (GP) integrated with metaheuristic methods, namely Archimedes optimization algorithm (AOA), Lichtenberg algorithm (LA), and Henry gas solubility optimization (HGSO). GP was used to construct the antenna capacitance fitness function based on 241 combinations of wire antenna radius and elevation, and dipole plate elevation, length, width, and thickness measurements. Minimization of antenna capacitance (approaching 1 nF) to achieve quasi-static condition was performed using GP-AOA, GP-LA, and GP-HGSO. The 3 metaheuristic-based antennas were 3D-modeled using Altair Feko and compared from the default antenna's electrical features. It was found that even with the smallest dipole geometry, hybrid GP-LA antenna model exhibited the most practical outputs at 5 kHz with correct directional propagation based on its radiation pattern, a realistic receiver voltage of -8.86 dBV which is close to the default model, and a high-power efficiency of 99.925%. While hybrid GP-AOA and GP-HGSO resulted in indirect coupled transceiver systems with unsuitable antenna characteristic capacitance inducing anomalous receiver voltages. The experimental results prove the validity of the developed technique for more accurate determination of optimal antenna geometry.

Keywords—antenna geometry modeling, Archimedes optimization, Henry gas solubility optimization, Lichtenberg optimization, subsurface imaging

I. INTRODUCTION

One of the most popular geophysical imaging techniques used in a broad range of applications is subsurface imaging [1]. It is a non-destructive imaging and evaluation tool based on transmission and reception of low spectral electromagnetic

(EM) waves [2]. This working concept is proven to be very effective in studying the electrical properties of soil and is broadly used in a variety of applications, including surveying underground infrastructures such as cables and pipes, land mines, road inspections, and archaeological and geological studies. [2-3]. Subsurface mapping techniques such as electrical resistivity tomography, use of ground-penetrating radar (GPR), accelerographs, and magnetic techniques can contribute valuable data on urban engineering and environmental sites such as industrial lands, runways, and paved roads [3]. However, these methods may intercept other active signals when utilized in urban areas because these typically operate from high frequency up to ultra-high frequency spectrum. With regards to the efficiency of data acquisition and practical aspects, these approaches still require optimization of the antenna geometry [3].

The quasi-static condition in subsurface imaging is the irreversible effect that is materialized when the frequency of operation from extremely low (ELF) to very low frequency causes dynamic effects to become negligible [3]. In other words, the induction number becomes almost zero and the electrostatic geometric factor reaches one [3]. This is a condition wherein the transmitting and receiving antenna or sensors are fully capacitively coupled to the ground but still can map the electromagnetic skin depth without any change in the signal waveform. The operating frequency, geometric conditions, and ground resistivity or conductivity are factors that affect the quasi-static condition in capacitive coupling to the ground surface [3]. However, only the geometric condition of the antenna can be modified and optimized to avoid the buildup of parasitic capacitance causing matrix residual current or leakage current, ringing of the transmitter terminal, and an increase in the reflected power on the antenna transmission line [3].

Most of the antennas used in industry have no analytical formulas for parameter assessment and optimization [4]. To optimize and simulate such antennas, it is essential to utilize electrostatics modeling software such as using XFDTD antenna design software which is also widely used in the field

of engineering [4]. For antenna effects such as antenna response mismatch and coupling between transmit and receive antennas, near-field calibration can be applied to GPR using the Higher-Order Basic Integral Equation Solver (HOBBIES) EM simulation software or incorporated into LINUX software. The amount of time it takes to complete a task has been reduced, and the level of stability has improved [5-6]. To achieve optimal conditions for the activated antenna pair, the signal type and dielectric properties of the embedded object are analyzed as independent parameters using the Subsurface Imaging (SSI) algorithm and the Finite-difference Time Domain (FDTD)-based virtual tool GrGPR. Comparatively, dipole linear antenna array models associated with towing system can be improved through CST Microwave Studio Simulation Software while optimization in slot antenna geometry is devised by folding the ground of the slot antenna along the slot length to enhance the gain and lessen the backward radiation using Agilent Technologies E8362B vector network analyzer [8-9]. Antennas with small gaps in topology and complex design geometry are becoming progressively prevalent in research and existing commercial equipment, which is believed to be uncertain to use with generators of high voltages [4]. The most recent antenna geometry optimization related to capacitive resistivity imaging uses capacitive dipole [3]. Properties of the antenna were studied and optimized by modeling the capacitance of their architectural parts [3]. This technique has used the modeling of sensor capacitances to expose that the idea of point poles claimed withinside the quasi-static system of capacitive resistivity has a practical understanding withinside the form of plate-wire sensors given that the plate dimensions are limited compared with its dipole length [3]. The most suitable geometry for capacitance resistivity measurements is the use of a dipole-dipole array and data accumulation is accomplished using plate-wire combinations equatorial geometry arrangement [3]. Genetic algorithm has been mainly used also for optimizing antenna geometry based on antenna bandwidth and sensitivity [10], properties of electromagnetic poles [11], S_{11} , and gain parameters [12] that are all extracted from another simulation software. Hence, these optimization techniques are believed to be more effective and efficient through the integration of computational intelligence methods.

Computational intelligence (CI) has offered several approaches for design optimization used for subsurface imaging through population-based metaheuristic algorithms [10-12]. Various applications of Archimedes Optimization Algorithm (AOA) were to attain the best performance of a wind energy generation system, to identify the optimal model of proton exchange membrane fuel cell (PEMFC) stacks based on an improved version of the Deep Belief Network (DBN) and for improving the performance of feedforward neural networks (FFNN) [13-15]. The Lichtenberg algorithm (LA) is shown to be a powerful tool for the identification of damage in mechanical structures constructed by composite materials, providing a more exact prediction that can be useful for modeling an aluminum alloy 2025 T6 metal matrix composite reinforced with red mud [16-18]. Moreover, Henry gas solubility optimization (HGSO) was employed in the field of the automotive industry wherein it was recorded that it can

design better optimal components as well as in obtaining the best parameters in terminal voltage control of a generator in an automatic voltage regulator (AVR) system utilizing a fractional-order proportional-integral-derivative (FOPID) controller [19-21]. Given some of their applications, AOA, LO, and HGSO are said to be high-performance optimization tools that deliver superior and competitive results when solving challenging engineering optimization problems compared to other existing solving algorithms [22-24]. Thus, these methods are proposed in this study to enhance the characteristic capacitance of a single-pair equatorial dipole-dipole subsurface imaging antenna through geometry modeling. The default mathematical method optimizes the antenna results in a prolonged approach as it requires the designer to perform computation, computer-aided designing, reiterated simulations, and extraction of simulated data [10-12].

Despite the advances in antenna geometry optimization techniques for subsurface imaging, other properties of the antenna are still can be optimized by modeling the plate-wire capacitance of the antenna. To work on this emerging challenge, this study has developed a new capacitance model for optimizing the geometry of a single pair equatorial dipole-dipole antenna for very low frequency capacitive coupling to the ground using genetic programming (GP) integrated with metaheuristic optimization algorithms, namely Archimedes, Lichtenberg, and Henry gas solubility optimizers. This is vital to keep the antenna in quasi-static condition for subsurface imaging based on minimum capacitance (approaching 1 nF) to successfully transmit adequate electrical current into the ground. The suitability of AOA, LA, and HGSO was explored in optimizing the antenna geometry parameters, namely elevation of wire segment, radius of wire antenna conductor, dipole plate elevation from the ground, width, length, and thickness of plate that have a direct impact on the plate-wire capacitance of the antenna. The output of this study is an equatorial dipole-dipole antenna with a minimized capacitance that can operate at 5 kHz having higher efficiency than the default model. This study contributes to the: (1) development and application of hybrid GP-AOA, GP-LA, and GP-HGSO in minimizing the capacitance of an antenna for underground imaging in a quasi-static condition; (2) innovation of an efficient technique for faster and more accurate determination of optimal antenna geometry avoiding long mathematical computation; and (3) quantifying the correlations of antenna geometry and its characteristic capacitance.

II. MATERIALS AND METHODS

This study includes four major phases: (1) modeling of single-pair equatorial dipole-dipole antenna geometry using the standard mathematical technique, (2) application of genetic programming to develop the capacitance model, (3) minimization of antenna capacitance through metaheuristic algorithms, (4) remodeling of antenna geometry based on optimized capacitance using industry-standard software, and (5) electrical parameter comparison and statistical evaluation (Fig. 1). Evaluation of the default and proposed CI-based capacitance models are one of the highlights of this work.

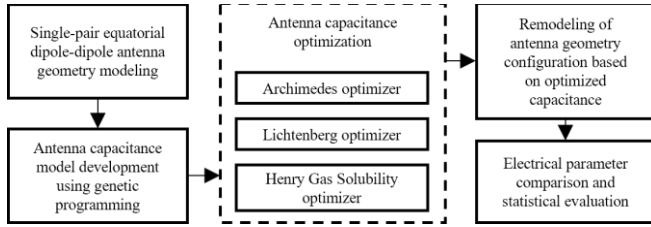


Fig. 1. Developmental framework for the optimization of subsurface imaging antenna capacitance through geometry modeling using Archimedes, Lichtenberg, and Henry Gas Solubility metaheuristics

A. Single-pair Equatorial Dipole-dipole Antenna Geometry Modeling

A 5 kHz single-pair equatorial dipole-dipole antenna was 3D modeled in Altair Feko software (Altair, Singapore) based on the standard mathematical approach in determining the radius of the wire antenna segment (a , m), wire antenna elevation (h_w , m), dipole plate elevation from the ground (h_p , m), dipole length (l), dipole width (w , m), and dipole thickness (t , m) (Fig. 2). These antenna geometries are computed to comply with the quasi-static condition in which the geometric factor (K_{eq} , unitless) is approaching 1 (1) with the dipole spacing (r , m) set to 1 m. The target capacitance for this antenna configuration was made close to 1 nF by computing the wire antenna segment (C_w , F) (2) and dipole plate (C_p , F) (3) capacitances combined making a total wire-plate capacitance of $0.5(C_w + C_p)$ with ϵ_o as the free space permittivity of 8.854 pF/m, ϵ_r as the permittivity of air equivalent to 1.001, and d as the distance of the center of wire to the ground that is maintained at 0.209 m. This is to reduce stray parasitics that could affect the reception sensitivity and may drive signal reflection. The dipole spacing (r , m) from the transmitter to the receiver was maintained at 1 m to achieve the ideal induction number (B , unitless) which is approaching 0.

$$K_{eq} = \frac{\frac{1}{\sqrt{r^2+4h_p^2}} \frac{1}{\sqrt{r^2+l^2+4h_p^2}}}{\frac{1}{r} \frac{1}{\sqrt{r^2+l^2}}} \quad (1)$$

$$C_w = \frac{2\pi\epsilon_o\epsilon_r}{\text{acosh}\left(\frac{d}{a}\right)} \quad (2)$$

$$C_p = \epsilon_o\epsilon_r \left[1.15 \frac{wl}{h_p} + 1.4 \left(\frac{t}{h_p}\right)^{0.222} (2w + 2l) + 4.1 \left(\frac{t}{h_p}\right)^{0.722} h_p \right] \quad (3)$$

B. Antenna Capacitance Model Development Using Genetic Programming

Genetic programming (GP) is an evolutionary algorithm that is a hybrid of genetic algorithm and regression tree [25]. In this study, a multigene symbolic regression GP was employed through MATLAB GPTIPsv2 tool to construct the capacitance fitness function as a function of a , h_w , h_p , l , w , and t . A dataset of 241 rows generated by series of 241 computations of different combinations of antenna geometries has been used as input to GP following a stratified sampling scheme of 56% of the dataset volume is set for training the

model, 24% for validation, and 20% for testing. GP starts with initializing 100 random antenna capacitance chromosomes

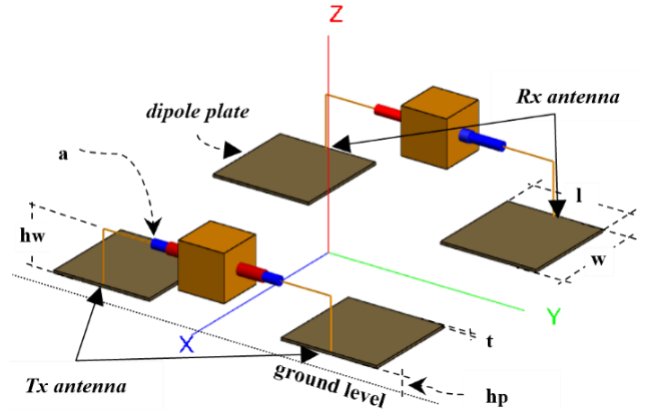


Fig. 2. 3D schematic of an equatorial dipole-dipole antenna for subsurface imaging operating at 5 kHz where a is the radius of the wire antenna, h_w is the wire antenna elevation, h_p is the dipole plate elevation from the ground, and l , w , and t are the length, width, and thickness of the dipole plate.

and then evaluates the fitness value of the chromosomes (Fig. 3). Next, symbolic regression was performed using mathematical functions with a maximum of 10 genes, and a maximum tree depth of 5. It is followed by selecting 3 capacitance chromosomes through a tournament with a size of 50, 0.1 elite fraction, and enabled lexicographic selection pressure. The fittest (minimum) capacitance chromosomes make the new generation and undergo crossover and mutation with a rate of 0.84 and 0.14, respectively, with an evolutionary rate covariation probability of 0.1. After 100 elapsed generations, GP converged resulting in a capacitance mathematical topology without dipole thickness as non-significant features are neglected in the tournament selection (4). This fitness expression was used for metaheuristic modeling of antenna geometry.

$$C = f(a, h_w, h_p, l, w) \quad (4)$$

C. Antenna Capacitance Optimization Using Metaheuristic Algorithms

In this study, three metaheuristic algorithms, Archimedes optimization algorithm (AOA), Lichtenberg algorithm (LA), and Henry gas solubility optimization (HGSO) were explored in optimizing the antenna capacitance through geometry modeling based on the GP-constructed fitness function (4) (Fig. 3). Inspired by the law of buoyancy, Archimedes optimization algorithm uses the population of solutions as the objects immersed in fluid considering the parameters density, volume, and acceleration to attain neutrally buoyant objects and identify the object with the best fitness score. Part of the iterative process is to update each object's acceleration using (5) to get the new positions when the transfer function operator (TF) ≤ 0.5 , where acc_i^{t+1} is the acceleration of object i for iteration $t+1$, den_{best} , vol_{best} and acc_{best} are the best recorded object's density, volume and acceleration, respectively, while den_i^{t+1} and vol_i^{t+1} are the density and volume of object i for iteration $t+1$, respectively. [23]. Here, the object number was set to 5 representing the antenna

parameters in the fitness function and the four ranges of normalization were set to 2, 6, 0.9, and 0.1.

$$acc_i^{t+1} = \frac{den_{best} + vol_{best} + acc_{best}}{den_i^{t+1} \times vol_i^{t+1}} \quad (5)$$

Lichtenberg algorithm is based on a natural phenomenon of intra-cloud lightning where Lichtenberg figures (LF) are bitmapped in a multi-dimensional space and the particles are the candidate solutions that will be clustered according to the Diffusion Limited Aggregation theory which considers the creation radius (R_c) and stickiness coefficient (S) that influence the fractal dimension (D) (6), and density, respectively, where $N_{cluster}$ and $R_{cluster}$ are the numbers of particles and the average radius of a cluster, respectively. [22]. In this study, LA was configured with 100 population size, $1e^6$ particles for 5D search space, stick probability of 1, creation radius of 150, and refinement of 0.2 (Fig. 3).

$$D = \ln(N_{cluster}) / \ln(R_{cluster}) \quad (6)$$

Henry gas solubility optimization algorithm treats the population of solutions as gases which will be clustered by gas

type based on Henry's constant value that is necessary to obtain the new solubility in (7) where $S_{i,j}(t)$ and $P_{i,j}(t)$ corresponds to the solubility and partial pressure of gas i in cluster j during iteration t , $H_j(t+1)$ is the new Henry's coefficient of cluster j for iteration ($t+1$) and K is a constant, and position of each gas. The equilibrium state of each gas determines the best gas for each cluster where the optimal gas is chosen [24]. In this study, HGSO was configured with 100 initial gases, 5 gas types, and 2 worst gas clusters (Fig. 3).

$$S_{i,j}(t) = K \times H_j(t+1) \times P_{i,j}(t) \quad (7)$$

For AOA, LA, and HGSO, the geometric parameter boundaries were set to [0.0004, 0.2008], [1.91e⁻¹⁷, 0.5524], [0.003, 0.007], [0.1, 0.5], and [2e⁻⁵, 0.002] for a , h_w , h_p , l , and w , respectively. The three algorithms converged after 80 generations and provided the global best solution for minimized antenna capacitance with specific antenna geometry parameter values.

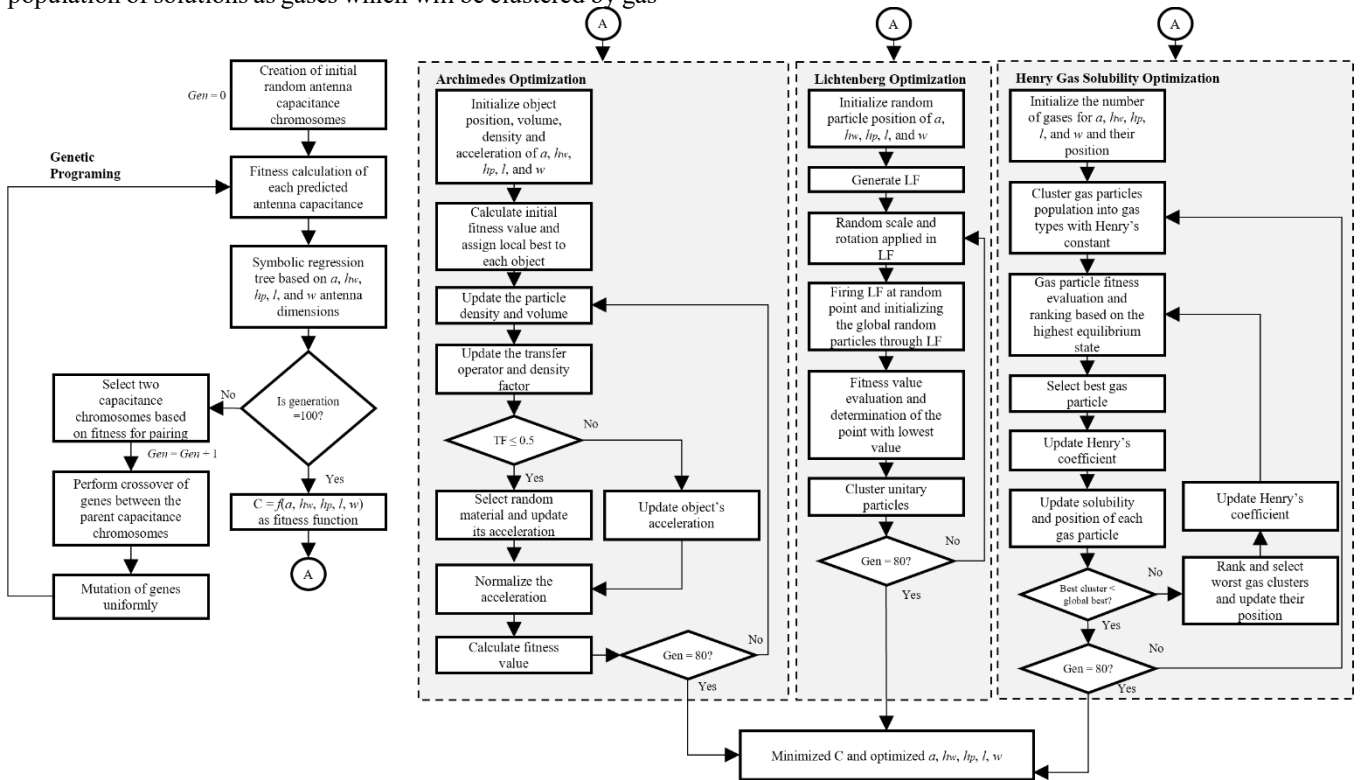


Fig. 3. Proposed hybrid multigene symbolic regression genetic programming and Archimedes, Lichtenberg, and Henry Gas Solubility optimizer framework in minimizing the value of antenna capacitance approaching 1 nF and determining the exact global best combination of the antenna geometry (radius of coaxial cable (a), antenna wire segment elevation (h_w), dipole plate elevation (h_p), and dipole plate (l) and width (w)).

D. Remodeling of Antenna Geometry Based on Optimized Capacitance Model

After a series of explorations and exploitations of AOA, LA, and HGSO, there were 3 new sets of a , h_w , h_p , l , and w of antenna geometry that influence the antenna capacitance. By using Altair Feko software, 3 3D antenna models were

designed with the following additional mechanical configurations: 1 m dipole spacing, graphite as the enclosure material, 0.003 mm enclosure thickness and 0.001 m dipole plate thickness. For the electronic configurations, antennas were simulated with 1kV and 10 W transmitter voltage and emitted power at 5 kHz operation, 5 MΩ receiver resistance. For the environment configuration, the ground plane was set

to planar multilayer substrate with soil resistivity of 3 units and the circuit box dimension to 0.2 m x 0.2 m x 0.2 m. Radiation pattern, electric and magnetic fields, voltage, and power efficiency at the receiver antenna were extracted and graphed using the same tool for the default antenna model (mathematical), GP-AOA, GP-LA, and GP-HGSO antennas. These electronic measurements served as the bases for evaluating which among the developed antenna models is the most recommended for 5 kHz subsurface imaging of roads.

E. Statistical Analysis

Pearson’s correlation analysis with $\alpha \leq 0.05$ was performed using Minitab (Minitab LLC) and surface graphs were constructed to show dynamic relationships among antenna geometry parameters and capacitance.

III. RESULTS AND DISCUSSION

A. Dynamic Relationships of Antenna Geometry and Capacitance

For all the developed antenna models, the total wire-plate antenna capacitance has a strong positive correlation to dipole plate length ($R = 0.907$) and a strong inverse correlation with plate elevation above ground ($R = -0.851$). As antenna wire segment and dipole plate width increase in size, antenna capacitance also increases moderately. On the other hand, the wire antenna’s radius weakly and inversely influences capacitance. Interestingly, if the plate elevation was increased greater than 0.006 m and the antenna wire elevation is below 0.175 m, capacitance approaches minimum (< 40 pF), and maximum capacitance of 100 pF was characterized when plate elevation is 0.006 m and wire elevation is 0.175 m (Fig. 4a). Another finding is that the antenna capacitance is almost constant given that dipole plate width ranges from 0.2 to 0.4 m and the antenna wire radius ranges from 0.02 to 0.03 m (Fig. 4b). It was also analyzed that there is a capacitance trough for the combination of dipole width below 0.2 m and antenna wire radius greater than 0.02 m. Hence, this exact issue in antenna modeling demands the potential of computational intelligence through metaheuristic optimizations.

B. Minimized Capacitance as a Requirement for Quasi-static Condition

By completing the 100 generations of 7-gene GP, it resulted in an antenna capacitance model (8) that considers the issue of non-linearity of antenna geometry. This sensitive GP-based antenna capacitance model is acceptable in this application as it scored a testing accuracy of 82.764% with negligible RMSE of $1.1049e^{-11}$ and MAE of $4.8668e^{-12}$. It was used as the fitness function for all metaheuristic optimizations.

$$C = 2.07e^{-11}h_w - 2.07e^{-11}a + 3.14e^{-10}l + 3.14e^{-10}w - 1.09e^{-10}\log(h_p) - 4.05e^{-13}\log(h_w)^3 + 1.06e^{-8}a^3 - 6.73e^{-10} \quad (8)$$

After 80 generations, GP-AOA, GP-LA, and GP-HGSO converged (Fig. 5) and provided their individual global best solution pertaining to the optimum combination of antenna’s a , h_w , h_p , l , and w for minimized capacitance (Table 1). The fitness curves in Fig. 5 show that only GP-LA and GP-HGSO capacitance models have exhibited substantial exploration and exploitation (marked with red vertical line) (Fig. 5). GP-LA

explored in advanced the 5D search space, thus, locating the cluster of particles in a constraint region as early as the 14th iteration and deliberately exploited to having a minimum capacitance of 63.5321 pF (Table 1). GP-HGSO finished its

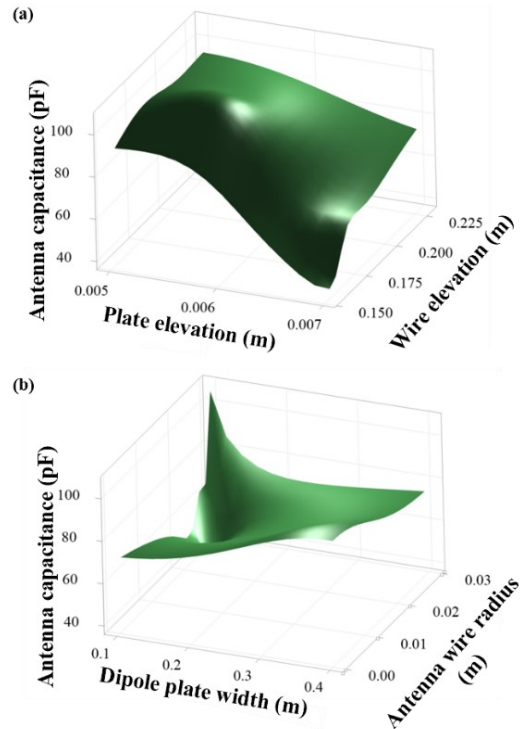


Fig. 4. Dynamics relationship of the capacitance of single-pair equatorial dipole-dipole antenna for 5 kHz operation with references to (a) plate and wire elevations and (b) dipole plate width and radius of antenna wire.

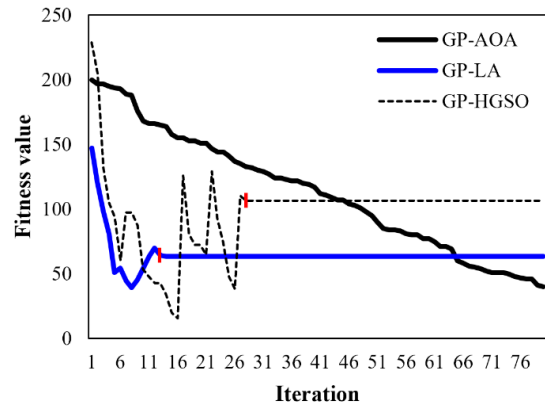


Fig. 5. Fitness curves of the developed capacitance model as optimized using Archimedes, Lichtenberg, and Henry Gas Solubility metaheuristics with a maximum generation of 80. The red vertical lines show the starting point of the exploitation stage for convergence.

exploration at a later iteration than GP-LA and has exploited at a minimum capacitance of 106.5584 pF (Table 1). On the other hand, GP-AOA did not exhibit any exploitation even if the iteration has been extended to 1000. It suggests that GP-AOA neither had a premature convergence nor converged properly even if the resulting capacitance turned to 40.103 pF (Table 1). As compared with the default model, the developed metaheuristic models’ optimum capacitances are just below

and above its value signifying that these hybrid models are acceptable. Likewise, these findings show that the extracted antenna geometry by GP-AOA, GP-LA, and GP-HGSO are within a realistic range (Table 1).

TABLE I. SUMMARY OF ANTENNA GEOMETRY CONFIGURATIONS AS RESULTS OF DEFAULT MODEL AND PROPOSED METAHEURISTIC ALGORITHMS AND THE CORRESPONDING ANTENNA CAPACITANCE (DIPOLE PLATE = 0.001 M)

Antenna Model	a (m)	h _w (m)	h _p (m)	l (m)	w (m)	C (pF)
Default	0.7813e ⁻³	0.2286	0.005	0.4000	0.4000	94.6080
GP-AOA	0.0227	0.1549	0.0069	0.1474	0.1233	40.1031
GP-LA	0.0277	0.1762	0.0070	0.1001	0.1012	63.5321
GP-HGSO	0.0292	0.1800	0.0056	0.5655	0.1000	106.5584

This study conforms to the trend of hybridization of multigene genetic programming with other population-based optimization metaheuristics [25-33]. Previous studies in equatorial dipole-dipole antenna optimization have only dealt with applying genetic algorithm to optimize the geometry based on bandwidth and sensitivity [10], properties of electromagnetic poles including S₁₁ and gain parameters [11, 12] that are all extracted from another simulation software. Hence, the proposed approach of employing computational intelligence is an efficient technique for faster and more accurate determination of optimal antenna geometry avoiding long mathematical computation. Here, instead of performing iterative manual computing of quasi-static condition requirements based on altered combinations of antenna geometry, the optimum characteristic capacitance can be determined basically by using the developed fitness function (8) and the intelligent metaheuristic model will optimize the independent parameters.

C. Enhanced Antenna Geometry Yielded through Capacitance Minimization

For an equatorial dipole-dipole antenna operating at a very low frequency (5 kHz) for subsurface imaging, considerably high characteristic capacitance may negatively influence the directivity of the antenna and produce reactance that will make the antenna resistive causing it not to emit the signal properly. The antenna geometries generated using the default technique, GP-AOA, GP-LA, and GP-HGSO were modeled resulting in radiation spectrums (Figs. 6a to 6d). Note that the dipole length and width are significantly varied in each model causing diverse radiation patterns (Figs. 6a to 6d). It is noticeable that even with the small dipole geometry of the GP-LA antenna, there is no signal reflection and there is directional emission (Fig. 6c).

Because the length recommended by GP-HGSO is 564.935% longer than the GP-LA, the dipole spacing between the transmitter and receiver becomes significantly closer causing widened electric field and magnetic field at the receiver (Figs. 6e to 6f). This instance is probably because of direct coupling due to close proximity. The electric and magnetic fields exhibited by the receiver of GP-LA antenna

mimic the detection of an object as they are condensed in specific proximity (Figs. 6e to 6f). One of the strong bases for evaluating which among these antenna models is the most recommended for subsurface imaging application is the receiver voltage in which the GP-LA records a -8.86 dBV (0.361 V) which is realistic and close to what the default model (standard) records that is -22.5 dBV (0.075 V) (Fig. 6g). This is supported by previous studies that when the transmitted signal is rated above 1 kV operating at very low to high frequency, the receiver voltage signal should just be around microvolt to 2 V [5, 7, 8, 12]. The pronounced receiver voltage values of GP-AOA and GP-HGSO, 38.459 V and 8.913 V respectively, are mainly because of indirect coupling of transmitter and receiver dipoles and unsuitable antenna characteristic capacitance. Moreover, in the case of power efficiency, all models got a high rating with a lowest of 98.658% (default model) and a highest of 99.953% (GP-HGSO) (Fig. 6h). This is to emphasize that GP-HGSO got a high power efficiency rating because of its coupling issue whereas the GP-LA antenna given its practical receiver voltage value, exhibited a 99.925% power efficiency considering this model as the most recommended antenna for VLF subsurface imaging. In subsurface imaging of utilities in urban roads, the characteristics exhibited by GP-HGSO antenna are not recommended as very shallow or probably no utilities will be mapped out, while GP-LA is highly recommended as there is correct directional propagation that is penetrating through the ground and the receiver is sensitive enough to detect the reflected signal at the same phase of the transmitted current. To sum it up, an equatorial dipole-dipole antenna does not require a high voltage reading at the receiver end, instead, it demands sensitivity and proper coupling through manifesting quasi-static condition. And quasi-static condition extendedly requires minimal characteristic capacitance close to 1 nF. This study has established an understanding that the capacitance of 63.532 pF is enough to provide high efficiency and sensitive receiver antenna. This study is limited in thorough experimentation and analysis of the possible impact of different transmitter voltage levels, which could be a good recommendation for future studies.

IV. CONCLUSION

This study proposed a new technique for optimizing the capacitance of an equatorial dipole-dipole antenna operating at a very low frequency for subsurface imaging of utilities embedded beneath the road through the integration of computational intelligence. Combining 7-gene genetic programming (GP) with Archimedes optimization algorithm (AOA), Lichtenberg algorithm (LA), and Henry gas solubility optimization (HGSO) resulted in varied antenna geometries and electrical characteristics. With this, 4 antenna models were developed, namely the default mathematical antenna model, GP-AOA, GP-LA, and GP-HGSO from which the GP-LA exhibited the most practical geometry, realistic sensitivity, and efficient power system. Antenna geometry and signal coupling present in GP-AOA and GP-HGSO negatively influences the electric and magnetic fields and received voltage signal. Dipole length and elevation greatly affect the characteristic capacitance of the antenna. Lastly, this study innovated an efficient technique for faster and more accurate

determination of optimal antenna geometry avoiding long mathematical computation as proven by the GP-LA antenna model. For future studies, variation of transmitted signal

electrical parameters including voltage, frequency, and signal waveform may be done to further improve the antenna model.

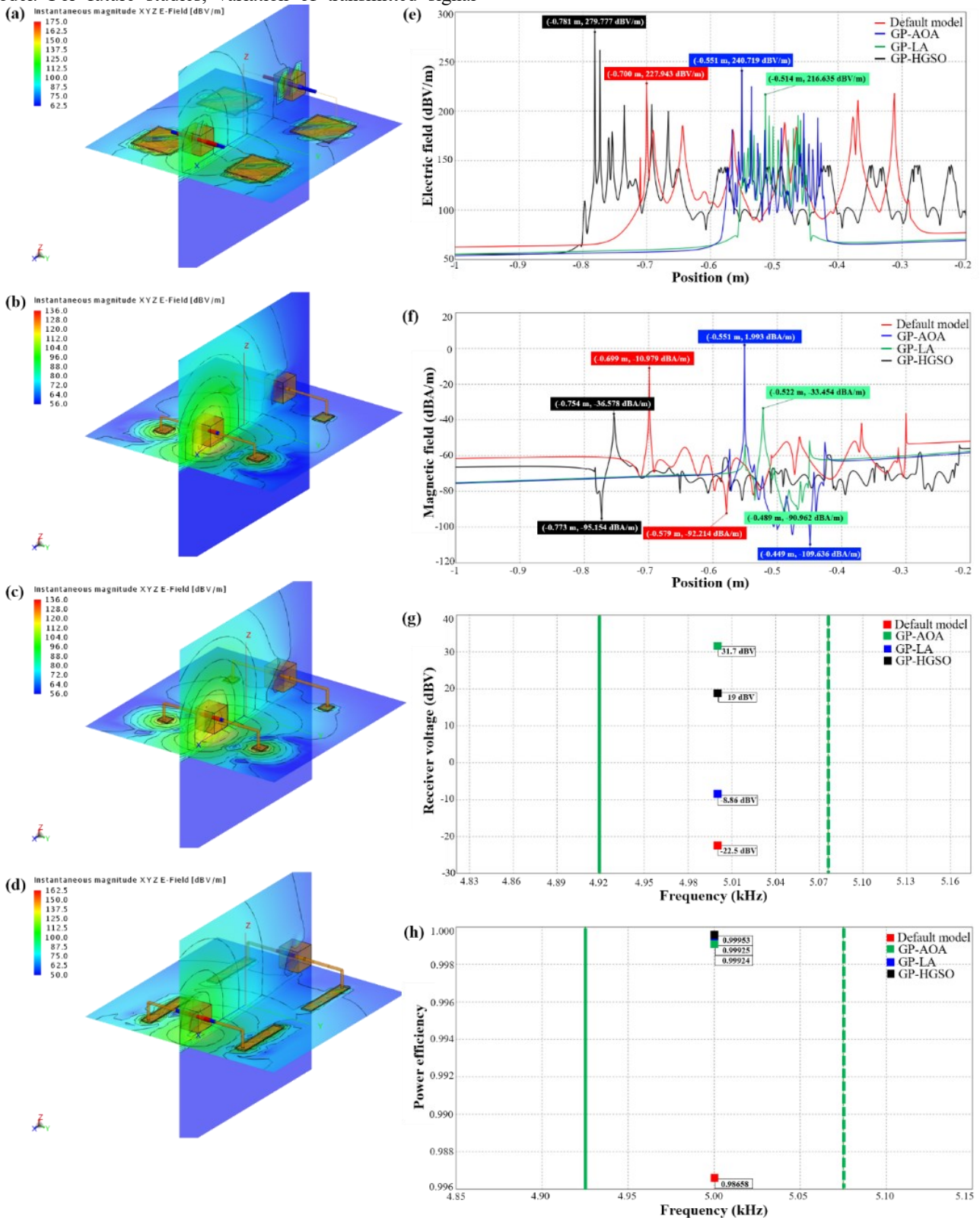


Fig. 6. Comparison of radiation pattern for the (a) default, (b) GP-AOA, (c) GP-LA, and (d) GP-HGSO models reflecting the optimized antenna geometry. (e) Electric and (f) magnetic fields of near field radiation with respect to position. (g) Receiver voltage and (h) power efficiency with respect to frequency of operation.

ACKNOWLEDGMENT

The authors would like to thank the Philippine Council for Industry, Energy and Emerging Technology Research and Development of the Department of Science and Technology and the Intelligent Systems Laboratory for the support granted.

REFERENCES

- [1] R. Deiana, D. Vicenzutto, G.P. Deidda, J. Boaga, and M. Cupitò, "Remote Sensing, Archaeological, and Geophysical Data to Study the Terramare Settlements: The Case Study of Fondo Paviani (Northern Italy)", *Remote Sensing*, vol. 12, no. 16, pp. 2617, 2020.
- [2] H. Ali, A. Z. A. Firdaus, M. S. Z. Azalan, S. N. A. M. Kanafiah, S. H. Salman, M. R. Ahmad, T. S. T. Amran, and M. S. M. Amin, "Classification of different materials for underground object using artificial neural network", *IOP Conference Series: Materials Science and Engineering, 5th International Conference on Man Machine Systems, Pulau Pinang, Malaysia*, vol. 705, 2013.
- [3] K. Oliver, "The Capacitive Resistivity Technique for Electrical Imaging of the Shallow Subsurface", *PhD thesis, University of Nottingham*, 2002.
- [4] A. V. Kochetov, G. V. Komarov, and D. Y. Kulikova, "Ultra-wideband Multidisc Antenna with Reconfigurable Polarization for Ground-Penetrating Imaging Radar", *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering*, 2019.
- [5] L. Bannawat, A. Boonpoonga, and S. Burintramart, "Resolution Improvement of GPR Image using Antenna Calibration for Object Detection", *2017 International Symposium on Antennas and Propagation (ISAP)*, 2017.
- [6] S. Dai, L. Liu, and G. Fang, "A Low-cost Handled Integrated UWB Radar for Shallow Underground Detection", *Proceedings of 2010 IEEE International Conference on Ultra-Wideband (ICUWB2010)*, 2010.
- [7] E. Karpat, "Subsurface Imaging Analysis for Multiple Objects", *2013 4th International Conference on Intelligent Systems, Modelling and Simulation*, 2013.
- [8] L. Wang, "Analysis and Development of Components of Dipole Linear Antenna Array", *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 2017.
- [9] W. Kang, S. Lee, and K. Kim, "A Ground-Folded Slot Antenna for Imaging Radar Applications", *IEEE Antennas and Wireless Propagation Letters*, vol. 10, 2011.
- [10] X. Zhuge, T. G. Savelyev, A. G. Yarovoy, and L. P. Ligthart, "Subsurface imaging with UWB linear array: Evaluation of antenna step and array aperture," *2007 IEEE Int. Conf. Ultra-Wideband, ICUWB*, pp. 66–70, 2007, doi: 10.1109/ICUWB.2007.4380917.
- [11] A. De Coster and S. Lambot, "Full-Wave Removal of Internal Antenna Effects and Antenna-Medium Interactions for Improved Ground-Penetrating Radar Imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 93–103, 2019, doi: 10.1109/TGRS.2018.2852486.
- [12] A. Srivastav, P. Nguyen, M. McConnell, K. A. Loparo, and S. Mandal, "A Highly Digital Multiantenna Ground-Penetrating Radar (GPR) System," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7422–7436, 2020, doi: 10.1109/TIM.2020.2984415.
- [13] A. Fathy, A. G. Alharbi, S. Alshammari, and H. M. Hasanien, "Archimedes optimization algorithm based maximum power point tracker for wind energy generation system", *Ain Shams Engineering Journal*, vol. 13, issue 2, no. 101548, 2022.
- [14] Z. M. Ali, I. M. Diaeldin, A. El-Rafei, H. M. Hasanien, S. H. E. A. Aleem, and A. Y. Abdelaziz, "A novel distributed generation planning algorithm via graphically-based network reconfiguration and soft open points placement using Archimedes optimization algorithm", *Ain Shams Engineering Journal*, vol. 12, issue 2, pp. 1923-1941, 2021.
- [15] W. Aribowo, S. Muslim, B. Suprianto, and S. Haryudo, "Intelligent Control of Power System Stabilizer Based on Archimedes Optimization Algorithm – Feed Forward Neural Network", *International Journal of Intelligent Engineering and Systems*, vol. 14, no.3, 2021.
- [16] J. L. J. Pereira, M. B. Francisco, S. S. da Cunha Jr., and G. F. Gomes, "A powerful Lichtenberg Optimization Algorithm: A damage identification case study", *Engineering Applications of Artificial Intelligence*, vol. 97, no. 104055, 2020.
- [17] M. Challan, S. Jeet, D. K. Bagal, L. Mishra, A. K. Pattanaik, and A. Barua, "Fabrication and mechanical characterization of red mud based Al2025-T6 MMC using Lichtenberg optimization algorithm and Whale optimization algorithm", *Materials Today: Proceedings*, vol. 50, part 5, pp. 1346-1353, 2022.
- [18] M. B. Francisco, D. M. Junqueira, G. A. Oliver, J. L. J. Pereira, S. S. da Cunha Jr., and G. F. Gomes, "Design optimizations of carbon fibre reinforced polymer isogrid lower limb prosthesis using particle swarm optimization and Lichtenberg algorithm", *Engineering Optimization*, vol. 53, issue 11, 2021.
- [19] B. S. Yıldız, A. R. Yıldız, N. Pholdee, S. Bureerat, S. M. Sait, and V. Patel, "The Henry gas solubility optimization algorithm for optimum structural design of automobile brake components", *Journal for Materials Testing*, vol. 62, no. 3, pp. 261-264, 2020.
- [20] S. Ekinici, D. Izci, and B. Hekimoğlu, "Henry Gas Solubility Optimization Algorithm Based FOPID Controller Design for Automatic Voltage Regulator", *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1-6, 2020.
- [21] S. Ekinici, B. Hekimoğlu, and D. Izci, "Opposition based Henry gas solubility optimization as a novel algorithm for PID control of DC motor", *Engineering Science and Technology, an International Journal*, vol.24, issue 2, pp. 331-342, 2021.
- [22] J. L. J. Pereira, M. B. Francisco, C. A. Diniz, G. A. Oliver, S. S. Cunha Jr., and G. F. Gomes, "Lichtenberg algorithm: A novel hybrid physics-based metaheuristic for global optimization", *Expert Systems with Applications*, vol. 170, no. 114522, 2021.
- [23] F. A. Hashim, K. Hussain, E. H. Houssein, M. S. Mabrouk, and W. Al-Atabany, "Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems", *Appl Intell*, vol. 51, pp. 1531–1551, 2021.
- [24] F. A. Hashim, E. Houssein, M. S. Mabrouk, W. Al-Atabany, and S. Mirjalili, "Henry gas solubility optimization: A novel physics-based algorithm", *Future Generation Computer Systems*, vol. 101, pp. 646-667, 2019.
- [25] R. Concepcion, S. Lauguico, J. Alejandrino, J. De Guia, E. Dadios, and A. Bandala, "Aquaphotomics determination of total organic carbon and hydrogen biomarkers on aquaponic pond water and concentration prediction using genetic programming," *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*, 2020.
- [26] M. G. Palconit, M. Pareja, A. Bandala, J. Espanola, R. R. Vicerra, R. Concepcion, E. Sybingco, and E. Dadios, "FishEye: A Centroid-Based Stereo Vision Fish Tracking Using Multigene Genetic Programming," *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, 2021.
- [27] R. Concepcion, S. Lauguico, J. Alejandrino, E. Dadios, E. Sybingco, and A. Bandala, "Aquaphotomics Determination of Nutrient Biomarker for Spectrophotometric Parameterization of Crop Growth Primary Macronutrients Using Genetic Programming," *Information Processing in Agriculture*, 2021.
- [28] I. Petrasova, P. Karban, P. Kropik, D. Panek, and I. Dolezel, "Optimization of selected operation characteristics of array antennas," *Journal of Computational and Applied Mathematics*, vol. 399, no. 113726, 2022.
- [29] S. N. K. Koc, and A. Koksall, "Wire antennas optimized using genetic algorithm," *Computers & Electrical Engineering*, vol. 37, no. 6, pp. 875-885, 2011.
- [30] R. Concepcion, B. Duarte, A. Bandala, J. Cuello, R. R. Vicerra, and E. Dadios, "Characterization of Potassium Chloride Stress on Philippine Vigna radiata Varieties in Temperature-stabilized Hydroponics Using Genetic Programming," *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2021.
- [31] V. Hosseini, Y. Farhang, K. Majidzadeh, and Ch. Ghojadi, "Customized mutated PSO algorithm of isolation enhancement for printed MIMO antenna with ISM band applications," *AEU – International Journal of Electronics and Communications*, vol. 145, no. 154067, 2022.
- [32] R. Concepcion, E. Dadios, A. Bandala, J. Cuello, and Y. Kodama, "Hybrid Genetic Programming and Multiverse-based Optimization of Pre-Harvest Growth Factors of Aquaponic Lettuce Based on Chlorophyll Concentration," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 6, pp. 2128-2138, 2021.
- [33] R. Mehta, "Optimal receive beamforming in spatial antenna diversity system using evolutionary genetic algorithm," *Array*, vol. 10, no. 100053, 2021.

Employing Nonlinear Model based PI like Controller for the Time Varying System

Samrat Banerjee¹, Atanu Panda², Indrajit Pandey³, Parijat Bhowmick⁴

^{1,3}Dept. of Appl. Electronics and Instrumentation Engg. Techno International New Town, Kolkata, India.

²Dept. of Electronics and Communication Engg., IEM, Kolkata, India.

⁴Dept. of Electronics and Electrical Engg. IIT Guwahati, India.

samrat.banerjee@tict.edu.in¹, atanu.686@gmail.com², indrajit.pandey@tict.edu.in³, parijat.bhowmick@iitg.ac.in⁴

Abstract — This work outlines an adaptive non-linear model-based PI like control scheme for the spherical tank level system. The efficiency of the model-based PI like control law is implemented by conducting simulation studies on the spherical tank system. The servo compliance with above said control problem exhibit satisfactory performances. Simultaneously, model state(s) and the influential model parameter(s) were estimated using unscented Kalman filter algorithm and the updated state values of the process model were employed to derive the control strategy. From the effective analysis, it has been observed that the model-based PI like control scheme utilizing non-linear Kalman filter strategy offers better performance over the traditional non-linear model-based PI like controller.

Keywords — Spherical tank, Non-linear model based PI like control, Derivative free Kalman filter, Parameter estimation.

1. INTRODUCTION

Adaptive control schemes like Gain-scheduling (GS), Self-tune-control (STR) and Model-reference-adaptive-control (MRAC) are the well-recognized strategies to control different plants, in which parameter(s) value changes with time. The plethora of designing technique based on closed-loop control performance and robustness criteria are currently available to design the above mentioned adaptive control schemes (Astrom and Wittenmark, 2011).

Literature survey reveals that the following control schemes have been already designed and implemented on the variable area tank systems (e.g. conical or spherical tank): gain scheduled adaptive controller (Astrom and Wittenmark, 1995; Anandanatarajan et al. 2006 and Greg McMillan, 2010). Robust STR (Tan et al. 2001), fuzzy logic controller (Sakthivel et al. 2008), optimal controller using dynamic programming (Bhuvanewari, 2009) and neural network based adaptive control scheme (Bhadra et al 2019, Panda and Panda 2018). To the best of our knowledge, adaptive non-linear model-based PI like control law for controlling liquid level in the spherical tank system has not been reported in the literature.

An optimization-based solution technique would necessarily be devoted to enhance controller performance and is therefore highly desirable. Designing prediction-based control algorithm relies under optimal trade-offs between objective derived on the basis of performances indices and the computational complexity (J.E.G. Refsnes, 2007). Since,

‘controller-observer’ type of combined logic fairly deals with intrinsic robust features in a straightforward manner (Wang et al. 2014) To construct a stable estimator for the continuous-time dynamical system, Kalman filter (KF) would be well suited with/without presence of the plant uncertainty (Subbotina & Tokmantsev, 2014).

Motivated by the increasing need to develop state-of-art ‘direct way of NMBC’ and ‘prediction logic-based’ control implementation for both SISO/MIMO types of processes, a newly developed control strategies were illustrated, taking into account effectiveness, easily deployable, reliability and robustness. Due to difficulties of obtaining measured values of the process parameter(s) from the real plants in several scenarios, an online estimation technique would necessary be exploited to synthesize the estimated parameter(s) value (Liu et al. 2014, Morari and Lee, 1999). The proposed attempt leads to a control framework, where predicted values of the effective model parameters (using non-linear KF approaches) were taking into consideration and the values of the model state(s) and the controller gain were derived implicitly by predicted values of the process parameter(s). Finally, the predicted and the corrected part of the manipulated variable have been determined.

The remainder of this article is given as: A detailed of the literatures and the motivational section was offered in section 1. Section 2 deals with process description of the spherical tank system. The control strategy has been described in section 3. Simulation based realistic results are reported in section 4, Section 5 concludes with an effective analysis.

2. PROCESS DESCRIPTION

The processes considered for the simulation study are spherical tank level process. The first principle equation for the spherical area tank system considered for the simulation study is given below:

$$A(h) \frac{dh}{dt} = F_{in} - F_{out} \quad (1)$$

Here, $A(h)$ indicate area of the tank. (F_{in} or q) and F_{out} signifies inflow and outflow rate of the tank respectively. h is the liquid level. The relationship exhibits between outlet rate (F_{out} or q_{out}) and the liquid height (h) in this tank is mentioned below:

$$F_{out} = c_v \sqrt{2gh} \tag{2}$$

Here, c_v specifies valve co-efficient. Note that, area of the variable tank varies with respect to liquid height in this tank. The area of the tank is computed using the below equation:

$$A(h) = \pi [2rh - h^2] \tag{3}$$

Here, r indicates maximum radius of the tank. Maximum radius and height are chosen as 30 cm and 60 cm respectively. The schematic diagram of the spherical tank process is presented in figure1. Disturbance can be introduced either through changing of downstream valve or adding liquid through q_0 .

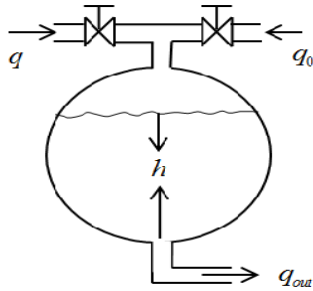


Figure1: Schematic diagram of spherical tank system

3. PROPOSED ADAPTIVE NON-LINEAR MODEL-BASED PI LIKE CONTROLLER

This work, introduces a non-linear model-based PI like controller to control the liquid-level in the spherical tank system. The control algorithm is formulated as:

$$F_{in}(i) = \alpha * \theta_{k_c}(i) [h_{sp}(i) - h(i)] + \theta_{bias}(i) \tag{4}$$

$$\theta_{k_c}(i) = \theta_{bias}(i) / \hat{h}(i|i-1) \tag{5}$$

Here, h_{sp} indicates set-point. $\hat{h}(i|i-1)$ is the one step ahead predicted height. Controller gain (θ_{k_c}), bias (θ_{bias}) and (α)

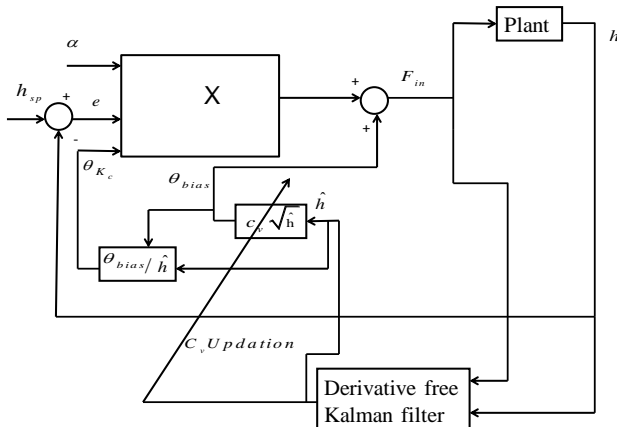


Figure2: Schematic diagram of the developed adaptive non-linear model-based PI like controller.

specifies parameters of the developed control methodology. The pictorial representation of the said control law is presented in figure2. The system equations with model parameters are as follows:

$$\hat{\theta}(i) = \hat{\theta}(i-1) + w(i-1) \tag{6}$$

$$\hat{\theta}(i) = [c_v] \tag{7}$$

$$h(i) = F[h(i-1), F_{in}(i-1), \hat{\theta}(i-1)] + v(i) \tag{8}$$

'h' represents measured output of the plant. 'w' and 'v' are zero mean, independent, Gaussian process noise and measurement noise respectively. 'Q', 'R' signifies noise covariance matrix of the process and measurement noise respectively. The predicted model parameter estimates can be computed as

$$\hat{\theta}(i|i-1) = \hat{\theta}(i-1|i-1) \tag{9}$$

The co-variance matrix with estimated errors in the predicted model parameter estimates are determined as:

$$P_{\theta}(i|i-1) = P_{\theta}(i-1|i-1) + Q \tag{10}$$

Here, the set of (2L+1) sigma-points with the associated weights $w(k)$ are chosen symmetrically about $\hat{\theta}(i|i-1)$ as:

$$\theta_s = [\hat{\theta}(i|i-1), \hat{\theta}(i|i-1) + \sqrt{(L+\kappa)P_{\theta}(i|i-1)}, \hat{\theta}(i|i-1) - \sqrt{(L+\kappa)P_{\theta}(i|i-1)}] \tag{11}$$

Since, measurement-prediction ($\hat{h}(i|i-1)$), computation-of-innovation ($e_{i|i-1}$) and covariance-matrix with innovation ($P_{ee}(i)$) and the cross-covariance-matrix between the predicted parameter estimate errors and innovation ($P_{\theta_e}(i)$) are computed as:

$$\hat{h}^i(i|i-1) = F(h(i-1), F_{in}(i-1), \theta^i(i|i-1)) \tag{12}$$

$$\hat{h}(i|i-1) = \sum_{i=0}^{2L} w_i (\hat{h}^i(i|i-1)) \tag{13}$$

$$e_{i|i-1} = \hat{h}(i) - h(i) \tag{14}$$

$$P_{ee}(i) = \sum_{i=0}^{2L} w_i (\hat{h}^i(i|i-1) - h_r(i|i-1))(\hat{h}^i(i|i-1) - h_r(i|i-1))^T + R \tag{15}$$

$$P_{\theta_e}(i) = \sum_{i=0}^{2L} w_i (\hat{\theta}_s^k(i|i-1) - \hat{\theta}(i|i-1))(\hat{h}^i(i|i-1) - h_r(i|i-1))^T \tag{16}$$

Where, $w_0 = \kappa / (L + \kappa)$ and $w_i = 1 / (2(L + \kappa))$. Here, κ is a tuning parameter. The Kalman-gain is derived as:

$$K = P_{\theta_e} (P_{ee}^{-1}) \tag{17}$$

The updated values of the model-parameter estimates are determined using following relation:

$$\hat{\theta}(i|i) = \hat{\theta}(i|i-1) + K e_{i|i-1} \tag{18}$$

The covariance-matrix of the estimated-errors with updated model-parameter estimation is derived by:

$$P_{\theta}(i|i) = P_{\theta}(i-1|i-1) - K P_{ee} K^T \tag{19}$$

In case of non-linear model based PI like controller, the control logic can be derived using Equation (4-5).

4. RESULTS AND DISCUSSION

1) Open loop study:

To identify the open-loop responses of the spherical-tank system, combination of positive and negative steps in the controller output were introduced. Figure 3(a) shows the step wise variation in the manipulated variable (f_{in}). The changes in the process variable (h) is depicted in figure 3(b).

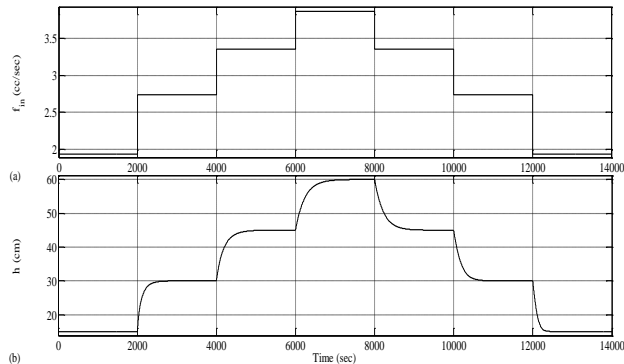


Figure3. Open-loop assessment of the spherical-tank level system: (a) change in manipulated variable, (b) process variable.

2) Servo-Regulatory response of spherical tank system with non-linear model based PI like controller:

To identify the servo performance and disturbance attenuation capability with non-linear model-based PI like control scheme, stepwise variation in the downstream valve co-efficient as reported in figure 4(c) was propounded. The servo-regulatory response of the spherical-tank level system with non-linear model-based PI like controller is portrayed in figure 4(a). From figure 3(a), it can be observed that the non-linear model-based PI like control method is capable of rejecting disturbances and also able to track the predetermined set value. Evolution of manipulated variable is depicted in figure 4(b). From the simulation study; we can conclude that the disturbance response of the proposed PI like control technique is found to be very sluggish in nature.

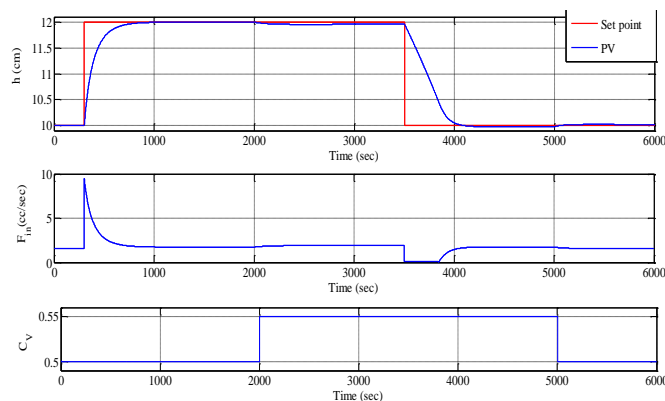


Figure4. Servo-regulatory compliance of the spherical-tank level process with non-linear model-based PI like controller (a) process variable, (b) manipulated variable, (c) variation in downstream valve co-efficient (PV: process variable).

3) Servo-regulatory compliance of the spherical-tank level process with proposed adaptive non-linear model-based PI like controller using UKF:

To identify the servo performance and disturbance elimination ability of the proposed adaptive non-linear model-based PI like controller, stepwise variations with downstream valve co-efficient as depicted in figure 5(c) was propounded. Table1 provides values associated with Kalman filter-based parameter estimation technique. The servo-regulatory

Table1. Filter specifications associated with EKF/EnKF/UKF based parameter estimation schemes employing on spherical tank system

Parameter	Value
Measurement noise covariance (R)	1e-8
Process noise covariance (Q)	1e-8
α_1, β and κ	1, 0 and 0
Initialization of parameter (b_0) i.e., $\hat{\Theta}(0 0)$	0.5
Initial error covariance $P(0 0)$	0.25

performance of the spherical-tank level system with adaptive non-linear model-based PI like controller utilizing UKF is presented in figure 5(a). From figure 5(a), it can be concluded that the proposed control law is capable of rejecting disturbances and also able to track the desired level. However, the controller preserves sluggish responses while attenuating disturbances. The variation in manipulated variable is depicted in figure 5(b). From the simulation study; we can conclude that the disturbance response of the non-linear model-based PI like control method is found to be satisfactory. Figure 5(c) and 5(d) shows evolution of the true and estimated value of the downstream valve co-efficient respectively.

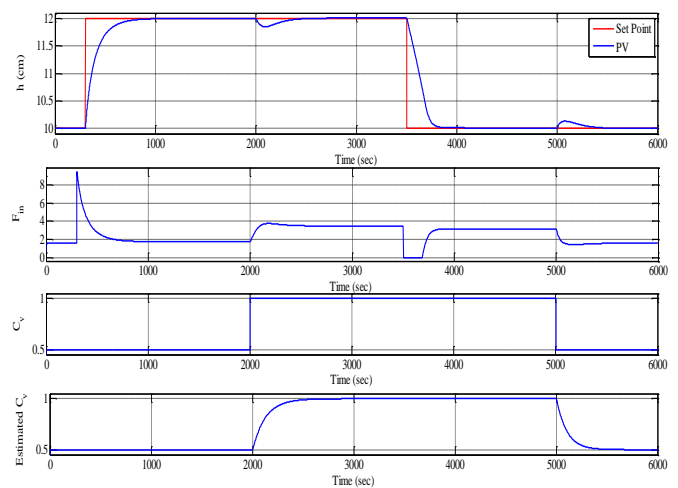


Figure5. Servo-Regulatory compliance of the spherical-tank level process with proposed adaptive non-linear model-based PI like controller (a) process variable, (b) manipulated variable, (c) evolution of downstream valve position, (d) estimated c_v .

3) Performance assessment of different controllers:

In order to assess qualitative performances with different level of control actions like servo-regulatory compliance, a performance-based chart like computation time (CT, Table 2), MSE (see Table 3)) were provided. Table 2 shows that proposed adaptive non-linear model based PI like controller using UKF strategy performs better over non-linear model based PI like control approach. As lesser MSE value resulted faster convergence and better accuracy, hence it can be embodied that newly developed control strategy provides faster set point tracking capabilities compared to the traditional model based PI control approach.

Table2. CT per sampling instant with different control schemes on spherical tank system.

Control scheme	Servo-Regulatory (F_{in})
Proposed scheme	1e-02 – 2.5
non-linear model based PI like controller	1e-02 – 2.9

Table3. MSE chart for different control schemes on spherical tank system.

Control scheme	Servo-regulatory
Proposed	5.0217e-05
non-linear model based PI like controller	8.6294e-05

5. CONCLUSION

In this paper, we have successfully designed and implemented an adaptive non-linear model-based PI like control algorithm. From the extensive study on the simulated models of the spherical tank level system, it can be inferred that the servo-regulatory performances of the proposed adaptive non-linear model-based PI like control strategy preserved satisfactory results compared to the traditional non-linear model-based PI like control framework. Further work is in progress to validate the proposed adaptive control algorithm on the experimental spherical tank test rigs available at the process control lab in the Department of Electronics & Communication Engineering, Institute of Engineering and Management, Kolkata.

REFERENCES

[1] Karl Johan Astrom and Bjorn Wittenmark, Adaptive Control Second Edition, Pearson Education, Inc., New York, 1995, pp. 392–398.
 [2] Karl Johan Astrom and Bjorn Wittenmark, A survey of adaptive control application, In Proc. 34th IEEE conference on Decision and Control, New Orleans, Louisiana, January 1995.
 [3] R. Anandanatarajan, M. Chidambaram, T. Jayasingh, Limitations of a PI controller for a first-order nonlinear process with dead time, ISA Trans., 45(2):185–199.
 [4] Greg McMillan, Sridhar Dasani and Prakash Jagadeesan, Adaptive Level of control, Control Magazine, February 2011.
 [5] K. K. Tan, S. N. Huang, H. F. Dou, T. H. Lee, S. J. Chin, S. Y. Lim, Adaptive robust motion control for precise trajectory tracking applications. ISA Trans. 40(1): 57–71.

[6] R Sakhivel, K Mathiyalagan and S Marshal Anthoni, Design of a Passification controller for uncertain fuzzy Hopfield neural networks with time-varying delays, Physica Scripta 84(4):045024.
 [7] N.S. Bhuvanewari, G.Uma, T.R. Rangaswamy, “Adaptive and optimal control of a non-linear process using intelligent controllers”, Appl. Soft Comput., 9(1):182-190.
 [8] S. Bhadra, A. Panda, P. Bhowmick, S. Goswami, R.C. Panda (2019), Design and application of nonlinear model-based tracking control schemes employing DEKF estimation, Optim. Control Appl. Meth., 40(5): 938–960.
 [9] Y.J Liu, D.J. Li and S Tong, Adaptive output feedback control for a class of nonlinear systems with full-state constraints (2014), Int. J. Cont., 87(2):281–290.
 [10] Morari M, Lee J. H. (1999), Model predictive past, present and future, Comp. Chem. Eng., 23(4-5):667–682.
 [11] J.E.G. Refsnes, Nonlinear Model-Based Control of Slender Body AUVs (2007), Ph.D. Dissertation, NUST, Trondheim.
 [12] N.N. Subbotina and T.B. Tokmantsev (2014), Optimal Synthesis to Inverse Problems of Dynamics, 19th World Congress of IFAC, Cape Town.
 [13] P. Wang, C. Yang, X. Tian and D. Huang (2014), Adaptive nonlinear model predictive control using an on-line support vector regression updating strategy, Chin. J. Chem. Eng., 22(7):774–781.
 [14] A. Panda, R.C. Panda, (2018), Adaptive nonlinear model-based control scheme implemented on the nonlinear processes, Nonlinear Dyn., 91(4):2735-2753.

A Novel Multifaceted Deep Learning-Based Mobile Application for Accurate and Efficient Waste Classification and Increased Composting Engagement in Communities

Samyak Shrimali
Jesuit High School
Portland, Oregon, USA
samyak.shrimali12@gmail.com

Abstract— Solid food waste is slowly accumulating around the world in landfills. This waste is hazardous to human health and our environment when it decomposes as it leads to widespread release of greenhouse gasses such as CO₂ and methane. Composting household waste actively can put wasted food to good use by creating arable soil and help mitigate the waste crisis that causes climate change. But currently, the general public finds it hard to manage and classify exactly what item is compostable and non-compostable. Furthermore, there is lack of motivation, incentive, and community support for composting waste actively. This paper proposes CompostAI, a novel deep learning-based mobile application targeted to make community participation in composting easy and socially engaging. This application uses a Xception convolutional neural network (CNN) model to classify waste into seven categories: compost, paper, cardboard, glass, metal, trash, plastic. The Xception model demonstrated optimal performance as it had the highest accuracy of 78.43% and F1 score of 81.22 out of the six CNN model trained, validated, and tested. CompostAI also has supplemental features that include allowing users to announce local sustainability events, find nearby composting centers, and learn new sustainable living techniques. CompostAI successfully makes community participation in composting easy and socially engaging, increasing composting rates and mitigating the detrimental waste crisis that leads to climate change.

Keywords—Artificial Intelligence, Convolutional Neural Networks, Waste Classification, Image Classification, Societal Computing

I. INTRODUCTION AND LITERATURE REVIEW

In the United States of America, more than 80 billion pounds of food waste is generated each year and more than 87% of this waste ends up in landfills. This leads to the release of over 3.3 billion metric tons of carbon dioxide [1]. Furthermore, this waste decomposes and releases methane into our atmosphere which has 21 times more warming potential than carbon dioxide [2]. This waste crisis is a prominent catalyst of climate change and can be easily mitigated if the general public composts their household waste actively. Composting can help divert food waste from landfills and help mitigate climate

change, while fostering environmental and economic benefits [3].

The primary reasons the general public does not compost is because of the hurdle of distinguishing between compostable/non-compostable material and lack of motivation, incentive, and community support. It is hard for people to read and understand often incomplete signs and boards near waste bins which cover only a few examples of the compostable/non-compostable material. Current practices for the general public to distinguish between compostable and non-compostable material are based upon significant research and support from the world wide web and local composting centers, this is a very labor some, manual, and error-prone process [4]. The general public is in dire need of an easy-to-use tool that can efficiently and accurately distinguish between compostable and non-compostable materials and also motivate them to compost regularly

A few researchers have approached the problem of waste classification using CNN image recognition models. Thung et al. designed TrashNet to classify waste into five classes of recycle and waste. Their model did not achieve a high accuracy; their accuracy was 63% showcasing high prediction error [5]. Researchers from the Indian Institute of Technology designed SpotGarbage uses GarbNet, a custom CNN, to identify garbage on streets. Instead of classifying waste into specific sectors, GarbNet just identified the presence of waste [6]. Both, TrashNet and GarbNet do not classify compostable material or food waste but are examples of potential of CNNs for waste classification applications.

Researchers like Salimi et al. and Gupta et al., have approached this problem in novel way by designing a smart disposal bin which uses several expensive industrial-grade sensors and hardware to classify waste as trash or plastic, but it is ineffective for large scale implementation around the globe [7] [8].

Smartphones have not been previously considered for this application but can be as since they provide high resolution displays, efficient computing power, and a built-in set of accessories (ex. cameras, speakers, etc.). They are also widely accessible, as of 2022, there are over 6.6

billion smartphone users around the world, and this number is projected to keep increasing [9]. The combined factors of widespread smartphone usage, access to high-definition cameras and computing processors in mobile devices lead to a situation where waste classification based on automated image recognition, if technically feasible, can be made available at an unprecedented scale and work to help mitigate climate change.

II. SOLUTION AND METHODOLOGY

This paper presents CompostAI, a novel deep learning-based mobile application targeted to make community participation in composting easy and socially engaging. This application will use state of the art convolutional neural network (CNN) and image classification techniques to identify an object into seven compostable (compost, paper, cardboard) and non-compostable (glass, metal, trash, plastic) classes. It will provide users with a platform to announce local sustainability events, find nearby composting centers, and learn new sustainable living techniques (such as self-composting, planting trees, etc.).

The development of this application was split into multiple phases. In phase one, an efficient and accurate deep-learning model was designed through a comparative analysis of several state-of-the-art convolution neural networks. To find and develop an optimal model, first, an image dataset was obtained and pre-processed. This dataset was acquired from the UC Santa Cruz data repository and consisted of 2,678 images referring to seven classes: compost, paper, cardboard, glass, metal, trash, plastic [10]. The goal was to determine the proper waste class for an item from an image and then classify the item as compostable/non-compostable.

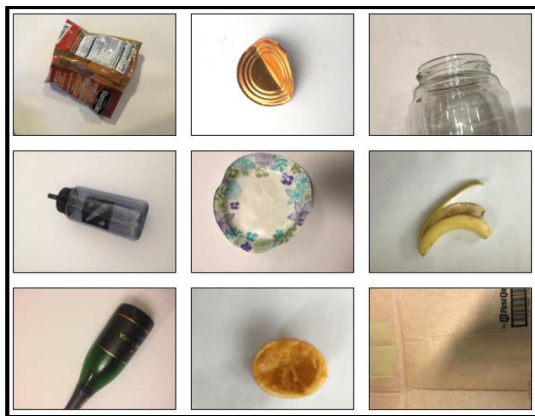


Fig. 1. Example images of a compostable and non-compostable items from the UC Santa Cruz waste dataset [10]

The images were originally 400x300 pixels and using the IMREAD and IMRESIZE functions of the OpenCV python library, they were resized to 224x224 pixels for the CNNs. The dataset was split into a ratio of 80%:15%:5% for training, validation, and testing, respectively.

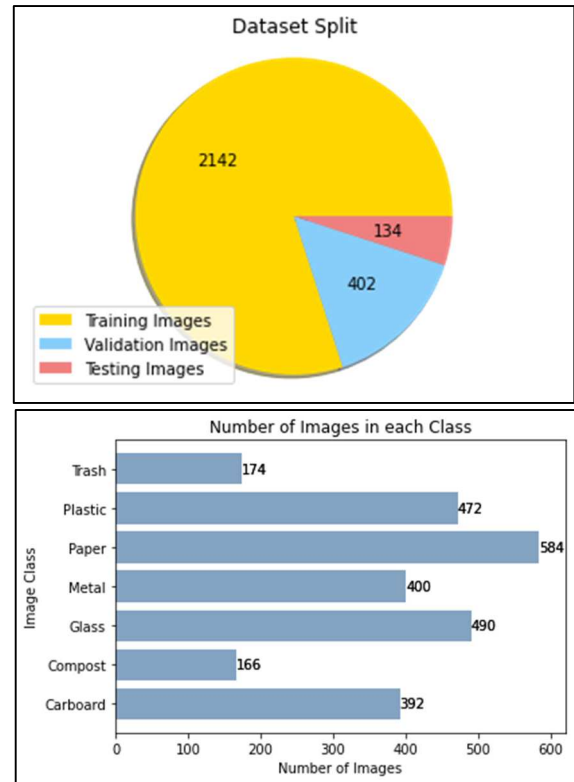


Fig. 2. Visual representation of the dataset split and number of images in each class

After the image dataset was ready, six different CNNs were designed to explore which led to efficient and accurate waste classification. Each CNN was based on a unique model architecture and varied in size. The CNNs used for the comparative analysis were: Xception, VGG-19, InceptionV3, InceptionResNetV2, MobileNetV2, NASNetMobile [11]. These are considered six of the most accurate, size-efficient, and effective model architectures for vision tools and multiclass classification problems. All these CNNs were initialized with the ImageNet dataset and used pretrained weights [11].

Model Comparison							
Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Time (ms) per inference step (CPU)	Time (ms) per inference step (GPU)
Xception	88	0.79	0.945	22.9M	81	109.4	8.1
VGG-19	549	0.713	0.9	143.7M	19	84.8	4.4
InceptionV3	92	0.779	0.937	23.9M	189	42.2	6.9
InceptionResNetV2	215	0.803	0.953	55.9M	449	130.2	10
MobileNetV2	14	0.713	0.901	3.5M	105	25.9	3.8
NASNetMobile	23	0.744	0.919	5.3M	389	27	6.7

Fig. 3. Comparison chart of the six CNNs considered for a comparative analysis [11]

The models used convolutional layers to extract specific features from the images for identification, max polling layers to down sample the size of the training/validation data, flatten layers to shorten the dimensions of the input, and dense layers to align feature-based predictions to certain categories. To ensure that the models did not pick up the noise between images, additionally to their architectures, dropout and regularization layers were added to monitor their speed of learning and restrict them from overfitting on the dataset.

Through the usage of the OpenCV library, due to the fairly limited dataset size, images were augmented during the training process. Each image was rotated an angle of 0, 45, 90, 180, and 270 degrees, reflected across the x/y axis, cropped, and blurred [12]. This ensured that the model would be able to extract features for a number of different scenarios and improve its downstream performance.

Xception [13] had the highest classification accuracy of 78.4% out of the five other CNNs trained, validated, and tested in this process, therefore, it was chosen as the most optimal model for CompostAI's mobile application.

In phase two, the Xception model was converted from a h5 file to a tflite file for usage on mobile devices using the TensorFlow Lite library. This tflite model was then deployed onto the Android Studio developing environment where it was integrated to make a camera-based classification algorithm.

CompostAI is multifaceted and consists of a number of unique features for its users. In phase three, its internal features and user interface were developed. CompostAI's main feature is that it allows users to scan an item and determine if it is compost, paper, cardboard, glass, metal, trash, or plastic, and compostable/non-compostable using the Xception CNN model. CompostAI's supplemental features include allowing users to announce local sustainability events, find nearby composting centers, and learn new sustainable living techniques. Figure 4 shows images of CompostAI's user interface and its various application features



Fig. 4. Screenshots from CompostAI's mobile application user-interface

III. RESULTS

CompostAI has been through multi-level testing and data analysis. In the initial CNN comparative analysis stage, six different CNNs were designed, trained, validated, and tested. The Xception model was chosen as the most optimal model for CompostAI's waste classification feature as it had the highest classification accuracy of 78.43% and F1 score of 81.22.

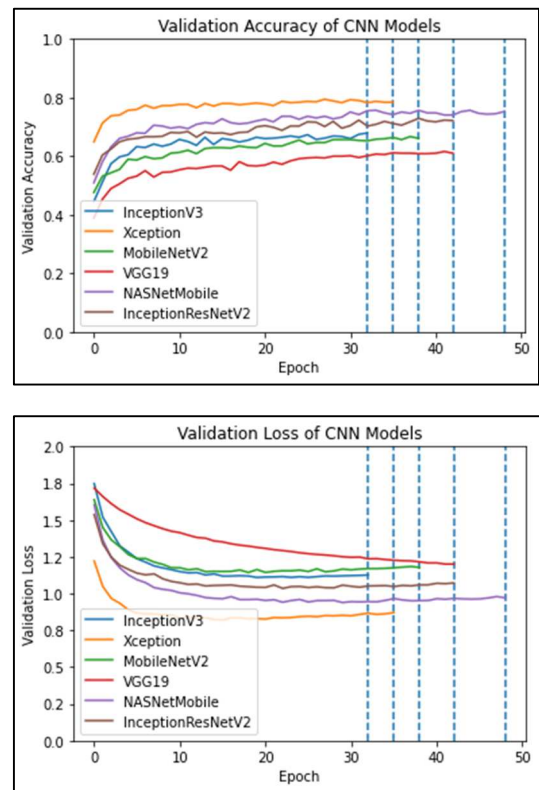


Fig. 5. Comparison of each CNNs validation accuracy and validation loss

In addition to validation accuracy and loss, precision, recall, and F1 score were also used to evaluate each CNN

model. Figure 6 shows the different evaluation metric values considered to determine the most optimal CNN.

A. *Precision*: True positive predictions divided by the total number of positive predictions, helps indicate the quality of a positive prediction made by the CNN model.

$$Precision = \frac{TP}{TP + FP}$$

B. *Recall*: True positive predictions divided by the number of true positive predictions and false negative predictions, helps indicate the quality of the CNN in identifying true positives.

$$Recall = \frac{TP}{TP + FN}$$

C. *F1 Score*: Harmonic mean of Precision and Recall, helps indicate the tradeoff between the CNN model's quality in making positive predictions and in identifying true positive

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Convolutional Neural Network - Evaluation Metrics					
Model Architecture	Validation Accuracy	Validation Loss	Precision	Recall	F1 Score
InceptionV3	77.91	1.126	75.23	78.32	77.21
Xception	78.43	0.807	79.92	81.75	81.22
MobileNetV2	66.1	1.178	64.63	67.32	66.83
VGG19	61.02	1.205	60.23	62.74	61.62
NASNetMobile	75.23	0.971	74.86	75.93	75.78
InceptionResNetV2	72.18	1.072	74.28	70.59	73.12

Fig. 6. Evaluation metrics used to determine the most optimal CNN for waste classification

After the Xception model was integrated with CompostAI and the application's user interface and internal algorithms were developed, real-time testing was conducted to further validate CompostAI's waste classification feature. 45 unique items were brought for diagnosis and CompostAI could predict that 41 of them were compostable/non-compostable correctly.

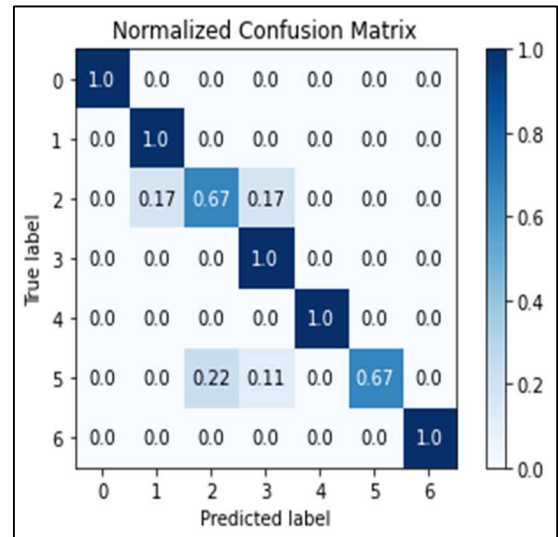


Fig. 7. Normalized confusion matrix of the results from the 45 real-time waste classifications

CompostAI was also presented to 25+ field experts at local composting centers and as a whole, they stated CompostAI significantly aids the general public in waste classification and will ensure they will compost regularly. They also stated that it has the ability to immensely mitigate the waste crisis and climate change.

IV. CONCLUSION

This paper presented CompostAI, a smart solution to make community participation in composting easy and socially engaging to all-in-all increase composting rates and mitigate the detrimental waste crisis that leads to climate change. In conclusion, through comprehensive testing, analysis of results, and expert feedback, CompostAI was a success. The waste classification feature can accurately and efficiently differentiate between more than 30 common compostable and non-compostable items. It is based on an effective and thoroughly trained, tested, and validated neural network that was designed through a comparative analysis of six different convolutional neural networks, each based on a unique model architecture. The chosen Xception model has a high classification accuracy of 78.43% and F1 score of 81.22 and works extremely well as validated with real-time testing. CompostAI's mobile application is thorough and comprehensive, it enables users with composting resources, provides community connection for sustainable choices, and allows users to educate themselves on environmental topics in a simple manner. CompostAI mitigates all the problems with current solutions and creates direct positive human impact.

Future plans for CompostAI include partnering with governments and other nonprofit organizations to enable a rewarding system that reward users for their composting contributions with monetary prizes and create a self-sustaining cycle.

ACKNOWLEDGEMENT

The author would like to thank his family, teachers, and mentors for their invaluable support during his research journey. He would also like to thank the various industry experts who were able to provide him with valuable feedback on his work during the testing and validation phase.

REFERENCES

- [1] "Food Waste in America in 2022: Statistics & Facts." *Recycle Track Systems*, <https://www.rts.com/resources/guides/food-waste-america/>.
- [2] "Basic Information about Landfill Gas." *EPA*, US Environmental Protection Agency, <https://www.epa.gov/lmop/basic-information-about-landfill-gas>.
- [3] "Reducing the Impact of Wasted Food by Feeding the Soil and Composting." *EPA*, US Environmental Protection Agency, <https://www.epa.gov/sustainable-management-food/reducing-impact-wasted-food-feeding-soil-and-composting>.
- [4] Wendeln, Jen. "I Want to Compost, but...!" *Sustainable America*, 3 June 2014, <https://sustainableamerica.org/blog/i-want-to-compost-but/>.
- [5] Thung, Gary, and Mindy Yang. *Classification of Trash for Recyclability Status*. Stanford CS229 Project Report, 2016, <http://cs229.stanford.edu/proj2016/report/ThungYang-ClassificationOfTrashForRecyclabilityStatus-report.pdf>.
- [6] Mittal, Gaurav, et al. *SpotGarbage: Smartphone App to Detect Garbage Using Deep Learning*. 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 12 Sept. 2016, <https://dl.acm.org/doi/proceedings/10.1145/2971648>.
- [7] Salimi, Irfan, et al. *Visual-Based Trash Detection and Classification System for Smart Trash Bin Robot*. International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), 2018, Oct. 2018, https://www.researchgate.net/publication/329164139_Visual-based_trash_detection_and_classification_system_for_smart_trash_bin_robot.
- [8] Gupta, Tanya, et al. "A Deep Learning Approach Based Hardware Solution to Categorize Garbage in Environment." *SpringerLink*, Springer International Publishing, 19 Nov. 2021, <https://link.springer.com/article/10.1007/s40747-021-00529-0>.
- [9] Gaubys, Justas. *How Many People Have Smartphones?* Oberlo, <https://www.oberlo.com/statistics/how-many-people-have-smartphones#:~:text=Latest%20figures%20show%20an%20increasing,rate%20is%20at%2045.4%20percent>.
- [10] Frost, Sarah, et al. *CompostNet Dataset*. UC Santa Cruz, 15 Oct. 2019, <https://github.com/sarahmfrost/compostnet>.
- [11] "Keras Documentation: Keras Applications." *Keras*, <https://keras.io/api/applications/>.
- [12] Agarwal, Vardan. "Complete Image Augmentation in OpenCV." *Medium*, Towards Data Science, 16 May 2020, <https://towardsdatascience.com/complete-image-augmentation-in-opencv-31a6b02694f5>.
- [13] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." *ArXiv.org*, 4 Apr. 2017, <https://arxiv.org/abs/1610.02357>.

Remote Elemental Analysis System for Liquid using sonoluminescence

Sardini Sayidatun Nisa Sailellah
 Department of Mechanical
 Engineering
 Tokyo Institute of Technology
 Tokyo, Japan
 nisa.s.aa@m.titech.ac.jp

Hideharu Takahashi
 Laboratory for Zero-Carbon
 Energy
 Tokyo Institute of Technology
 Tokyo, Japan
 takahashi.h.av@m.titech.ac.jp

Hiroshige Kikura
 Laboratory for Zero-Carbon
 Energy
 Tokyo Institute of Technology
 Tokyo, Japan
 kikura.h.aa@m.titech.ac.jp

Abstract—The severe accident that occurred in 2011 brought a significant concern to the world, especially in nuclear reactor technology, one of which is decommissioning. Given the radioactive contamination, the decommissioning process requires extensive information, including the knowledge of the chemical composition of the liquid inside the reactor post-accident. The harsh environment limits the choice of technology. The observation must be done using a remote technique. This work discusses the development of a remote elemental analysis technique to visualize as well as analyze the chemical composition of the liquid inside the reactor, using ultrasound technique, namely sonoluminescence. Spectral analysis of alkali metal and alkaline earth solutions (such as NaCl, SrCl₂, and CaCl₂) was conducted to confirm the feasibility of remote elemental analysis using sonoluminescence. The optical system using fiber optic and spectral analysis methods enables remote analysis. The simultaneous confirmation using a camera and developed image analysis algorithm based on the HSV method was conducted to confirm the presence of sonoluminescence.

Keywords—Remote analysis, sonoluminescence, ultrasound, spectrum analysis

I. INTRODUCTION

Post-Fukushima Nuclear Power Station accident in 2011, more attention was put on the post-accident treatment: removing spent fuel, identifying and retrieving fuel debris, and treating contaminated water [1]. Over the past 11 years, researchers have proposed and developed various methods such as radioactive level analysis [2], detection of fuel debris [3], and inspection using a robot with a video camera [4]. This has successfully removed the spent fuel from unit 4 in December 2014 and from unit 3 in February 2021 [5,6]. However, the retrieval of fuel debris and the treatment of contaminated water are still undergoing the process [7]. Some elemental analysis systems were proposed to enhance the efficiency of the process [8]. The elemental analysis system is required to have high radiation durability. The existing elemental analysis devices are mainly huge and have low radiation durability [8]. Therefore, the development of remote elemental analysis using the well-known ultrasound phenomena, Sonoluminescence (SL), is proposed to tackle the harsh conditions inside the reactor. The idea is to attach the ultrasound sensor to the robot with fiber optic light

transmission, enabling remote analysis. A combination of robot and ultrasound technology to perform the remote analysis has been developed [9, 10]

While the study of SL has been widely known among ultrasound researchers, the development of remote elemental analysis serves particular challenges. The basic principle of SL elemental analysis is that a high-intensity ultrasonic wave produces cavitation bubbles that repeatedly grow and are adiabatically compressed in an argon saturated solution, hence the plasma production. The plasma production then induced ionization reactions and photon emission. The photon emitted a particular wavelength for each element. [11, 12]

The development of the purpose as mentioned above requires careful measures in every step. Knowing the nature of SL requires saturated argon gas [13] and an appropriate optical system to analyze the light emission remotely. The conceptual design of remote elemental analysis using SL is shown in Figure 1.

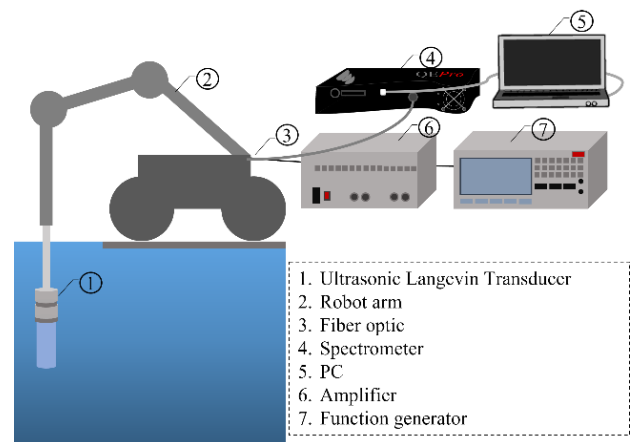


Fig. 1. Conceptual design of remote elemental analysis using SL

The designing process of the remote elemental analysis was divided into four stages: The fundamental study using a commercial bath-type ultrasound transducer, in which the

specific light emission for each element was confirmed and the operating parameter was obtained [14]; the initial steps of remote analysis performed using Langevin transducer, which allows the analysis using a smaller probe, hence, enable the remote sonication [15]; the operation of SL under a flowing solution, to confirm the feasibility on the actual application; finally, the light transmission analysis using the optical system and fiber optics. While the first three steps have been successfully carried out, the last step is still underway.

In this present work, the preliminary results of the development of a remote light analysis system using fiber optic and the spectrometer, including a confirmation using an additional method, are presented.

II. EXPERIMENTAL METHOD

A. Material

Figure 2 shows the schematic of the experimental apparatus used in this work. The experimental apparatus consists of 2 main parts: ultrasound generation and optical analysis systems. The former consists of an ultrasound transducer unit, an ultrasound controller with three operating frequencies (26 kHz, 78 kHz, and 130 kHz), and a power range around 0-50 W (commercial product by KAIJO, QUAVA mini QR-001), sample container.

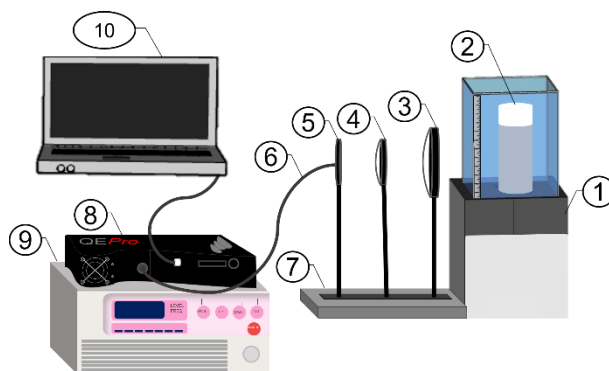
As of the latter, the SL spectrum analysis system was performed based on a standard OES spectrum analysis principle to analyze the chemical components of the sample. A focusing optic is a convex lens with a diameter of $\varnothing 75$ mm (Thorlabs, Inc.) and $\varnothing 50$ mm (Thorlabs, Inc.), a multimode optical fiber bundle (bundle of seven core fibers with a total light-receiving area of 14 mm^2), and a Czerny-Turner spectrometer (Ocean Insight – QEPRO HC-1 with wavelength range 200-950 nm, grating 600-1200 lines/mm, slit $10 \mu\text{m}$, FWHM $1.6 \mu\text{m}$) were used.

The samples used in the experiment are strontium chloride hexahydrate $\text{SrCl}_2 \cdot 6\text{H}_2\text{O}$ (FUJIFILM Wako Pure Chemical Corp, guaranteed reagent grade, assay 99.0 %), calcium chloride dihydrate $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (FUJIFILM Wako Pure Chemical Corp., guaranteed reagent grade, assay 99.0%), and sodium chloride NaCl (FUJIFILM Wako Pure Chemical Corp., guaranteed reagent grade, assay 99.5%), The purified water (ion-exchanged water) was used for the preparation of each sample. Solutions were filled into a 100 ml cylinder container made of PET material. The glycerol (KENEI Pharmaceutical Co., Ltd., type P) was used as an acoustic coupling

B. Data Acquisition Method

Our previous work has thoroughly explained a complete explanation of this method [14, 15]. A brief description of procedures is as follows. The aqueous solution with a certain concentration was prepared 1 hour before the experiment in the 100 ml solution container. The argon injection was done using Takeda Rika Kogyo Co., Ltd., argon 99,99% to attain argon saturated condition. In this process, argon acts as a cavitation agent. The gas flow rate was measured with an airflow meter (Tokyo Kaisei Co., LTD). The saturating solution was placed inside the sono-reactor container for ultrasound sonication.

In the spectrum analysis method, the analysis was performed based on a standard OES spectrum analysis principle to analyze the chemical components of the sample.



- | | |
|--|--------------------------|
| 1. Ultrasound Transducer Unit | 6. Fiber optic |
| 2. Container | 7. Lens stage |
| 3. Convex lens $\varnothing 75\text{mm}$ | 8. QEPro spectrometer |
| 4. Convex lens $\varnothing 50\text{mm}$ | 9. Ultrasound controller |
| 5. Fiber optic input + holder | 10. PC |

Fig. 2. Schematic of the experimental setup

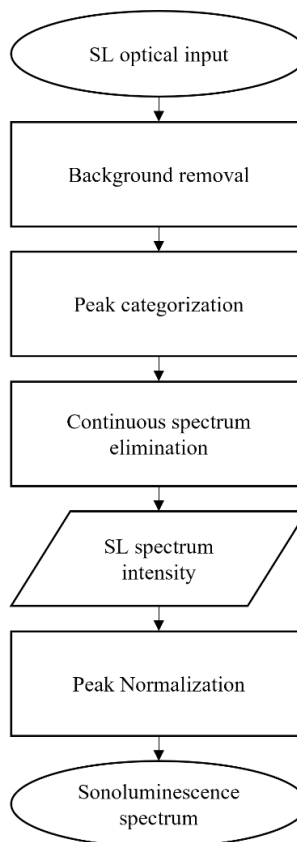


Fig. 3. flowchart of SL spectrum analysis method

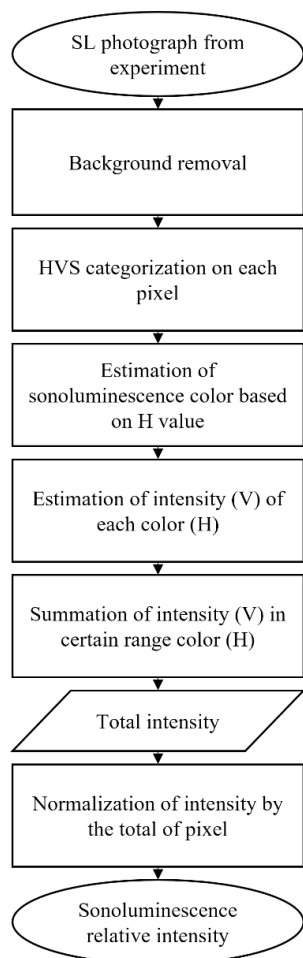


Fig. 4. Flowchart of HSV image analysis method

Once the sono reactor was operated and the SL emitted, the SL emission was then collected through a focusing optic; a convex lens with a diameter of 75 and 50 mm was used to focus the light emission to the fiber optic surface. The fiber optic then delivers the intensified light to the spectrometer. The spectrometer's integration time was set to 3s with 100 average per acquisition. The acquisition was started by measuring the background light without the SL operation, and then the value was used to correct the obtained spectrum. Finally, the background-corrected spectrum was obtained. Further data analysis methods will be explained in the following sub-section.

Moreover, to simultaneously confirm the existence of SL, a still camera (Nikon, D850) with an exposure time of 30 sec was used. A photograph of SL was taken from the side of the reactor.

C. Data Analysis Method

Although the background spectrum was used to eliminate the surrounding light and capture the actual signal during the experiment, unintended spectrum broadening was obtained from the water hydrolysis and argon reaction. Therefore, a data analysis method was developed to isolate the intended peaks. Finally, the data was presented in normalized value for each

acquisition to its own highest peak. The flowchart of the SL spectrum analysis method is shown in Figure 3

An HSV-based image analysis method was developed on MATLAB to analyze the photograph from the experiment. The analysis was done pixel by pixel on the photograph. On every pixel, the analysis was based on three characteristics, hue (H), saturation (S), and value (V). Essentially, hue shows the color distribution, saturation measures the degree of greyness, and value shows the intensity of the color. The analysis of each pixel was focused on the H and V, which correspond to color and intensity, respectively. The total of V for each pixel in *i* or *j* direction represents the intensity as shown in equation (1). The normalization of the results gives the value of relative intensity. The flowchart of image processing using the HSV method is shown in Figure 4.

$$I = \sum_i \sum_j V_{i,j} \quad (1)$$

The analysis of intensity and color from the raw data obtained from the still camera provides a resolution of around 3621x1846 pixels. In addition, using HSV-based color identification, 3% uncertainty of SL intensity can be found from this data acquisition method.

III. RESULT AND DISCUSSION

A. Remote spectrum analysis

The feasibility test on remote spectrum analysis using SL was conducted. The SL generated using the ultrasound equipment was focused using two convex lenses with 75- and 50-mm diameters. The double convex lenses were used to decrease the measurement focal length while maximizing the light intensity. The focused light is then directed to the fiber optic surface to be analyzed using the spectrometer. The existence of double convex lenses increases the visibility of SL by roughly 60%.

The spectra of SL from argon gas-saturated strontium, sodium, and calcium were studied in this paper. The spectra were investigated around 200-1000 nm. The spectrum was corrected using a background and continuum spectrum removal. However, the argon influence on the solution was assumed to be negligible.

Figure 5 shows the strontium, sodium, and calcium spectrum taken under 3s integration time with 100 averaging per acquisition. The integration time was selected after studying SL peaks' behavior on every iteration.

Figure 5 (a) shows the 407 and 421 nm strontium peaks. At the same time, Figures 5 (b) and (c) show the peaks of sodium (589 nm) and calcium (393 and 396 nm), respectively. These results aligned with the line spectrum database of NIST [16]. Some broadening was still observed in the case of calcium, even after the broadening removal. This is presumed to be caused by the resolution of the spectrometer. This is expected since the gratings chosen in this experiment were relatively lower in resolution. Also, some background noises were observed due to the long integration time of the experiment. However, it is

insignificant to the desired wavelength. These results demonstrate the feasibility of developing a remote elemental analysis system using SL.

The analysis of multi-element in the aqueous solution is shown in Figure 6. The figure shows a peak of Na and Ca observed under ultrasonic irradiation normalized to the Na peaks in a single element condition. It can be seen that the presence of mixed solution damped the intensity of the detected peaks compared to single element analysis. This is expected due to the possible quenching phenomenon [17]. Also, a spectrum broadening was observed in the mix condition. This is presumed due to the water hydrolysis that occurred. However, a broadening observed in the single element of Ca did not exist in the multi-element case. The real reason for this behavior is not yet observed.

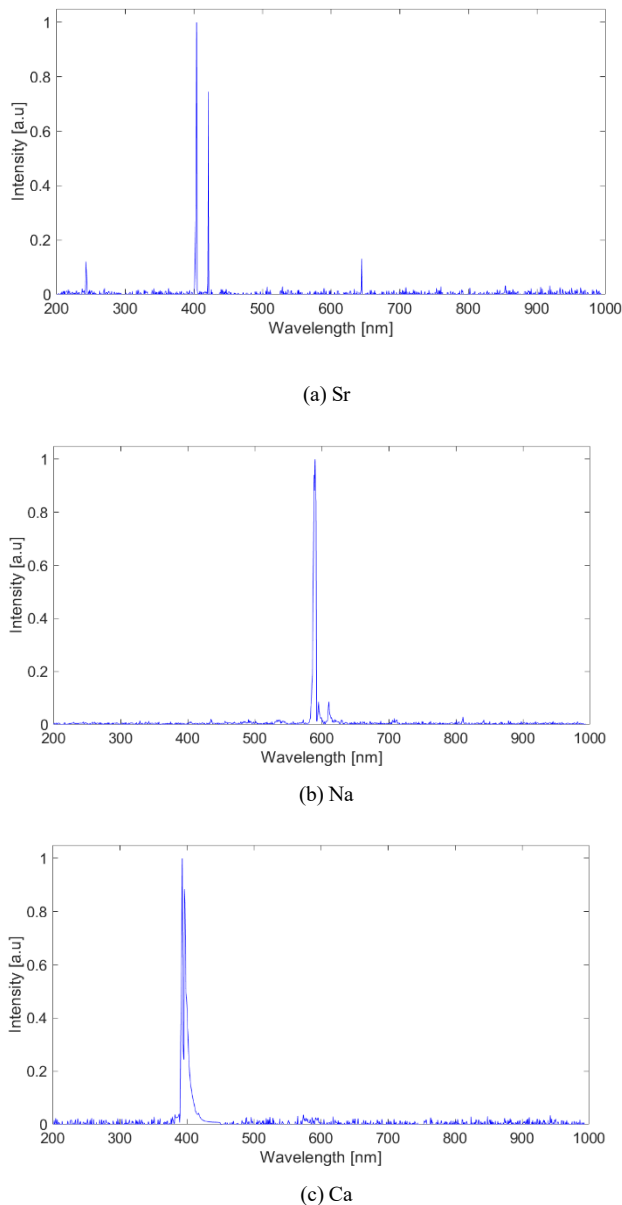


Fig. 5. SL spectrum of various elements

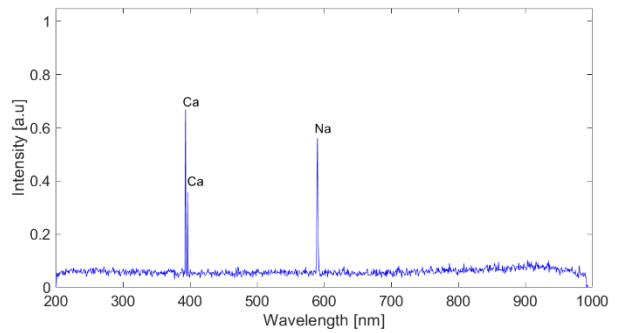


Fig. 6. SL spectrum of mixed elements

Furthermore, the mixed situation and the presence of other solid particles are expected in the actual situation. With this, a test on the presence of solid particles was conducted, and the result indicates the SL intensity was not highly affected by the presence of solid particles. Also, a filtering treatment to remove solid particles could be used as an alternative.

As for the limit of detection, a calibration curve to calculate the limit of detection has been conducted. The result demonstrates that SL intensity was proportional to the concentration of the analyzed element. The lowest element concentration that could be detected in a sufficient number intensity is in order of hundreds of ppm using the current setup and parameters. A possible increase in the detection limit could be obtained using a continuous injection of ionization agents into the solution over the measurement time. The limit of detection for each element varies and depends on several factors such as the temperature of the solution, metal boiling point, ionization energy, and excitation energy of the corresponding elements. Previous studies showed that higher ultrasound power was required to obtain the maximum possible SL intensity for elements with higher ionization energies [18].

B. SL visualization of various elements

A simultaneous observation using a still camera was conducted to confirm the visualization of SL emission. A 30s exposure time photograph of various elements taken in a dark environment is presented in Figure 7. The results show different colors of each component and demonstrate a pattern showing the high and low-intensity area resulting from the ultrasound interaction in the liquid.

Furthermore, the HSV analysis performed on the obtained photograph determines the photograph's color distribution. The use of HSV for image analysis, especially in image segmentation and histogram, has been widely used [19]. The color distribution is on a 360 degrees rotation basis. The advantage of HSV compared to other color schemes is that it has only one dimension of color (H). Roughly, the color is categorized as follows: 0-90 shades of red to yellow, 90-180 for shades of green, 180-270 for blue, 270-360 pink to red.

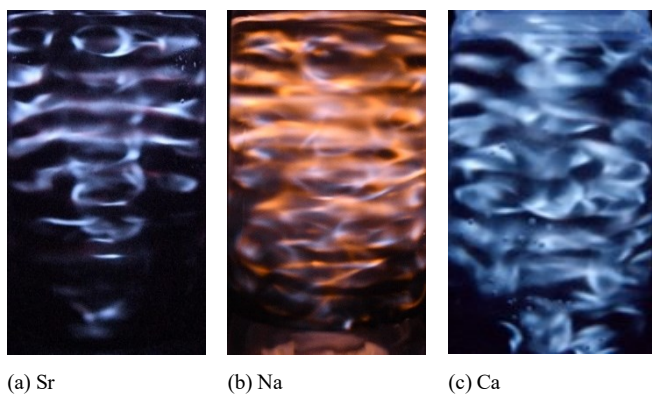


Fig. 7. SL of various element

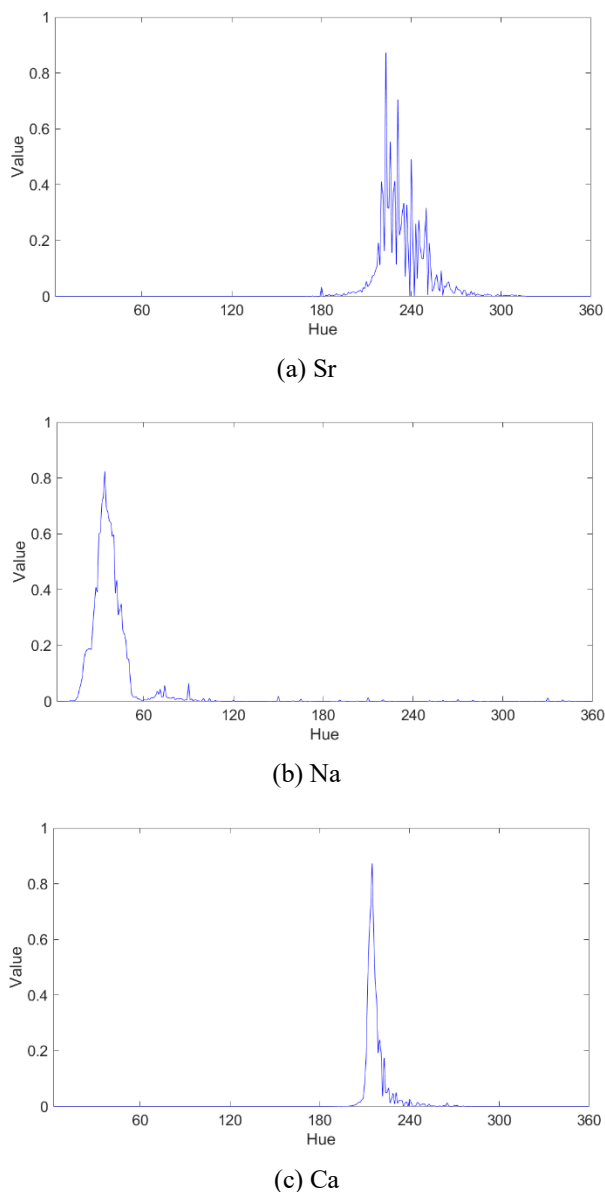


Fig. 8. HSV analysis results of various elements

The algorithm has been tested to various test images before the analysis. The results indicate that the obtained intensity tends to be narrow with the increase in pixel number. This is expected due to the convergence of the data with the increase in pixel number.

The HSV analysis results on strontium, sodium, and calcium are demonstrated in Figure 8. The results match the visual observation of the photograph.

In these results, a broadening peak was observed. Mainly, the one from strontium was the most noticeable, which is expected due to the broad distribution of the color. Nevertheless, these results support the reliability of the results obtained using spectrometers. Hence the feasibility of developing a remote elemental analysis system using SL.

IV. CONCLUSION

The feasibility test on remote spectrum analysis using SL was conducted. Strontium, sodium, and calcium in an aqueous solution were used as a sample in the experiment. The optical arrangement enables the remote spectrum analysis using fiber optic and spectral analysis. A simultaneous observation using a still camera was conducted to confirm the visualization of SL emission. The results obtained in this experiment demonstrate the ability to perform remote elemental analysis using SL.

REFERENCES

- [1] K. Yamashita, "Fukushima Daiichi. Post-accident countermeasures and mid-and-long-term action plan." *Nippon Genshiryoku Gakkai-Shi* 54, no. 6, pp. 381-385, 2012.
- [2] Y. Satou, K. Sueki, K. Sasa, H. Yoshikawa, S. Nakama, H. Minowa, Y. Abe, I. Nakai, T. Ono, K. Adachi, and Y. Igarashi, "Analysis of two forms of radioactive particles emitted during the early stages of the Fukushima Dai-ichi Nuclear Power Station accident." *Geochemical Journal*, 52(2), pp.137-143, 2018.
- [3] M. Nancekievill, J. Espinosa, S. Watson, B. Lennox, A. Jones, M.J. Joyce, J.I. Katakura, K. Okumura, S. Kamada, M. Katoh, and K. Nishimura, "Detection of simulated Fukushima Daiichi fuel debris using a remotely operated vehicle at the naraha test facility." *Sensors*, 19(20), pp. 4602, 2019.
- [4] Y. Sato, Y. Terasaka, W. Utsugi, H. Kikuchi, H. Kiyooka, and T. Torii, "Radiation imaging using a compact Compton camera mounted on a crawler robot inside reactor buildings of Fukushima Daiichi Nuclear Power Station." *Journal of Nuclear Science and Technology*, 56(9-10), pp.801-808, 2019.
- [5] Tokyo Electric Power Company, 2014, <https://www.tepco.co.jp/en/hd/decommission/progress/removal/unit4/index-e.html> [retrieved, May 3, 2022]
- [6] Tokyo Electric Power Company, 2021 <https://www.tepco.co.jp/en/hd/decommission/progress/removal/unit3/index-e.html> [retrieved, May 3, 2022]
- [7] Kinoshita, H., Tayama, R., Kometani, E.Y., Asano, T. and Kani, Y., 2014. Development of new technology for Fukushima Daiichi nuclear power station reconstruction. *Hitachi Review*, 63(4), pp.183-190.
- [8] A. Ruas, A. Matsumoto, H. Ohba, K. Akaoka, and I. Wakaida, "Application of laser-induced breakdown spectroscopy to zirconium in aqueous solution." *Spectrochimica Acta Part B: Atomic Spectroscopy*, 131, pp.99-106, 2017.
- [9] G. Endo, H. Takahashi, and H. Kikura, "Challenge to Investigation of Fuel Debris in RPV by an Advanced Super Dragon Articulated Robot Arm: Design and Prototyping of a Lightweight Super Long Reach Articulated Manipulator." In *International Conference on Nuclear*

- Engineering, American Society of Mechanical Engineers, Vol. 83761, pp. V001T04A013, August 2020.
- [10] Takahashi, N. Shoji, A. Ito, and H. Kikura, "Fundamental Study on Ultrasound Sensing Technology using Parametric Sound." WIT Transactions on Engineering Sciences, Vol. 128, pp.103-111, 2020.
- [11] S. Hilgenfeldt, S. Grossmann, and D. Lohse, "A simple explanation of light emission in sonoluminescence." *Nature*, 398(6726), pp.402-405, 1999.
- [12] O.I. Yurchenko, O.S. Kalinenko, A.N. Baklanov, E.A. Belov, and L.V. Baklanova, "Sonoluminescence spectroscopy as a promising new analytical method." *Journal of Applied Spectroscopy*, Vol. 83(1), pp.105-110, 2016.
- [13] C. Sehgal, R.P. Steer, R.G. Sutherland, and RE Verrall, "Sonoluminescence of argon saturated alkali metal salt solutions as a probe of acoustic cavitation." *The Journal of Chemical Physics*, Vol. 70(5), pp.2242-2248, 1979.
- [14] S.S.N. Sailellah, H. Takahashi, and H. Kikura, "Fundamental study for elemental analysis using sonoluminescence in aqueous solution." In *The Proceedings of the International Conference on Nuclear Engineering (ICONE) 2019*. The Japan Society of Mechanical Engineers. Vol. 27, pp. 2187, 2019.
- [15] S.S.N. Sailellah, H. Takahashi, and H. Kikura, "Sonoluminescence Elemental Analysis using External Transducer in Aqueous Solution." *Advanced Experimental Mechanics*, Vol. 5, pp.80-85, 2020.
- [16] P.J. Linstrom and W.G. Mallard, Eds., "NIST Chemistry WebBook, NIST Standard Reference Database Number 69", National Institute of Standards and Technology, Gaithersburg MD, 20899, <https://doi.org/10.18434/T4D303>, (retrieved May 3, 2022).
- [17] P.M. Kanthale, A. Brotchie, F. Grieser, and M. Ashokkumar, "Sonoluminescence quenching and cavitation bubble temperature measurements in an ionic liquid." *Ultrasonics sonochemistry*, 20(1), pp.47-51, 2013.
- [18] O.I. Yurchenko, O.S. Kalinenko, A.N. Baklanov, E.A. Belov, and L.V. Baklanova, "Sonoluminescence spectroscopy as a promising new analytical method." *Journal of Applied Spectroscopy*, 83(1), pp.105-110, 2016.
- [19] P. Ganesan, V. Rajini, B.S. Sathish, and K.B. Shaik, "HSV color space based segmentation of the region of interest in satellite images." In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 101-105. 2014

Detailed Bond Graph Modeling of PV-Battery System

S. Arash Omid
 Department of Engineering and
 applied science
 Memorial University of Newfoundland
 St John's, Canada
 saomidi@mun.ca

Geoff Rideout
 Department of Engineering and
 Applied science
 Memorial University of Newfoundland
 St John's, Canada
 g.rideout@mun.ca

M. Tariq Iqbal
 Department of Engineering and
 Applied science
 Memorial University of Newfoundland
 St John's, Canada
 tariq@mun.ca

Abstract— A bond graph modelling of a Photovoltaic system is presented in this study. A Photovoltaic system's four main components are Photovoltaic generator, DC-DC converter, battery, and DC-AC inverter. This study shows the bond graph models of the aforementioned part. A five-parameters PV generator is chosen based on high accuracy and illustrating of the effects of temperature and solar irradiation. To develop a link with reality, the bond graph model is created using the specs of the CS1U-400 Canadian solar module. The effects of different duty cycles have been examined using a bond graph model of a synchronous Boost converter and a synchronous Cuk converter. For a Lead Acid Battery, a novel bond graph model was developed. This model investigates the impacts of temperature and current in charging and discharging scenarios. Finally, a 3-phase inverter bond graph is designed to connect the PV system to the grid. The 20-Sim software is used to simulate the developed models, and the results are discussed.

Keywords—Dynamic modeling, Bond Graph, Photovoltaic System, Renewable Energy System

I. INTRODUCTION

The photovoltaic (PV) power industry is quickly expanding, notably in the field of distributed generation, due to growing interest in renewable energy supplies. As a result, designers require a versatile and trustworthy tool to precisely anticipate the electrical power generated by PV arrays of varied sizes [1]. The difficulty of predicting energy output from a PV cell system is particularly difficult since the electrical current yield is highly dependent on weather and environmental conditions, particularly temperature and global irradiance [2], hence the use of batteries for photovoltaic solar power. The main purpose of using batteries is storing extra energy in sunny days and supply load during night and cloudy days, also stabilizing and regulating voltage and frequency to increase the reliability of system. Moreover, some electrical circuits are needed to achieve this purpose of producing reliable electricity using solar energy and supplying load and connecting to grid. Following parts are essentials of a Photovoltaic System (PV system), Figure 1:

- Solar Panel
- Energy Storage System (ESS)
- DC-DC Converter
- DC-AC Inverter

A photovoltaic system can be modelled in a variety of ways. Semiconductor physics may be used to model a photovoltaic cell at a high level, encompassing phenomena like charge

carrier density, mobility, and recombination in the charge space [3].

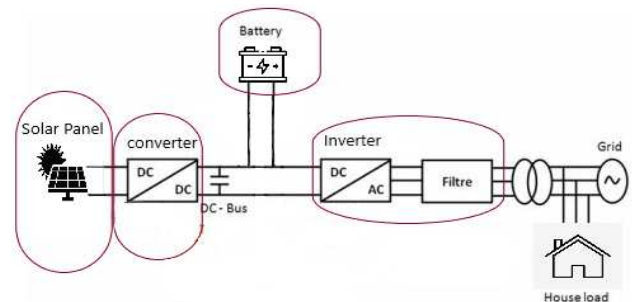


Figure 1. Four essential parts of a PV system

Analytical methodologies are difficult to use for this type of modelling, hence numeric processing is required to run simulations. It means that a lot of computational power is required. Researchers working on novel solar cells, on the other hand, will find such modelling valuable and required. However, such extensive modelling is not necessary for engineers interested in solar cell applications. Among the analytical approaches for computing parameters, the one described in [4] allows for the computation of five parameters using data from the datasheet of the module under investigation. This method is appealing since it does not need any prior testing or measurements. This work used the four-parameter equivalent of one diode with a series resistor [5]. The parameters are calculated analytically, using a technique described in [6]. Other approaches need iterative procedures, thus explicit expressions are employed instead. Because of the five controlling variables, this model[7] is commonly referred to as a "five-parameter model." Its comparable circuit consists of a photovoltaic current source I_{ph} in series with a light-sensitive diode (D) and a shunt resistance (R_p) expressing a leakage current, as well as a series resistance (R_s) indicating internal resistance to current flow. This circuit may be used for a single cell, a module made up of numerous cells, or an array made up of multiple modules [8]. The PV cell's electrical equivalent circuit may also be used to calculate the mathematical equation for PV modules/generators. Internal parameters, voltages, and currents will be a function of the number of cells, modules, and their combination in series and/or parallel, depending on the application [9].

Variable structure systems include switched mode power converters (SMPCs). The structure of the system's dynamical components is determined by the power switch's instantaneous positions. In the literature on mathematical

modelling of SMPCs, several analytical and graphical techniques have been developed. Large signal discrete time models for computer simulation and continuous time models for physical understanding have been presented as analytical approaches [10]. The switching dc-dc converters aid in increasing the voltage from a low voltage source, smoothing the path to a synchronized DC output voltage that would otherwise need a large number of battery voltage sources [11]. Generally, dc-dc converters are used to boost the output of renewable energy systems. High step-up output voltage gain, low input current and low current ripples, high output voltage and low voltage ripples, and improved efficiency are all required for dc-dc converters [12]. The most common DC-DC converters are Buck converter, Boost converter, Buck-Boost converter, Cuk converter, Zeta converter, and Sepic converter, which the most used ones in PV systems are Boost converter, Buck-Boost converter, and Cuk converter.

For all types of power electronics systems, including solar systems, power supply dependability and power quality have become critical challenges. When connecting a solar system to the utility, the PV system must fulfil the harmonic standard as well as the active power supply requirements. Because PV generates DC voltage, an inverter is required to convert the DC electricity to AC before connecting it to the grid. Grid is an infinite-capacity voltage source. The inverter's output voltage and frequency should match the grid's frequency and voltage.

II. INTRODUCTION TO BOND GRAPH

The bond graph is a framework for dynamic modelling. Paynter [13] presented it first, then Karnopp and Rosenberg expanded on it subsequently. This formalism allows the power transfers between the many subsystems of a physical system to be visually depicted in a coherent manner [14]. A bond graph is a multidisciplinary tool for modelling complicated systems that employs a unified methodology [15]. A bond graph model is a graph $G(S, A)$ in which the nodes S represent physical components, subsystems, and other fundamental elements like junctions, and the arcs A reflect power exchanges between the nodes S . A half-arrow carrying two conjugate variables named effort (e) and flow (f) represents this power exchange. The instantaneous power exchanged between two nodes S is represented by the product of these variables. A half arrow represents the transfer of power. Each physical domain studied has its own effort and flow variables [16]. The bond graph formalism is based on the power conservation and energy continuity principles. Energy cannot be generated or destroyed within a system; it can only be stored, dispersed, altered, or exchanged with the outside. As a result, the bond graph formalism only has a few parts. The element R allows for the representation of energy dissipation, whether it is reversible or irreversible. The factors C and I make it feasible to represent different energy storage forms. The transformation components TF and GY allow for the representation of energy transformation and/or movement from one physical domain to another. It is feasible to simulate the conditions of the system with which the latter would carry out the energy exchanges using the source elements of effort Se and flow Sf , which may be assimilated to limitless sources [17]. The junction elements, which are utilized to link components with the same effort (junction

"0") or the same flow (junction "1"), are added to these diverse elements. The idea of causality, which controls the interactions between components at the level of efforts and flows, is one of the most significant structural aspects of bond graph formalism. Indeed, the cause-effect linkages may be derived directly from the system's graphical representation [18].

In Electrical Engineering, the bond graph approach is also beneficial. Designers will be able to create a variety of dynamic systems with appropriate features using this method. All forms of electrical components and parasitic phenomena can be described by specifying linear and/or non-linear R , I , C , TF , and GY components. As a result, creating accurate and realistic dynamic models of various circuits and topologies will no longer be an issue.

III. LITRATURE REVIEW

Many electrical models have been presented to simulate PV cells working under various situations [19,21]. The number of parameters to be identified determines the model's complexity. Each model is simply an upgrade on the ideal model, which includes a current source to represent the incident solar power and a diode to represent the P-N junction. Additional features can be included to improve the description of the PV cell's performance in different operating modes. The most common model is the one-diode model. Because of its simplicity and accuracy in power generating mode, it is employed for PV cells and PV modules. The Bishop model, which represents the performance of a PV cell under reverse polarization, is a result of the development of the one-diode model. The two-diode model improves on the one-diode model by accounting for resistive losses and recombination processes in the circuit's different electrical components. Since this model is unable to express the PV voltage as an explicit function of the current, it must be solved iteratively or by using the Lambert-W function. As a result, high precision comes at the expense of a complicated formalization of the issues and a high computational cost. In [21], a dynamic battery model is provided that accounts for temperature variations and may be implemented into electric vehicle simulation for computer simulation testing purposes; However, there is no consideration of battery behavior in charging and discharging situations. In [22], the study presents a bond graph technique to model switched mode power converters such as, H-bridge inverter, Boost converter, and Cuk converter. The notion of switched power junctions is used to simulate the switching phenomena. The bond graph modelling of switching power converters is discussed in order to achieve the big signal averaged, steady-state, and tiny signal AC models. [23] proposes a bond graph model of a photovoltaic construction consisting of a solar generator, a Maximum Power Point Tracking converter, and a three-phase inverter powering an asynchronous motor-pump. The system responses were significantly enhanced after a vectorial control study of the closed loop system.

In this paper, a bond graph model of a photovoltaic-battery system is presented, which includes a five-parameter PV cell, a Boost converter, a Cuk converter, a lead acid battery, and a

three-phase inverter. The solar cell and power switches were constructed using datasheet parameters.

This paper structure includes six sections. Section I gives an overview of the concept of PV system and four essential parts. In section II, a brief introduction of bond graph modeling is proposed. A review of some recent related research is presented in section III. Section IV illustrates the bond graph models of mentioned parts of a PV system and discusses the equations and dominant parameters of each part. In section V, the simulation results are discussed. And in the last section a brief conclusion is presented.

IV. SYSTEM MODELING

A. Bond graph model of the PV generator

A model with five parameters is picked up for this work. This model includes a current source that provides a current I_{ph} proportional to the illumination, a diode for the cell's polarization phenomenon through which the current I_d flows, and a series resistor R_s and a parallel R_p for the losses, Fig. 2.

The five parameters of this model are extracted using numerical iterative approaches such as the Levenberg-Marquardt algorithm. Analytical approaches [24,25] can also be utilized to solve the implicit nonlinear equation that reflects the five-parameter model.

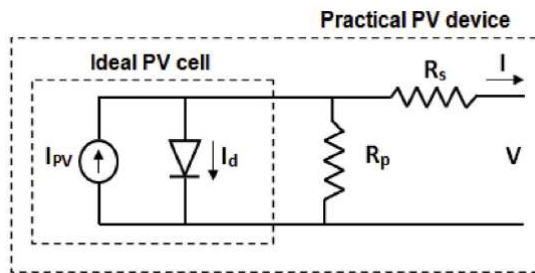


Figure 2. 5-parameter Practical PV cell model

Equation (1) and (2) illustrate the internal PV cell produced current and diode current, respectively:

$$I_{ph} = n_p(I_{sc} + K_i(T - T_r))(S_i/100) \tag{1}$$

$$K_i = I_{sc}I_{sc} \tag{2}$$

$$I_d = n_p I_{sat} \left(\exp\left(\frac{qV}{AKTn_s}\right) \right) \tag{3}$$

which I_d , R_s , R_p , I_{ph} , n_p , n_s , I_{sc} , I_{sc} , I_{sat} , T , T_r , q , A , K , and S_i are output current, diode current, series resistor, parallel resistor, Photo Current, number of cells connected in parallel, number of cells connected in series, short circuit current coefficient, short circuit current, saturation current, temperature, temperature reference, electron charge, ideality factor, Boltzmann constant, and Solar illumination (Irradiation), respectively.

Following equation shows the output current of the PV cell, based on electrical circuit:

$$I = I_{ph} - I_d - \left(\frac{V + I R_s}{R_p} \right) \tag{4}$$

The equation (4) represents the modified diode shield voltage as the temperature changes, where k_v is the temperature coefficient of open circuit voltage:

$$V_t = K_v(T - T_r) \tag{5}$$

$$K_v = C_t V_{oc} \tag{6}$$

which V_t , K_v , T , T_r , C_t , and V_{oc} are, respectively, modified diode shield voltage, temperature coefficient of open circuit voltage, temperature, temperature reference, temperature coefficient, and open circuit voltage.

The I-V characteristic of this module is represented by the following equation:

$$I(1 + R_s/R_p) = -n_p I_{sat} \left\{ \exp\left(\frac{q}{AKTn_s}\right) \left(\frac{V}{n_s} + IR_s\right) - 1 \right\} + n_p I_{ph} - \frac{V - n_s}{R_p} \tag{7}$$

The CS1U-400 Canadian solar module is utilized to simulate in this investigation. Table 1 lists the CS1U-400 specs that were used in this investigation.

The bond graph model of the PV cell has been illustrated in Figure 3.

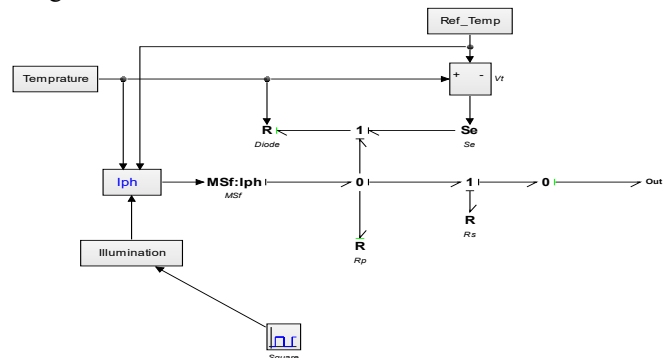


Figure 3. Bond graph model of PV generator

The aforementioned model, Figure 3, contains the implementation of three equations, with the one that calculates the photocurrent based on two factors, sun irradiation and temperature, being the most important, equation (7). The external I_{ph} block includes these parameters, allowing users to create C code and link it to SYMBOLS. To generate an I_{ph} current, equation (1) is expressed as a "Iph.dll block" in C code.

The photovoltaic generator is connected to DC bus through a step-up DC-DC converter.

B. Bond graph model of the DC-DC converters

To increase the DC voltage to standard level, stabilizing the transferred power, and providing the Maximum Power Point Tracking option, using a step-up converter is necessary. In this work, two most popular switch mode converters,

Table 1. CS1U-400 Canadian solar module main characteristics

ELECTRICAL DATA | NMOT*

CS1U	400MS	405MS	410MS	415MS	420MS
Nominal Max. Power (Pmax)	296 W	300 W	304 W	307 W	311 W
Opt. Operating Voltage (Vmp)	40.8 V	41.0 V	41.2 V	41.4 V	41.5 V
Opt. Operating Current (Imp)	7.26 A	7.32 A	7.37 A	7.43 A	7.48 A
Open Circuit Voltage (Voc)	49.9 V	50.0 V	50.1 V	50.2 V	50.3 V
Short Circuit Current (Isc)	7.75 A	7.79 A	7.83 A	7.87 A	7.91 A

* Under Nominal Module Operating Temperature (NMOT), irradiance of 800 W/m² spectrum AM 1.5, ambient temperature 20°C, wind speed 1 m/s.

TEMPERATURE CHARACTERISTICS

Specification	Data
Temperature Coefficient (Pmax)	-0.37 % / °C
Temperature Coefficient (Voc)	-0.29 % / °C
Temperature Coefficient (Isc)	0.05 % / °C
Nominal Module Operating Temperature	43±3 °C

figure 4, which are synchronous Boost converter and synchronous Cuk converter, have been modeled in 20-Sim software, Figure 6. In designing these converters, IRF150 characteristics as the power switch has been used, Figure 5.

Following equations illustrate design characteristics of Boost Converter[26]:

$$\frac{V_{out}}{V_{in}} = \frac{1}{1-D} \tag{8}$$

$$\frac{I_{out}}{I_{in}} = 1 - D \tag{9}$$

$$L = \frac{V_{out}D}{\Delta i_L f} \tag{10}$$

$$C_o = \frac{D}{R(\Delta V_{out}/V_{in})f} \tag{11}$$

frequency. Also design characteristic of Cuk converter have been shown below[26]:

$$\frac{V_{out}}{V_{in}} = -\frac{D}{1-D} \tag{12}$$

$$\frac{I_{out}}{I_{in}} = -\frac{1-D}{D} \tag{13}$$

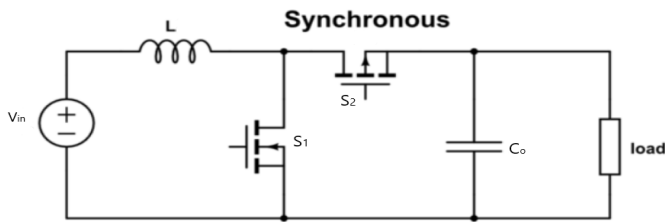
$$L_1 = \frac{V_{out}D}{\Delta i_{L1}f} \tag{14}$$

$$L_2 = \frac{V_{out}D}{\Delta i_{L2}f} \tag{15}$$

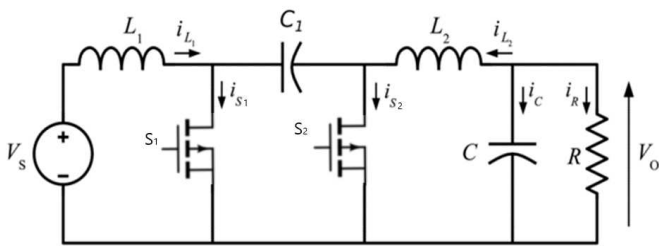
$$C_o = \frac{D}{R(\Delta V_{out}/V_{in})f} \tag{16}$$

$$C_1 = \frac{D}{R(\Delta V_{C1}/V_{out})f} \tag{17}$$

which $L_1, L_2, \Delta i_{L1}, \Delta i_{L2}, C_o, C_1,$ and ΔV_{C1} are, respectively, input inductor, output inductor, input inductor current ripple, output inductor current ripple, output capacitor, series capacitor, and series capacitor voltage ripple.



(a) Synchronous Boost converter



(b) Synchronous Cuk converter

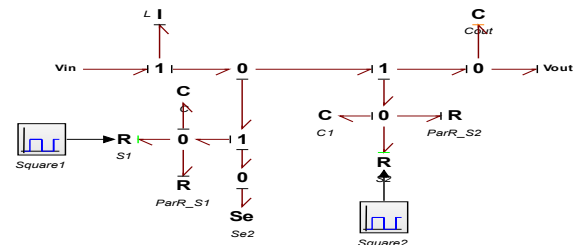
Figure 4. Converters. (a)Boost converter (b)Cuk converter

Product Summary

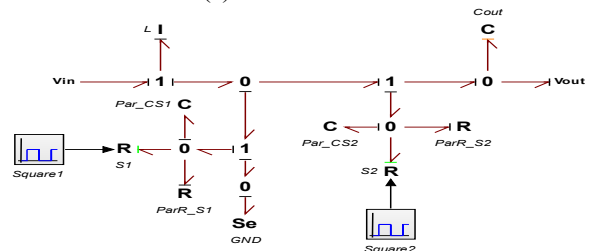
Part Number	BVDSS	RDS(on)	ID
IRF150	100V	0.055Ω	38A

Figure 5, IRF150 parameters summary

which $V_{out}, V_{in}, D, I_{out}, I_{in}, L, \Delta i_L, C_o, R, \Delta V_{out},$ and f are, respectively, output voltage, input voltage, duty cycle, output current, input current, Inductor, inductor current ripple, output capacitor, load resistor, output voltage ripple, and switching



(a) Boost Converter



(b) Cuk converter

Figure 6. Bond graph model. (a) Boost converter (b)Cuk converter

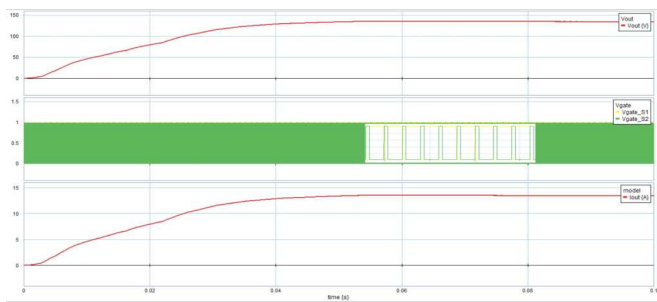
C. Bond graph model of the Lead Acid battery

There are some parameters which play significant roles in lead acid battery performance, such as temperature, State of Charge (SOC), charging current, and discharging current. A model which can predict the battery behavior in different values of mentioned parameters correctly, is a desirable model. Figure 7 shows the battery equivalent model which is the best option to satisfy aforementioned criteria. In this work, Battery is “T-105 with Bayonet Cap” specifications have been used to design bond graph model.

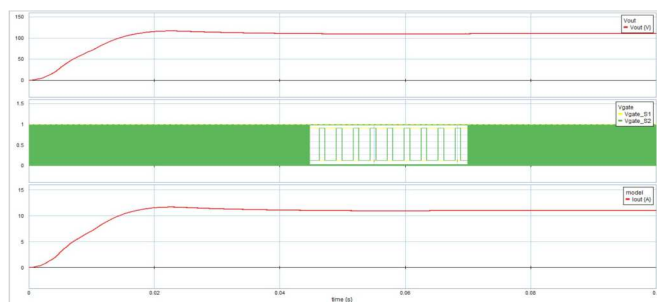
Any change in solar irradiation and temperature will have direct impact on PV generator performance.

By running the Boost converter in a situation where PV panel is on nominal situation, the results of the converter output will be shown in Figure 11. In case of changing duty cycle, the output voltage and output current will change, based on (8) & (9). Switching frequency of both converters are 50 KHz.

Charging and discharging of the battery has been shown in Figure 12. In Figure 12 and 13, the outputs of a single Lead Acid battery is shown at 25 and -10 degrees, respectively. Therefore, in lower temperature the battery capacity decreases and with same currents in lower temperature it discharges and charges faster. But, in order to add the ESS to the PV system, the output voltage must be equal to the DC bus voltage. By putting number of batteries in series this purpose will be achieved. Based on the design, the ESS can be put in lower DC voltage (PV output) or high DC voltage (converter output). In most cases the energy storage system will be at high DC voltage part. In this work also, because of next works in future and further developments, it is connected to high DC voltage. Therefore, by putting 25 single batteries in series, it will possible to attach them to DC bus.



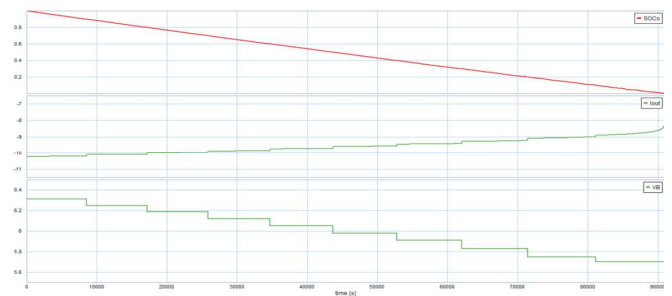
(a) D=0.75.



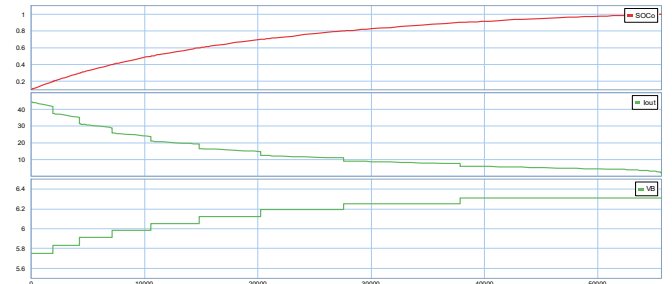
(b) D=0.6.

Figure 11. Boost converter waveforms. (a) D=0.75. (b) D=0.6.

As mentioned before, to connect the PV system to grid, an inverter is needed. In North America, the standard voltage is 120 v and 60 Hz. To achieve this purpose and based on (24), DC voltage must be 150V, also switching frequency has to be 60 Hz. The results of simulation of the 3-phase inverter with mentioned situation have been shown in Figure 14.

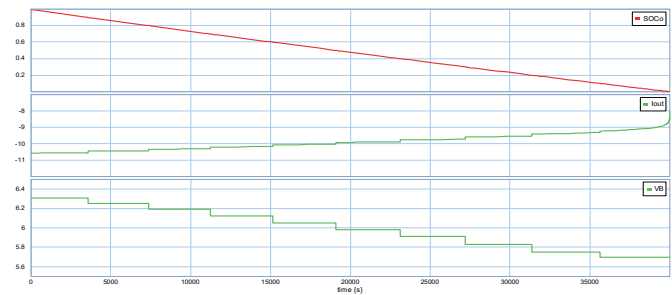


(a) Discharging by 10A

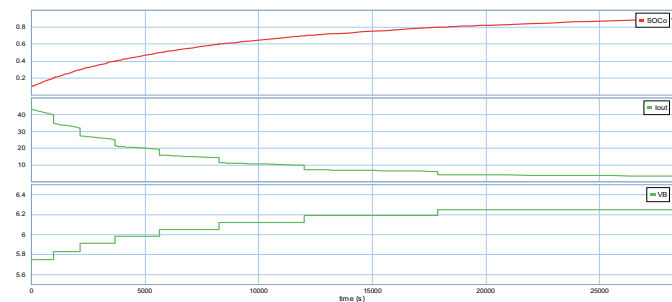


(b) Charging by 30A

Figure 12. Battery outputs at 25 degrees. (a) Discharging by 10A. (b) Charging by 30A.



(a) Discharging by 10A



(b) Charging by 30A

Figure 13. Battery outputs at -10 degrees. (a) Discharging by 10A. (b) Charging by 30A.

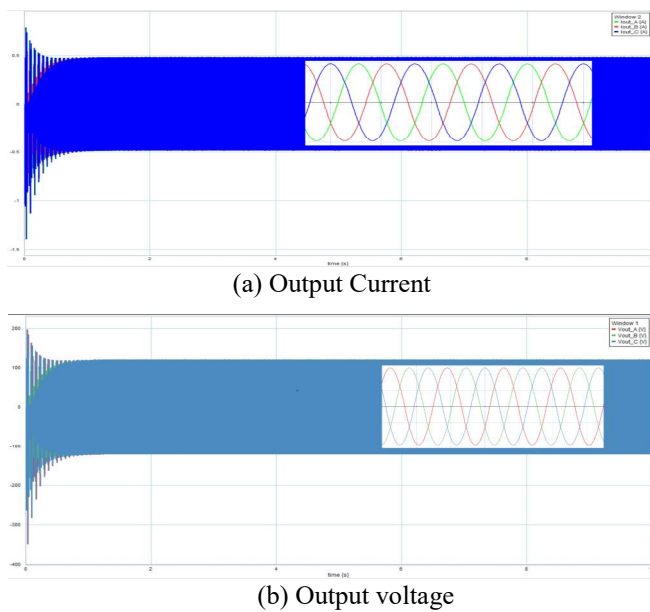


Figure 14. Inverter outputs. (a) Current. (b) Voltage.

VI. CONCLUSION

A bond graph modelling of a PV system is provided in this paper. PV generator, DC-DC converter, Battery, and DC-AC inverter are the four fundamental components of a PV system that have been modelled in this article. Because being more accurate and showing the effects of temperature and solar irradiation more precisely, a five-parameters PV generator is chosen. The specifications of the CSIU-400 Canadian solar module are utilized to create the bond graph model in order to make a connection with reality. Bond graph model of a synchronous Boost converter and a synchronous Cuk converter, as two of the most popular candidates of PV applications, are designed and the effects of different duty cycles have been studied. A new bond graph model for a Lead Acid Battery was created. The effects of output current and temperature in charging and discharging scenarios are investigated in this model. Finally, to link the PV system to the grid, a 3-phase inverter bond graph is designed.

REFERENCES

[1] Tian, Hongmei, Fernando Mancilla-David, Kevin Ellis, Eduard Muljadi, and Peter Jenkins. "A cell-to-module-to-array detailed model for photovoltaic panels." *Solar energy* 86, no. 9 (2012): 2695-2706.

[2] Ciulla, Giuseppina, Vincenzo Franzitta, Valerio Lo Brano, Alessia Viola, and Marco Trapanese. "Mini wind plant to power telecommunication systems: A case study in sicily." In *Advanced Materials Research*, vol. 622, pp. 1078-1083. Trans Tech Publications Ltd, 2013.

[3] Pietro, P. "Altermatt. Models for numerical device simulations of crystalline silicon solar cells: a review." *J. Comput. Electron* 10, no. 3 (2011): 314. K. Elissa, "Title of paper if known," unpublished.

[4] De Soto, Widaly, Sanford A. Klein, and William A. Beckman. "Improvement and validation of a model for photovoltaic array performance." *Solar energy* 80, no. 1 (2006): 78-88.

[5] Chin, Vun Jack, Zainal Salam, and Kashif Ishaque. "Cell modelling and model parameters estimation techniques for photovoltaic simulator application: A review." *Applied Energy* 154 (2015): 500-519.

[6] Townsend, Timothy U. "A method for estimating the long-term performance of direct-coupled photovoltaic systems." PhD diss., 1989.

[7] Besheer, Ahmad H., Ahmed M. Kassem, and Almoataz Y. Abdelaziz. "Single-diode model based photovoltaic module: analysis and comparison approach." *Electric Power Components and Systems* 42, no. 12 (2014): 1289-1300.

[8] De Soto, Widaly, Sanford A. Klein, and William A. Beckman. "Improvement and validation of a model for photovoltaic array performance." *Solar energy* 80, no. 1 (2006): 78-88.

[9] Carrero, C., D. Ramirez, J. Rodriguez, and C. A. Platero. "Accurate and fast convergence method for parameter estimation of PV generators based on three main points of the I-V curve." *Renewable Energy* 36, no. 11 (2011): 2972-2977.

[10] Erickson, Robert W., Slobodan Cuk, and R. D. Middlebrook. "Large-signal modelling and analysis of switching regulators." In *1982 IEEE Power Electronics Specialists conference*, pp. 240-250. IEEE, 1982.

[11] Tan, Siew-Chong, Y. M. Lai, and Chi Kong Tse. "Implementation of pulse-width-modulation based sliding mode controller for boost converters." *IEEE Power Electronics Letters* 3, no. 4 (2005): 130-135.

[12] Palanisamy, R., A. U. Mutawakkil, and K. Vijayakumar. "Hysteresis SVM for coupled inductor z source diode clamped 3-level inverter based grid connected PV system." *International Journal of Power Electronics and Drive Systems* 7, no. 4 (2016).

[13] Paynter, H. M., and P. Briggs. "Massachusetts Institute of Technology, Analysis and design of engineering systems: class notes for MIT course 2.751." (1961).

[14] Badoud, Abd Essalam, Bertrand Raison, Luiz Lavado Fernando Vila, Belkacem Ould Bouamama, and Mabrouk Khemliche. "Modeling, simulation and hardware implementation of a bond graph-maximum power point tracker for a photovoltaic panel under partially shaded conditions." *Simulation* 92, no. 7 (2016): 687-707.

[15] Umarikar, A. C., and L. Umanand. "Modelling of switched mode power converters using bond graph." *IEE Proceedings-Electric Power Applications* 152, no. 1 (2005): 51-60.

[16] Abdallah, Ibrahim, Anne-Lise Gehin, and Belkacem Ould Bouamama. "Event driven hybrid bond graph for hybrid renewable energy systems part i: Modelling and operating mode management." *International journal of hydrogen energy* 43, no. 49 (2018): 22088-22107.

[17] Karnopp, Dean, Ronald Rosenberg, and Alan S. Perelson. "System dynamics: a unified approach." *IEEE Transactions on Systems, Man, and Cybernetics* 10 (1976): 724-724.

[18] Benabdelaziz, Kawtar, and Mohammed Maaroufi. "Battery dynamic energy model for use in electric vehicle simulation." *International Journal of Hydrogen Energy* 42, no. 30 (2017): 19496-19503.

[19] Villa-Villaseñor, Noé, and René Galindo-Orozco. "Bond graph modelling of a 4-parameter photovoltaic array." *Mathematical and Computer Modelling of Dynamical Systems* 24, no. 3 (2018): 275-295.

[20] Madi, Saida, and Aissa Kheldoun. "Bond graph based modeling for parameter identification of photovoltaic module." *Energy* 141 (2017): 1456-1465.

[21] Benabdelaziz, Kawtar, and Mohammed Maaroufi. "Battery dynamic energy model for use in electric vehicle simulation." *International Journal of Hydrogen Energy* 42, no. 30 (2017): 19496-19503.

[22] Umarikar, A. C., and L. Umanand. "Modelling of switched mode power converters using bond graph." *IEE Proceedings-Electric Power Applications* 152, no. 1 (2005): 51-60.

[23] Mezghanni, Dhafer, R. Andoulsi, Abdelkader Mami, and Geneviève Dauphin-Tanguy. "Bond graph modelling of a photovoltaic system feeding an induction motor-pump." *Simulation Modelling Practice and Theory* 15, no. 10 (2007): 1224-1238.

[24] Douiri, Moulay Rachid. "A predictive model for solar photovoltaic power based on computational intelligence technique." *Arabian Journal for Science and Engineering* 44, no. 8 (2019): 6923-6940.

[25] Jain, Amit, and Avinashi Kapoor. "Exact analytical solutions of the parameters of real solar cells using Lambert W-function." *Solar Energy Materials and Solar Cells* 81, no. 2 (2004): 269-277.

[26] Selvabharathi, P., S. Veerakumar, and V. Kamatchi Kannan. "Simulation Of Dc-Dc Converter Topology For Solar Pv System Under Varying Climatic Conditions With Mppt Controller." In *IOP Conference Series: Materials Science and Engineering*, vol. 1084, no. 1, p. 012084. IOP Publishing, 2021

Design and Simulate a 500 MW Grid-Connected PV Farm for Labrador

Sayed Arfat Alam Quadri¹, Mohamad Mahdi Baalbaki², Andrew Chacko³, M. Tariq Iqbal⁴

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

St. John's, Canada

¹saaquadri@mun.ca, ²mmgbaalbaki@mun.ca, ³achacko@mun.ca, ⁴tariq@mun.ca

Abstract—This paper investigates the system sizing, schema, modeling, and simulation of a 500 MW, grid-connected PV farm at a site close to the Churchill Falls Airport in Labrador. The objective is to understand the PV farm's technical and economic feasibility. The system is sized manually and with the help of PVWatts. The outputs from this calculation and PVWatts are used to select system components, such as solar panels and MPPT inverters. The plant is divided into four blocks of 125 MW each for ease of maintenance, control, and redundancy for planned and unplanned outages. Pooling transformers, high voltage switchgear, grid transformers, and the export transmission line are sized accordingly. System performance is analyzed using NREL's System Advisor Model (SAM) software. The grid-connected PV system, with the inverter, is modeled in Simulink (MATLAB) for dynamic simulations. Protection and control scenarios are also modeled in Simulink. The system's output parameters are then analyzed to understand power output and stability. Finally, recommendations on feasibility and further work are discussed.

Keywords—*photovoltaics, solar farm, MPPT, protection, control, grid, Simulink, MATLAB, SAM, PVWatts.*

I. INTRODUCTION

Presently, there is no large PV installation in Labrador, NL. The population of Labrador is more than 30,000. Electricity is mainly supplied from large hydropower plants. Remote communities are powered by diesel generators. Labrador has a large area that could be utilized for large-scale solar power generation. This paper presents the feasibility of a 500 MW solar power generation in Labrador. Large-scale solar power generation will help decrease greenhouse gas emissions. Capacity addition near the high voltage grid will help improve the utilization of transmission infrastructure and electricity export to inter-provincial and international markets.

This report includes an investigation of other literature on the implementation of grid-connected PV farms; an examination of the details of implementing a grid-connected PV farm at a location close to the Churchill Falls Airport in Labrador, Eastern Canada; modeling and dynamic simulations of the project in Simulink; and control and protection design. The report describes site selection, selection of the main components, system

sizing calculations, and the overall schematic. Modeling and simulation are performed using various software tools to derive the annual and monthly DC/AC energy production, system losses, and system behavior in varying conditions. This report aims to understand the technical and economic feasibility of a large-scale PV farm in Labrador, an area with low average solar irradiation.

II. LITERATURE REVIEW

A previous study was done to design a 50 MW utility-scale Solar PV farm in India [1]. The study objective was to design the complete power system, from PV panels to the overhead transmission line connection, to deliver the power to the 132 kV utility grid. The ratings selected for the PV panels and the string inverter were 330 Wp and 160 kW, respectively. The power plant was divided into eight blocks of 6.25 MW each. Within a block, a number of inverters were paralleled to feed an 800 V LT Panel. This power was stepped-up through a three-winding, inverter-duty transformer rated as 33 kV/0.8 kV/0.8 kV, Dy11y11, 5/6.25 MVA. The 33 kV output from each block was connected to a main 33 kV switchgear through a number of incomers. A single outgoing feeder of the 33 kV switchgear was connected to a 132 kV/33 kV, YNyn0, 32/40/50 MVA pooling substation transformer. The output of the pooling substation transformer was connected to the utility grid through an overhead link [1]. PV farm layouts, cable route layouts, transformer yard layouts, 800 V, 33 kV, and 132 kV Single Line Diagrams were prepared as the output of the study project [1]. PV Syst software was used for simulation in this project. Based on the selected configuration and with 45% overloading capacity of inverters, the total number of panels was calculated as 6864. The total land requirement for the project was estimated at 250 acres [1].

In their paper 'Design and Simulation of a 10 MW Photovoltaic Power Plant using MATLAB and Simulink', the authors describe the components of a PV farm power generation system [2] consisting of:

- (1) an array of PV panels grouped in series or parallel to achieve maximum power output
- (2) a DC-DC boost converter, used as a load regulator and to convert PV array output voltage to a voltage suitable for the inverter

- (3) a three-phase DC-AC converter to convert from DC to AC and to supply this electrical energy to the three phase grid, a three-phase step-up transformer to convert the low voltage output of the inverter to the voltage rating of the grid, and
- (4) a PV power generation system controller containing the Maximum Power Point Tracking (MPPT) controller for the DC-DC boost converter and inverter controller [2].

To design the PV array, they considered several manufacturers of PV cells, including Mitsubishi, Sharp, Sanyo, BP Solar, and Suntech. They selected Sanyo’s HIP-225HDE1 PV module with a maximum power rating of 225 Wp to fit PV array design requirements [2]. The authors used the Simulink module in MATLAB to simulate the system. The PV field was built in software using elements from the SimPowerSystems library. The Simulink model included

- (1) the Signal Builder Tool to build solar radiation and panel temperature signals to test the system under different conditions,
- (2) the PV farm with PV cells,
- (3) the DC-DC converter, along with the MPPT control system,
- (4) the three-phase inverter with its dedicated control system,
- (5) the step-up transformer to connect the inverter to the grid and the grid into which energy is transferred.

To simulate real-world conditions, the authors adjusted parameters such as solar irradiance and PV cell temperature to understand their corresponding effect on the PV array’s output power and voltage [2].

Keiichi K. et al. mentioned in [5] that the initial cost for a very large-scale PV plant is 1-3 MUSD/MW, with OPEX of 1% of the initial cost. The Levelized cost of energy (LCOE) for utility-scale PV generation globally is estimated at approximately 0.15 USD/kWh, 0.10 USD/kWh, and 0.05 USD/kWh for the higher initial cost case (3 MUSD/MW), middle case (2 MUSD/MW) and lower case (1 MUSD/MW), respectively. However, there are some differences depending on regional conditions.

III. SITE SELECTION

The site selection was based on various criteria such as solar irradiance, terrain, vicinity to the high voltage transmission line for power evacuation, and vicinity to means of transport. The selected location has the following characteristics and advantages. (1) Solar irradiation of 3.5 kWh/m²/day, (2) vicinity to high voltage grid for power evacuation with the option to connect to 230 kV, 138 kV, and 69 kV grids, (3) vicinity to means of transport viz. Trans-Labrador Highway 500 and the Churchill Falls Airport, (4) the selected site has a flat terrain, and (5) the Churchill Falls community and the Churchill Falls Hydro-Electric Substation are nearby, which can prove

advantageous for personnel management during project phases of construction, operation, and maintenance.

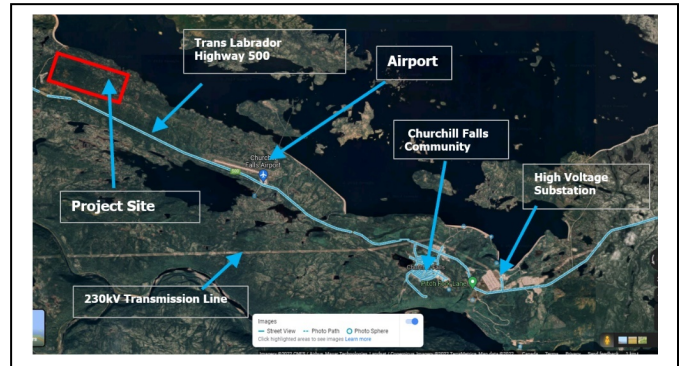


Fig. 1. 500 MW PV Farm Site Selection

IV. SYSTEM COMPONENTS

A. System Components

The following are the main components of the system. (1) photovoltaic modules, (2) DC combiner boxes, (3) inverters, (4) step-up pooling transformers, (5) pooling switchgears and (6) step-up grid transformers.

B. PV Module Selection

Based on utilization in previous utility-scale solar power plants and high-efficiency criteria, the Canadian Solar CS6U-340M PV module is selected. The module has a 17.49% efficiency with a rated output power of 340 W, open-circuit voltage of 37.9V, and a short circuit current of 9.48A.

C. Inverter Selection

Based on utilization in previous utility-scale solar projects, high power rating, and high efficiency, the ABB PVS980 2000 kVA-K inverter is selected. The inverter has a rated output power of 2000 kVA, output voltage of 660 V, maximum DC input power of 3200 W, maximum DC input voltage of 1500 V and MPPT voltage range between 935-1500 V at 35°C.

V. SYSTEM SIZING AND CALCULATION

A. PVWatts Calculation

After site selection, the preliminary kWh energy output from the proposed site was calculated using PVWatts, an online tool developed by National Renewable Energy Laboratory (NREL). The output results are shown in the table below.

TABLE I. PVWATTS CALCULATION RESULT

Property	Unit	Value
Land Requirement	km ²	3.3
Energy Output	kWh/yr	596078656
Average Annular Radiation	kW/m ² /day	4
Energy Value	\$	74629010
Capacity Factor	%	13.6

B. System Structure

The overall plant is divided into four blocks, with all blocks having an identical arrangement for ease of operation, maintenance, and control. The power system of each block starts from an individual PV module and ends at the 230 kV terminals of its grid transformer. To get the desired DC voltage per the inverter specification, 32 no. of PV modules are connected in series to form a string. Further, 288 strings are connected in parallel to supply input DC power to a single inverter. The scheme is designed such that each string voltage (operating and open circuit) shall be below the rated maximum input voltages allowed at the input of the inverter.

Also, the total input power of parallel strings shall be below the specified limit by the inverter vendor. Since the inverter can take only 24 DC inputs, the intermediate DC Combiner Boxes (DCBs) shall be installed between PV module strings and the inverters. Each DCB takes DC power from 12 strings and combines it in the form of a single DC output for further connection to the DC side of the inverter.

C. Power Evacuation

Due to the vicinity of the Churchill Falls Generating Station and its substation, there are three voltage-level options available for power evacuation: 230 kV, 138 kV, and 69 kV. Considering the criteria of shortest distance and low ampacity requirement, the existing 230 kV overhead transmission line is chosen for power evacuation. The direct length from the project site to the 230 kV transmission line is approx. 5 km as measured on google maps. Each block is designed to export power individually to the existing transmission grid. AC power output from the inverter is available at 60 Hz, 660 V level, which is connected to the low voltage primary side of the pooling transformers. There are twenty pooling transformers per block, each accepting power from two inverters. Pooling transformers are inverter-rated, three winding, 660 V/660 V/35 kV, DY5Y5, 7.5 MVA step-up transformers. Output from all 20 pooling transformers is combined at a 35 kV, 3-phase, 3-wire, gas-insulated, pooling switchgear. A single feeder from each 35 kV pooling switchgear feeds the power to the 35 kV / 230 kV, YNyn0, 125 MVA, step-up grid transformer. In each block, there is one pooling switchgear and a grid transformer (totaling 4). Finally, the 230 kV side of the grid transformers is connected to the existing transmission grid through a new 230 kV transmission link and air-insulated gantry structure. This link comprises of 3-wire, 4-circuit overhead transmission link (1-circuit per grid transformer).

D. System Schematic

The following figures show the schematic of the DC combiner box and the four system blocks that make up the 500 MW PV capacity.

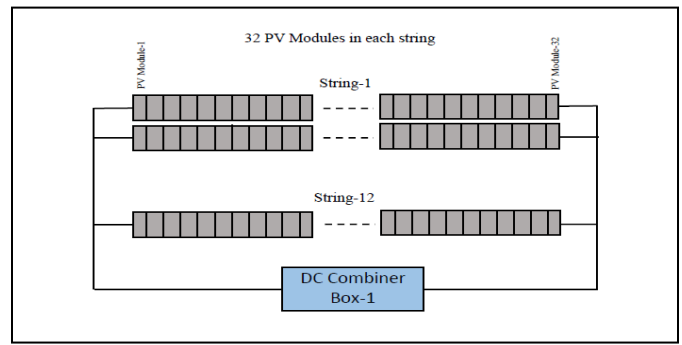


Fig. 2. Schematic of each DC combiner box

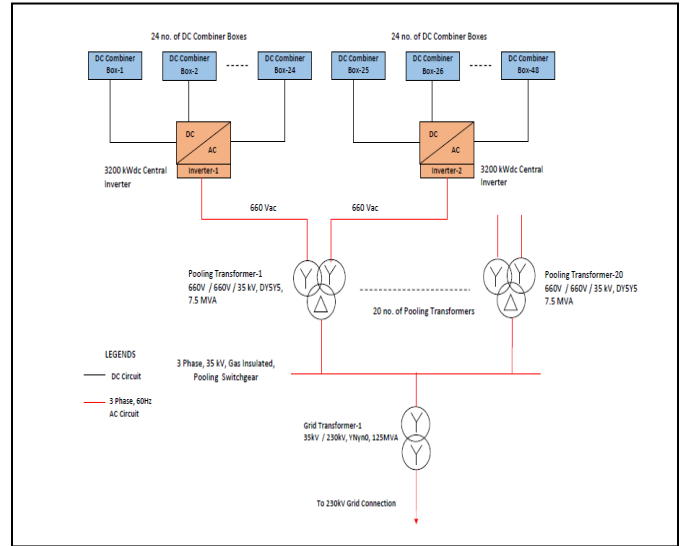


Fig. 3. Scheme diagram of each block

VI. SYSTEM ADVISOR MODEL ANALYSIS

The modeling and simulation were done in System Advisor Model (SAM), a free techno-economic program model that enables individuals in the renewable energy business such as policy analysts, researchers, project managers, engineers, and technology developers to make better decisions. SAM can simulate a wide range of renewable energy systems, including photovoltaic systems, battery storage, industrial process heat from parabolic, marine energy wave systems, wind power, solar water heating, geothermal power generation, fuel cells, biomass combustion, and high concentration photovoltaic systems. The system designed is located near the Churchill Falls Airport, and the resources were downloaded in the SAM software from National Solar Radiation Database (NSRDB), as shown below [3].

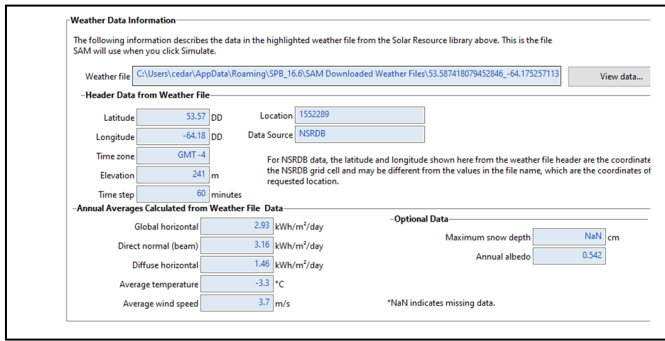


Fig. 4. Selected site's weather data in SAM UI

A. PV Module Analysis

The PV solar panel model used is Canadian Solar Max Power CS6U-340M with an output rated power of 340 Watts and 17.485% efficiency, while in the SAM software, the output power was calculated to be 339.963 Watts. The model was not available in the software library; hence it was manually entered.

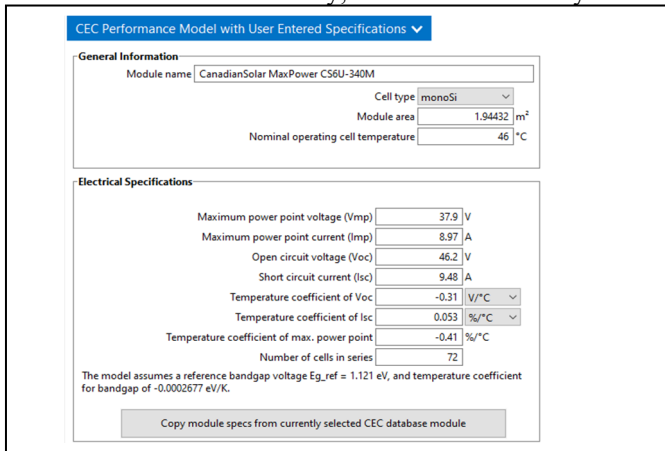


Fig. 5. CS6U-340M module specifications in the SAM UI

B. Inverter Analysis

The inverter model used is ABB's PVS980-58-2000 kVA-K. The model wasn't available in the library of the software, so it was manually entered.

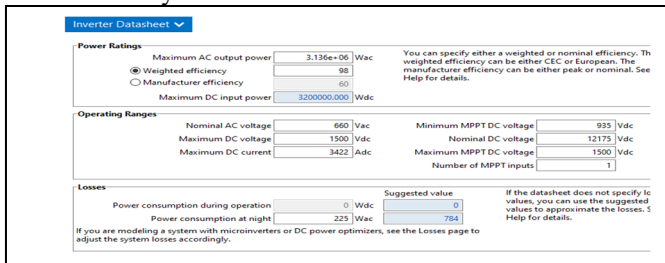


Fig. 6. PVS980-58-2000 kVA-K inverter specifications in the SAM UI

C. System Design

The system design uses the following parameters.

TABLE II. SYSTEM PARAMETERS

Property	Value
Number of inverters	250
Nameplate DC capacity	499,935.989 kW
Total AC capacity	784,000 kW
Total inverter DC capacity	800,000 kW
Number of Modules (PV)	1,470,560
Number of strings	45,955

Several losses are taken into consideration.

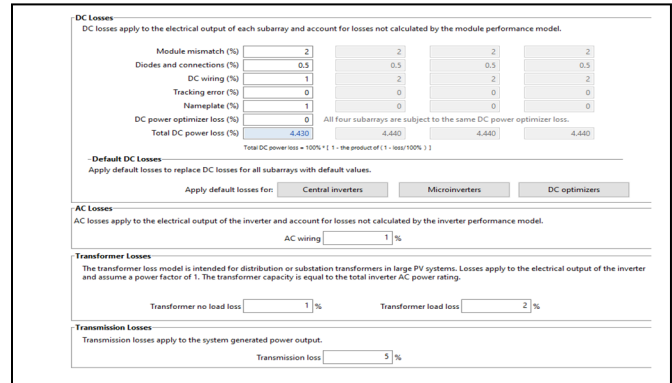


Fig. 7. System losses entered in SAM UI

D. Analysis Results

Results from running the system simulation in SAM are shown below.

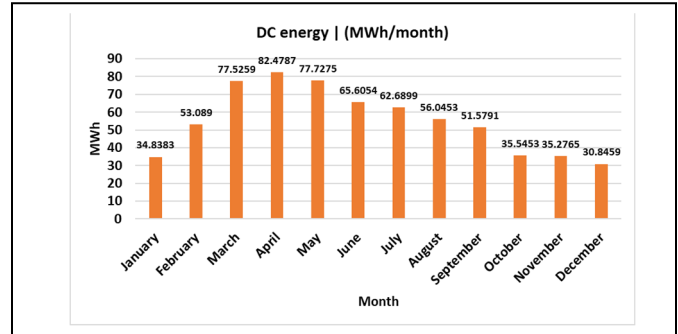


Fig. 8. Annual DC energy production

As expected, DC energy production is higher during the spring and summer months, where days are longer, and lower where winter days are shorter. The annual AC energy production and average hourly energy production charts are shown below.

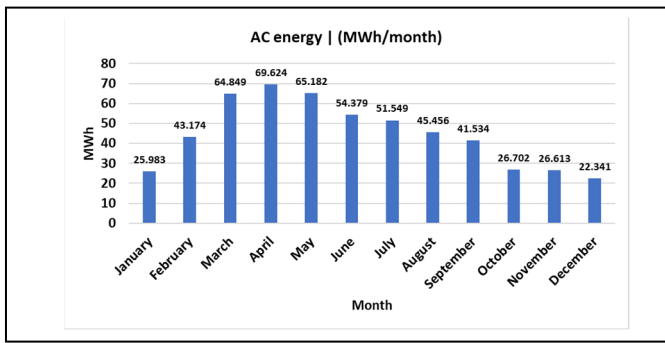


Fig. 9. Annual AC energy production

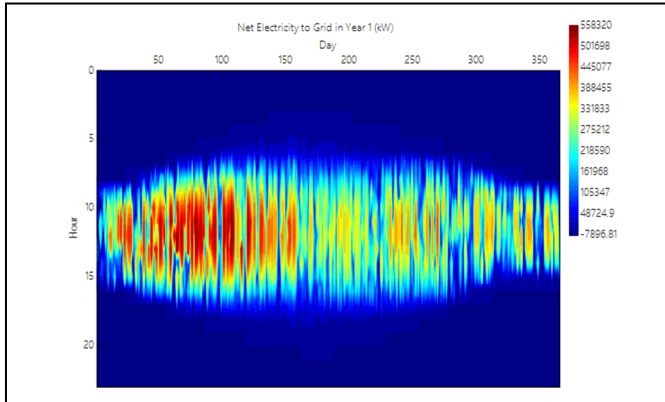


Fig. 10. Energy production averaged hourly over the year.

VII. SYSTEM DYNAMIC SIMULATION

The objective of the dynamic simulation is to simulate PV array output under the MPPT regime and plot the output characteristics on a time scale. The PV output voltage needs to be controlled to get the maximum power at any given irradiance. Also, since the power output from the PV arrays is DC, it needs to be converted and filtered to get 60 Hz AC for the purpose of exporting to the grid. These two functionalities are executed through DC-DC converters and DC-AC inverters, respectively. For this study, central inverter topology is considered, wherein the number of PV strings is paralleled as the input to an inverter.

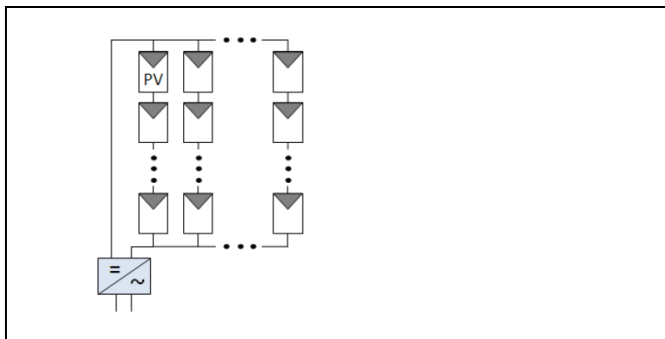


Fig. 11. Central inverter topology

As per the ABB datasheet, each inverter has one MPPT. For dynamic simulation purposes, DC Combiner Box level PV array size is considered, i.e., 12 strings with 32 modules in each string. The selected PV module (Canadian Solar, CS6U-340M) is readily available in the Simulink library. The irradiance is

considered as variable between 600 W/m² and 250 W/m², whereas the temperature was kept constant at -4.57 °C (the yearly average temperature at the project location), respectively. PV array output is connected to a DC-DC boost converter. Boost converter output voltage is stabilized to 1200 V by selecting the value of Vnom_dc as 1200 V in the VSC controller block. This is based on the mean value of the allowed input range for ABB inverter (935 to 1500 V DC).

The MPPT algorithm in the Simulink model is of Incremental Conductance type. DC output from boost converter is converted to AC through a DC-AC, 3 phase inverter and further stepped up by a Ygd11, 660 V/35 kV transformer. To match with PV array size modeled in Simulink, the transformer rating is selected as 3500 kVA for this simulation. The inverter output filter capacitance rating is 10% of the transformer rating, i.e., 350 kVAR. A swing bus with ratings of 235 kV, 2500 MVA is modeled along with a 30 MW, 2 MVar local load to represent the grid. The simulation was run for a period of 2.5 sec. The voltage, current, and power plots are included below.

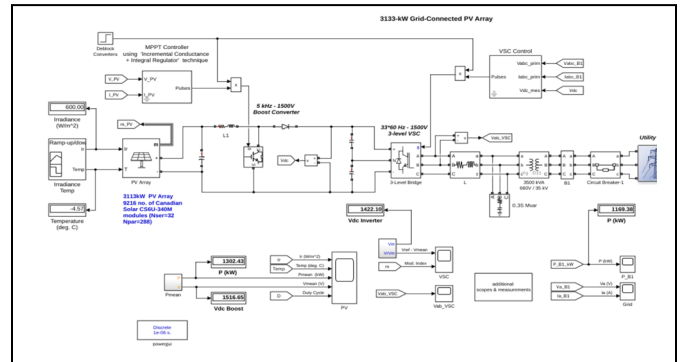


Fig. 12. MATLAB Simulink model

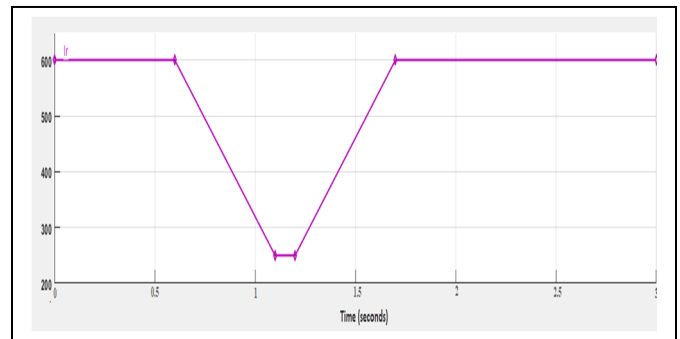


Fig. 13. Variation in solar irradiance

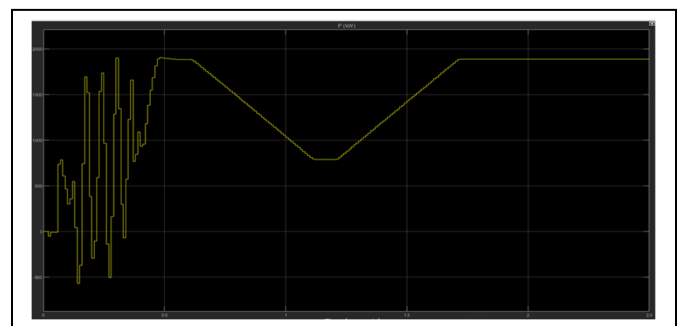


Fig. 14. Power output on 35 kV side of Pooling Transformer

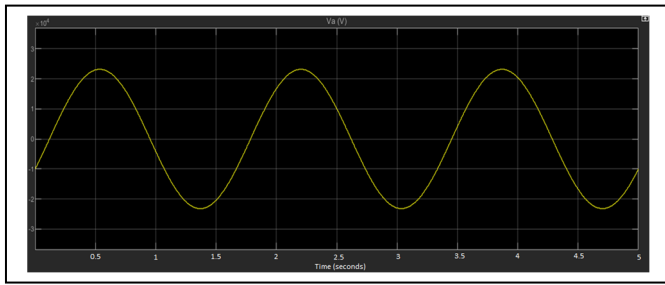


Fig. 15. Voltage output on 35 kV side of Pooling Transformer

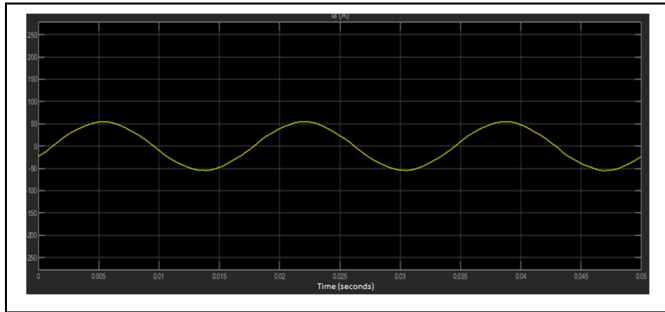


Fig. 16. Current output at the 35 kV side of Pooling Transformer

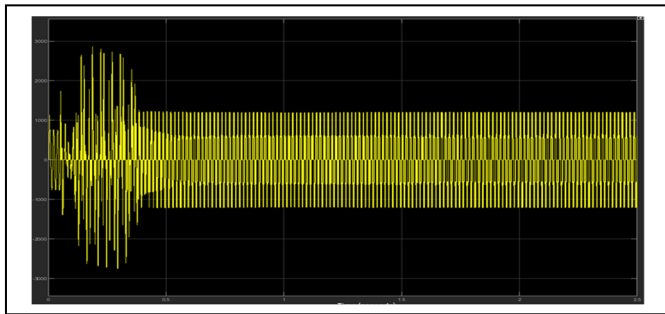


Fig. 17. Inverter output voltage waveform

As the result of the simulation, the PV array output is stabilized at 1945 kW, at 1.7 sec., with the input of 600 W/m² irradiance and -4.57°C temperature. The generated PV array voltage (V_{mean}) stabilized at 1210 V. The boost converter stabilizes the voltage to 1200 V, which is less than the maximum allowed voltage limit of 1500 V DC at ABBs PVS980 inverter incoming side. The 0.35 MVar capacitor is used at the output port of the inverter to remove the harmonics. The power exported on the 35 kV side of the pooling transformer is 1914 kW at the end of the 2.5-sec simulation duration.

TABLE III. SIMULATION RESULTS

Parameter	Result
PV capacity connected	3133 kW
Solar power generated	1945 kW
Power exported	1886 kW
PV voltage output	1210 V DC
Inverter output voltage	683V RMS ac
Voltage output for export	16.38 kV (RMS ac)

Parameter	Result
Overall simulation efficiency	60.5%

VIII. SYTEM PROTECTION AND CONTROL

The objective of this section is to describe protection and control requirements during variations in inputs or in the grid conditions. Earlier configuration of grid is retained, which includes a swing bus of 235 kV, 2500 MVA_{sc}, and a local load at Churchill Falls substation of rating 235 kV, 30 MW, 2 MVar.

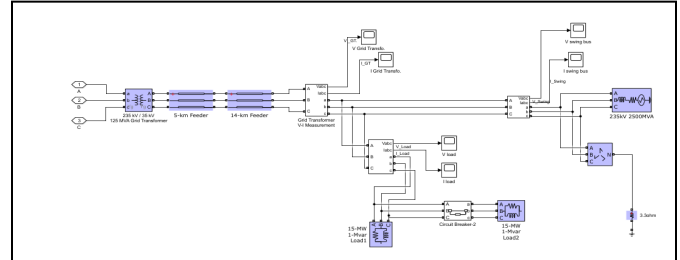


Fig. 18. Simulink model of utility grid and local load

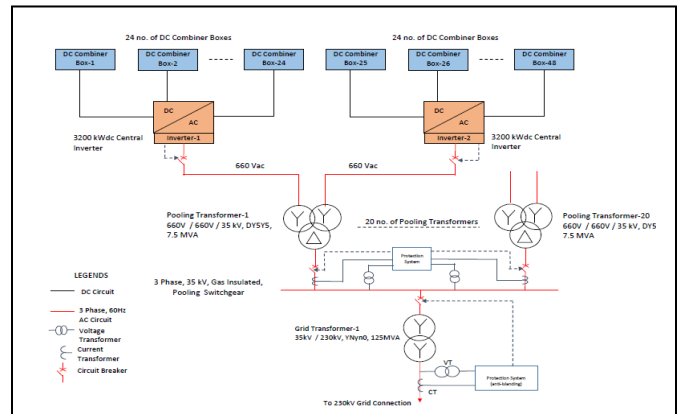


Fig. 19. Scheme diagram with protection and control system

The following cases are simulated – (1) variation in load, (2) grid islanding, and (3) variation in capacitance.

A. Variation in Load

For this simulation, the case load is split into two equal blocks, each of 15 MW, 1 MVar. One of the loads is switched off at 1 sec by opening the Circuit Breaker-2. The simulation is run for 2.5 sec. Irradiance and temperature are kept constant at 600 W/m² and -4.57°C, respectively.

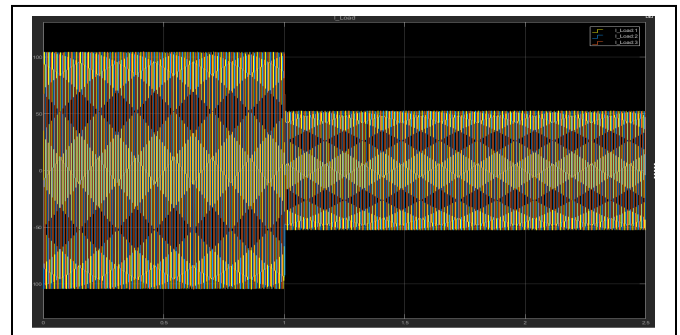


Fig. 20. Local-load current profile

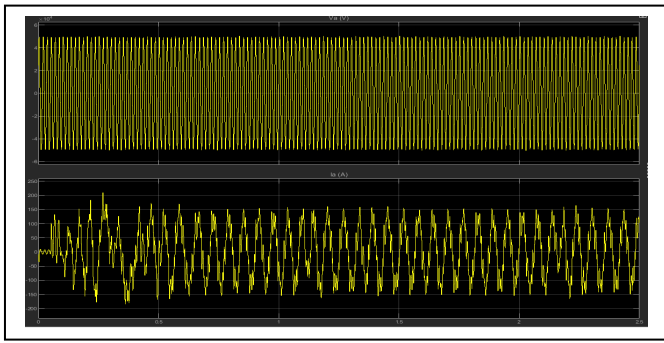


Fig. 21. Voltage (phase to phase) and current at 35 kV side of pooling transformer

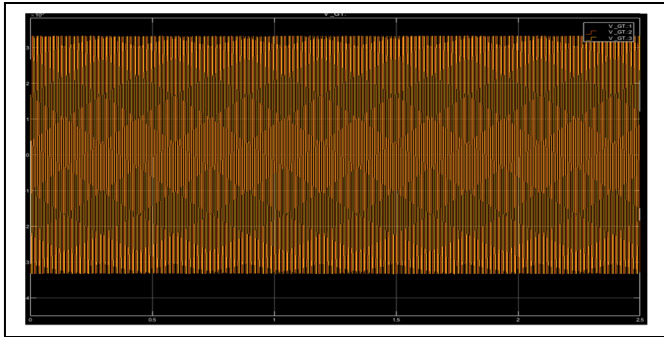


Fig. 22. 235 kV Transmission Line voltage (ph-ph)

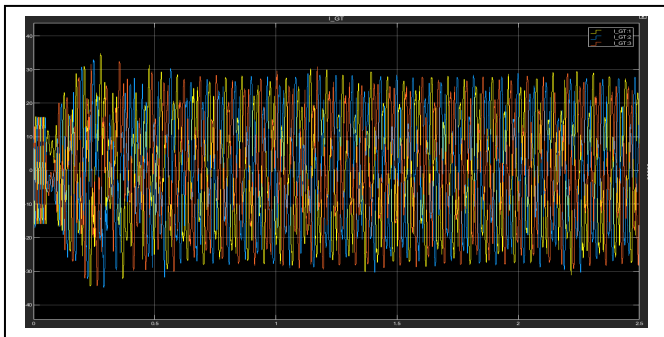


Fig. 23. Transmission Line current, on 235 kV side of Grid Transformer

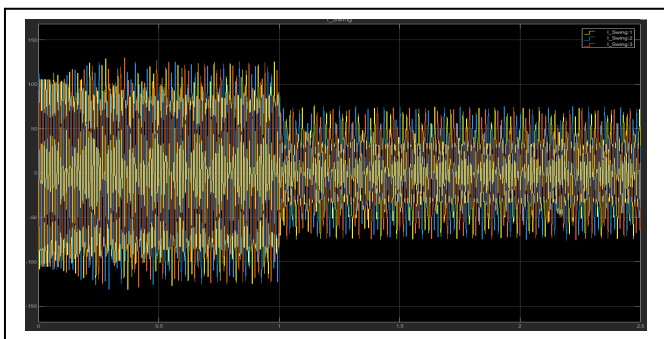


Fig. 24. Transmission line current, on the swing bus side

It is observed that 50% load loss does not affect the load shared (i.e., current) by the PV plant. However, the current profile measured at the swing bus side (I_{Swing}) drops to half the value. This denotes that, since the rated capacity considered

for the PV system (3133 kW) and the Grid Transformer (125 MVA) are much weaker than the swing bus, the majority of load current is supplied by the swing bus. Also, even during loss of 50% of the load, the voltage of the transmission line is held stable by the swing bus at close to nominal capacity, i.e., 235 kV ph-ph. (Note- Current value from swing bus is higher than load bus as swing bus also supplies the charging currents for transmission lines and the grid transformer. In this case, since the voltage and current at the output of the PV system (i.e., on the 35 kV side of pooling transformer) are stable and within range, the protection system does not need to operate or isolate the PV system.

B. Grid Islanding

In the islanding scenario, the grid is lost completely on the occurrence of some major disturbances on the grid or unexpected opening of the connecting circuit breaker. To simulate this case, Circuit Breaker-1 on the 35 kV side of the pooling transformer is made open at 1 sec. This results in the isolation of the grid while the PV system is still generating power.

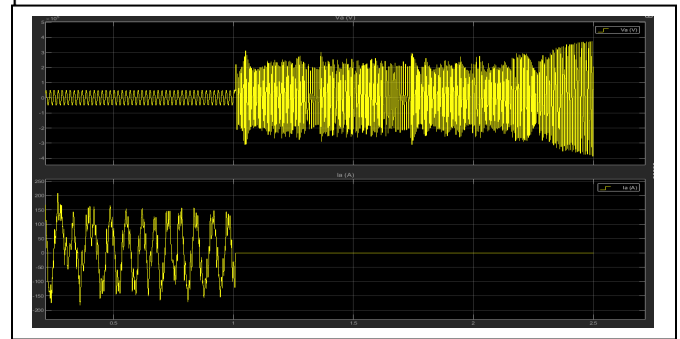


Fig. 25. Voltage and current on 35 kV side of pooling transformer

Around 0.5 sec, the PV system output current stabilizes with an RMS value of 85.59A between 0 to 1 sec. However, on the occurrence of grid loss at 1 sec., the PV system current output goes down to zero. On the other hand, voltage is held stable at 34.97 kV from 0 sec to 1 sec. After 1 sec., at the grid loss instant, overvoltage scenario occurs with the major rise from 2.25 sec. onwards. Generally, power system equipment is suitable for 10% overvoltage in normal running conditions. As observed from graphs, grid islanding poses a major overvoltage abnormality. Near the end of the simulation cycle, values as high as 235 kV can be seen on 35 kV rated secondary of the pooling transformer. To avoid this scenario, fast-acting, anti-islanding protection is necessary. Protection relays on the PV system side shall get inputs from transmission line voltage and current transformer to monitor the healthiness of the grid. On the occurrence of loss of grid or grid-fault, the protection system should act instantaneously to shut down the generation. Solar power plants shall be isolated to avoid supplying power to the faulty grid and thereby safeguarded from major insulation failures.

C. Variation in Capacitance

The output filter capacitance of the inverter is reduced to 0.15 MVAR from the normal value of 0.35 MVAR.

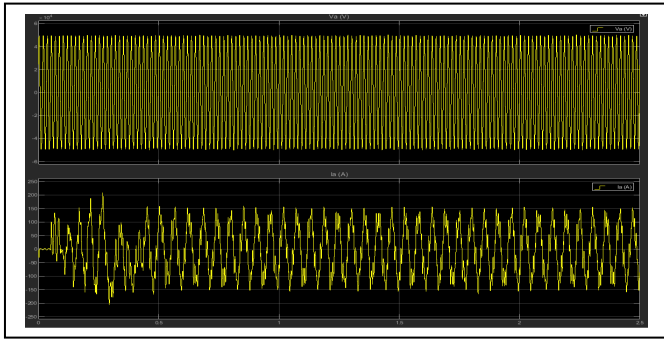


Fig. 26. Voltage and current on 35kV side of pooling transformer

Since no significant excursions were observed from normal voltage and current values, a protection system is not required to operate in reduced capacitance of the inverter filter. However, reduced capacitance may give rise to current harmonics.

IX. RESULTS AND DISCUSSION

This research investigated the system sizing, schema, modeling, and simulation of a 500 MW, grid-connected PV farm at a site close to the Churchill Falls Airport in Labrador to understand the PV farm's technical and economic feasibility. The system was sized using manual calculations and the PVWatts. A solar PV module and an MPPT inverter were selected using this information. The overall plant was divided into four blocks of 125 MW each for ease of maintenance, control, and redundancy in case of planned and unplanned outages. Pooling transformers, high voltage switchgear, grid transformers, and the export transmission line was sized accordingly. System performance was analyzed using System Advisor Model. A grid-connected PV system with the inverter was modeled in Simulink, MATLAB for dynamic simulations. The grid was modeled as a swing bus along with a local load.

System output parameters were analyzed for varying conditions.

In Canada, the average installation cost of a PV farm is \$1.25/Watt [4]. For this 500 MW system, the installation cost would be approximately \$625 million, with additional operational and maintenance costs over the system's lifespan. The land area requirement for this system, coming up to nearly 3.33 sq. km, is quite large. Our observation is that utility-scale PV plants are more practical in provinces with higher average annual solar irradiance, like Ontario and Alberta. Future technological improvements in PV cell efficiency could make this installation more practical in areas with lower average annual solar irradiance, such as in Labrador.

Future work in this project will include tuning the inverter output filters to reduce THD (total harmonic distortions), performing a detailed sizing of other equipment such as DC & AC cables, transmission lines; plant layout design; and detailed cost analysis of the solar power plant.

REFERENCES

- [1] Shah, Dr. Sweta; Hindocha, Krunal; "Design of 50 MW Grid Connected Solar Power Plant," International Journal of Engineering and Technology, vol. 09, no. 04, pp. 885-897, 2020.
- [2] D. Popa, M. Nicolae, P. Nicolae and M. Popescu, "Design and simulation of a 10 MW photovoltaic power plant using MATLAB and Simulink," 2016 IEEE International Power Electronics and Motion Control Conference (PEMC), 2016, pp. 378-383.
- [3] "Provincial and Territorial Energy Profiles - Newfoundland and Labrador," Canada Energy Regulator, 17 March 2021.
- [4] "CER – Economics of Solar Power in Canada – Appendix A: Methods", Cer-rec.gc.ca, 2017.
- [5] Komoto, K., Ehara, T., Xu, H., Lv, F., Wang, S., Sinha, P., Cunow, E., Wade, A., Faiman, D., Araki, K., Perez, M., Megherbi, K., Enebish, N., Breyer, C. and Bogdanov, D., 2015. Energy from the Desert: Very Large Scale PV Power Plants for Shifting to Renewable Energy Future.

Coordinated Motion and Force Control of Multi-Rover Robotics System with Mecanum Wheels

S. Kalaycioglu
 Department of Aerospace
 Engineering
 Toronto Metropolitan University
 Toronto, Canada
 skalay@ryerson.ca

A. de Ruiter
 Department of Aerospace
 Engineering
 Toronto Metropolitan University
 Toronto, Canada
 aderuiter@ryerson.ca

Abstract—This paper presents a novel optimal control algorithm for coordinated force and motion control of multi rover robotics system with mecanum wheels while manipulating a common payload. Such a system with kinematical rolling conditions lead to non-holonomic constraints. The proposed control algorithm focuses on the minimization of joint torques, the rover-mecanum wheel moments as well as the contact force / moments made with the payload. A quadratic cost function in terms of the joint torques, the wheel moments and the contact forces and moments are minimized to overcome the so called joint torque saturation problem commonly seen while manipulating a common payload and also to provide an optimum solution for such an underdetermined system with non-holonomic constraints. Furthermore, the proposed control algorithm provides an on-line trajectory generation capability while manipulating a common payload for both the rovers and the arms simultaneously. The computer simulation results show that the control algorithm works efficiently and the minimum joint torques, and the contact forces and moments can be obtained while the end-effectors are manipulating and tracking a desired payload trajectory.

Keywords— *optimal control, coordinated motion and force control, multi rover system, mecanum wheels*

I. INTRODUCTION

There has been a considerable amount of interest in mobile rovers working in complex environments including space, mining and construction and military.

Initial technological challenges included mechanical design, especially related to the mechanics of locomotion, collision free trajectory generation, dynamic control of rover and mounted robotics manipulators.

Necsulescu et al [1] and [2] studied free and contact motion for rovers and developed impedance control techniques for real-time collision free motion generation as well as force control.

Control of mobile rovers subject to nonholonomic constraints were demonstrated using Differential Wheeled Mobile Robots (DWMR) in [3] and [4]. These nonholonomic constraints are often experienced when the kinematic constraints cannot be written in terms of time derivatives of some functions of the generalized coordinates.

The control of a mechanical system with nonholonomic constraints has been studied extensively and quite often,

kinematic control is designed ignoring the dynamics of the system [5]. It has been illustrated in [6] that a mechanical system with nonholonomic constraints can still be controlled despite the structure of the nonholonomic constraints. It was also demonstrated that a nonholonomic system cannot be stabilized to a single equilibrium point by a smooth time-invariant feedback [7].

Control of Multiple Robot Manipulators with Optimal Force Distribution was demonstrated in Kalaycioglu [8]. However, this study did not include rovers and was limited to two cooperating arms.

Recently, mechanics of wheeled locomotion drew considerable amounts of attention [9] – [14]. Several studies focused on the kinematics and dynamics modeling of the mecanum wheel, a special type of omnidirectional wheel [15] – [20].

Although, there is a compilation of extensive research studies, publications, and system development in the areas of single rover trajectory control, motion control of the arm, etc., the research is still at its infancy when it comes to payload sharing multiple rovers and mounted arms, their coordinated real-time collision-free trajectory generations and gross and fine control of the motions while operating a common payload.

The paper is structured into four sections. Section II presents the theoretical formulations, including the kinematics, dynamics model of the compound system which consists of two rovers and two n-degree redundant manipulators and a common payload. An optimal control algorithm is formulated by minimizing joint torques, wheel moments and the contact force/moments, respectively. Simulation results and discussion are presented in Section III and some concluding remarks are provided in Section IV.

II. THEORETICAL FORMULATIONS

A. Description of the System

The system under consideration consists of two identical four-wheeled mobile rovers with mecanum wheels and two n-DOF redundant manipulators which are mounted on these rovers operating a common payload. An example of the system of mobile rovers with two n-degree robots is illustrated in Fig. 1.

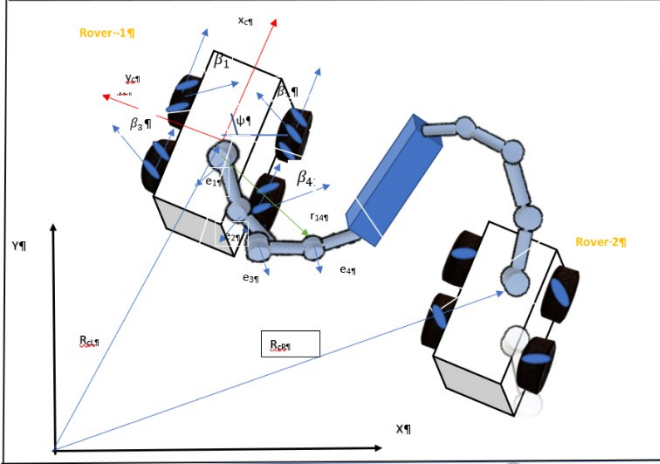


Fig. 1. Rover Robotics System Description

The system parameters of the robots and rovers used in the simulations are provided in Table 1. Let the R_{ci} , the position vector and the orientation angle ψ_i define the location of the centre of mass C_i of the rover- i with respect to the inertial coordinate system, X, Y, Z . A set of coordinate axes x_{ci}, y_{ci}, z_{ci} fixed to the center of the rover is obtained from the inertial coordinate system through a rotation around Z -axis with an angle of ψ_i .

The masses of the rover and the wheels are denoted by m_{ci} and m_{wij} while the distances between the centers of wheels along the y_{ci} and x_{ci} -axes are defined by $2a$ and $2b$, respectively.

Each mecanum wheel has a radius of s while the angle of rotation and the angular rate for each wheel are identified as ϕ_{ij} and ω_{ij} respectively, $j=1...4$. As shown in Fig. 1, the mecanum wheels have rollers attached to their outer rims. The angle between the x_{ci} and the axis of rotation of the roller is defined by β_{ij} for each wheel, $j=1...4$ and $i=1,2$ for each rover.

B. Kinematics Model of the System

The velocity of the centre of the wheel, V_{mi} can be calculated by

$$V_{mij} = V_{ci} + \Omega_{ci} \times r_{wij} \tag{1}$$

$$\Omega_{ci} = \dot{\psi}_i e_z \tag{2}$$

where V_{ci} is the velocity of the rover's centre of mass, Ω_{ci} is the angular rate of the rover along the z_{ci} , and e_z is the unit vector along the z_{ci} and finally r_{wij} is the displacement vector measured from the rover's centre of mass to the centre of the wheel.

The velocity of the point P, the centre of the roller can be calculated as

$$V_{pij} = V_{mij} + \omega_{ij} \times \rho \tag{3}$$

where ρ is the displacement vector measured from the centre of the wheel to the centre of the roller, point P.

If there is no slip, V_{pj} can not have a component along the axis of roller rotation $e_{\beta j}$, namely

$$V_{pij} \cdot e_{\beta ij} = 0 \tag{4}$$

where $e_{\beta ij}$ is the unit vector along the axis of the roller rotation. Substituting (3) into (4), carrying out the cross product, one can obtain the following relationship:

$$\begin{aligned} V_{mij} \cdot e_{\beta ij} + (\omega_{ij} \times \rho) \cdot e_{\beta ij} &= 0 \\ (\omega_{ij} \times \rho) &= -\omega_{ij} s e_x \\ V_{mij} \cdot e_{\beta ij} &= s (e_x \cdot e_{\beta ij}) \end{aligned} \tag{5}$$

where e_x is the unit vector along the x_{ci} axis and s is the radius of the wheel.

Also, substituting (1) into (5), one can rewrite the constraint equation as follows:

$$\begin{aligned} V_{ci} \cdot e_{\beta ij} + (\Omega_{ci} \times r_{wij}) \cdot e_{\beta ij} &= \omega_{ij} s (e_x \cdot e_{\beta ij}) \\ V_{ci} \cdot e_{\beta ij} + (r_{wij} \times e_{\beta ij}) \cdot \Omega_{ci} &= \omega_{ij} s \cos(\beta_{ij}) \end{aligned} \tag{6}$$

where β_{ij} is the angle between the unit vectors of e_x and $e_{\beta ij}$

$$\begin{aligned} e_{\beta i1}^T &= [\cos(\beta_{i1}), -\sin(\beta_{i1}), 0] \\ e_{\beta i2}^T &= [\cos(\beta_{i2}), \sin(\beta_{i2}), 0] \\ e_{\beta i3}^T &= [\cos(\beta_{i3}), \sin(\beta_{i3}), 0] \\ e_{\beta i4}^T &= [\cos(\beta_{i4}), -\sin(\beta_{i4}), 0] \\ r_{wi1}^T &= [a, b, 0] \\ r_{wi2}^T &= [a, -b, 0] \\ r_{wi3}^T &= [-a, b, 0] \\ r_{wi4}^T &= [-a, -b, 0] \end{aligned} \tag{7}$$

Substituting (7) into (6) and assigning β_{ij} as $\pi/4$, one can obtain the following equations:

$$V_{ci} = \begin{bmatrix} V_{cxi} \\ V_{cyi} \\ V_{czi} \end{bmatrix} = \begin{bmatrix} s(\omega_{i1} + \omega_{i2})/2 \\ s(\omega_{i3} - \omega_{i1})/2 \\ 0 \end{bmatrix}$$

$$\Omega_{ci} = \begin{bmatrix} \Omega_{cxi} \\ \Omega_{cyi} \\ \Omega_{czi} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ s(\omega_{i3} - \omega_{i1})/(2(a+b)) \end{bmatrix}$$

$$\omega_4 = \omega_1 + \omega_2 - \omega_3 \quad (8)$$

The rotational transformation from the inertial to the rover body axes is represented by the following rotation matrix:

$$\Psi_z = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (9)$$

Similarly, each homogeneous transformation matrix T_i^j that transforms the coordinates of a point from frame j to frame i on a robot manipulator is calculated using Denavit-Hartenberg (D-H) convention.

$$T_i^j = \underbrace{A_{i+1}} \underbrace{A_{i+2}} \dots \underbrace{A_{j-1}} A_j \quad i < j$$

$$A_i = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \cos\alpha_i & \sin\theta_i \sin\alpha_i & a_i \cos\theta_i \\ \sin\theta_i & \cos\theta_i \cos\alpha_i & -\cos\theta_i \sin\alpha_i & a_i \sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

where the four quantities $\theta_i, \alpha_i, d_i, a_i$ are parameters of link-i and joint-i and a_i is the length of the link, α_i is the twist, d_i is the offset while θ_i is the joint angle.

The Jacobian matrices and their first-time derivatives between the centre of the rover and a point-k on the robotics manipulator can also be calculated as follows:

$$\begin{pmatrix} \dot{j}_c^k \\ \dot{L}_c^k \end{pmatrix}_L = \begin{bmatrix} e_z & e_1 & \dots & e_5 & e_7 \\ e_z \times r_{ck} & e_1 \times r_{1k} & \dots & e_5 \times r_{5k} & e_7 \times r_{7k} \end{bmatrix} \quad (11)$$

$$\begin{pmatrix} j_c^k \\ L_c^k \end{pmatrix}_L = \begin{bmatrix} e_z & e_1 & \dots & e_7 \\ e_z \times (\Omega_c \times r_{ck}) & e_1 \times (\theta_1 \times r_{1k}) & \dots & e_7 \times (\theta_7 \times r_{7k}) \end{bmatrix} \quad (12)$$

where e_z is the unit vector along Ω_c rover's fixed rotation axis, e_i is the unit vector along the rotation axis of the i^{th} joint and r_{ck}, r_{ik} are the displacement vectors from the centre of rover and i^{th} joint to the point k, respectively.

The angular and linear velocities and accelerations of any point on the first rover arm can be determined by using the Jacobian matrix:

$$\begin{bmatrix} \Omega_k \\ V_k \end{bmatrix}^L = \begin{pmatrix} j_c^k \\ L_c^k \end{pmatrix}_L \begin{bmatrix} \Omega_c \\ V_{cL} \end{bmatrix} + \begin{bmatrix} 0 \\ V_{cL} \end{bmatrix} \quad (13)$$

$$\begin{bmatrix} \dot{\Omega}_k \\ \dot{V}_k \end{bmatrix}^L = \begin{pmatrix} j_c^k \\ L_c^k \end{pmatrix}_L \begin{bmatrix} \dot{\Omega}_c \\ \dot{\theta}_L \end{bmatrix} + \begin{pmatrix} j_c^k \\ L_c^k \end{pmatrix}_L \begin{bmatrix} \Omega_c \\ \theta_L \end{bmatrix} + \begin{bmatrix} 0 \\ V_{cL} \end{bmatrix} \quad (14)$$

where $\dot{\theta}_L^T = [\dot{\theta}_1, \dot{\theta}_2, \dot{\theta}_3, \dots, \dot{\theta}_n]$ is the first rover-arm joint rates, Ω_k and V_k are the (3x1) angular and linear velocity vectors of the k point on the first arm, respectively. Again, the angular and linear velocity of any point of the second arm can also be calculated by simply changing the index from L to R.

C. Dynamics Model of the System

The Lagrangian formulation is employed to obtain the dynamics equations of motions of the robotics system. The total kinetic energy of the system T_t can be calculated by adding both the translational and rotational kinetic energies of the rovers and the robot manipulators.

$$T_t = T_{tra} + T_{rot} \quad (15)$$

The translational and rotational velocities of the centre of mass of the rovers and the robot links are provided in (8) and (14). These expressions are substituted into (15) to obtain the total kinetic energy of the system.

The Lagrangian equations can be written as

$$\frac{d}{dt} \left(\frac{\partial T_t}{\partial \dot{q}_n} \right) - \frac{\partial T_t}{\partial q_n} = Q_n, \quad n = 1, \dots, 2m \quad (16)$$

where q_n are the generalized coordinates as

$q^T = [\phi_{1L}, \phi_{2L}, \phi_{3L}, \phi_{1R}, \phi_{2R}, \phi_{3R}, \theta_{1L}, \dots, \theta_{nL}, \theta_{1R}, \dots, \theta_{nR}]$ and Q_n are the generalized forces and $m = (n+3)$, n is the number of degrees of freedom of the redundant manipulators.

After substituting (15) into (16) and carrying out the differentiation, one can obtain the following equations of motions for the system.

$$\begin{bmatrix} G_{WL} & G_{WLR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_L \\ G_{WLR}^T & G_{WR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_R \\ G_{W\theta L}^T & G_{W\theta L}^T & G_{\theta L} & G_{\theta LR} & \ddot{\theta}_L \\ G_{W\theta R}^T & G_{W\theta R}^T & G_{\theta LR}^T & G_{\theta R} & \ddot{\theta}_R \end{bmatrix} \begin{bmatrix} c_L \\ c_R \\ c_{\theta L} \\ c_{\theta R} \end{bmatrix} + \begin{bmatrix} \dot{\Omega}_L \\ \dot{\Omega}_R \\ \ddot{\theta}_L \\ \ddot{\theta}_R \end{bmatrix} = \begin{bmatrix} M_L \\ M_R \\ \tau_{\theta L} \\ \tau_{\theta R} \end{bmatrix} \quad (17)$$

where \underline{G} , the inertia / mass matrix is a positive definite matrix and, $\dot{\Omega}_L, \dot{\Omega}_R$, are the angular accelerations of the wheels of the two rovers, $\ddot{\theta}_L, \ddot{\theta}_R$ are the joint angular accelerations for the two arms respectively. $c_L, c_R, c_{\theta L}$, and $c_{\theta R}$ are the non-linear terms while M_L, M_R are the wheel control moments for the rovers and $\tau_{\theta L}, \tau_{\theta R}$ are the joint control torques for the arms, respectively.

where $\Omega_L^T = [\omega_{L1}, \omega_{L2}, \omega_{L3}]$ consists of the angular rates of the three wheels of the first rover, similarly Ω_R^T can be written in terms of the second rover's wheel angular rates. The

fourth wheel angular rate is dependent on the first three if there is no slip as shown in (8). The sub-index L is referred to the first and R is to the second rover and robot manipulator, respectively.

D. A Novel Optimal Control Technique

The system consisting of two redundant robot manipulators mounted on two rovers manipulating a common payload is mathematically an underdetermined system due to the redundant number of actuators and sensors available to control the rotational and translational motions of the rovers and the links.

In this section, a new two-stage optimal control technique is developed and demonstrated in a control system block diagram in Fig. 2.

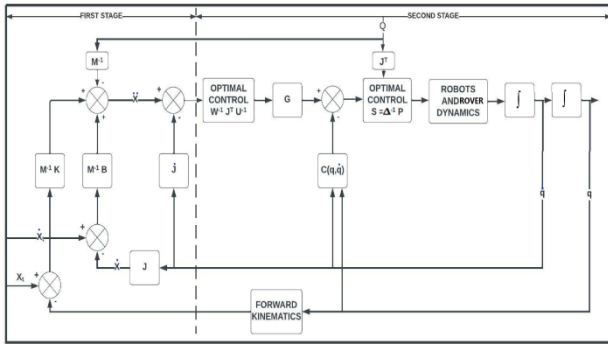


Fig. 2. Two stage optimal control system block diagram

The first stage of the block diagram consists of a well-known impedance control technique which facilitates the generation of the desired trajectories for the two end-effectors for a given trajectory of the payload and is presented in (18).

$$\ddot{X}_i = \underline{M}_i^{-1} \underline{B} \{ \dot{X}_{di} - \ddot{X}_i \} + \underline{M}_i^{-1} \underline{K} \{ \tilde{X}_{di} - \tilde{X}_i \}, \quad i = L, R \tag{18}$$

where, \underline{M}_i , \underline{B} , \underline{K} are 6x6 positive definite impedance matrices and can be selected to meet the requirements of the tracking performance while \tilde{X}_i . (i varies between L and R for each arm) are the desired end-effector trajectories for the two arms and \tilde{X}_{di} are the given trajectories for the arm attachment point on the payload. The matrix selection process will not be discussed here for the sake of brevity.

Equations (13) and (14) can be rewritten by substituting Ω_L and Ω_R and eliminating Ω_C and V_C from the equations.

$$\tilde{X}_L = \begin{bmatrix} \Omega_k \\ V_k \end{bmatrix}^L = \begin{bmatrix} J_c^k \\ \end{bmatrix}_L \begin{bmatrix} \Omega_L \\ \dot{\theta}_L \end{bmatrix} \tag{19}$$

$$\dot{\tilde{X}}_L = \begin{bmatrix} \dot{\Omega}_k \\ \dot{V}_k \end{bmatrix}^L = \begin{bmatrix} J_c^k \\ \end{bmatrix}_L \begin{bmatrix} \dot{\Omega}_L \\ \dot{\theta}_L \end{bmatrix} + \begin{bmatrix} J_c^k \\ \end{bmatrix}_L \begin{bmatrix} \Omega_L \\ \dot{\theta}_L \end{bmatrix} \tag{20}$$

Carrying out the least-square minimization of joint rates, one can solve for the inverse kinematics to obtain Ω_L and $\dot{\theta}_L$ and their time derivatives from (19) and (20).

$$\begin{bmatrix} \Omega_i \\ \dot{\theta}_i \end{bmatrix} = \underline{W}_i^{-1} \left(J_c^k \right)_i^T \underline{U}_i^{-1} \dot{\tilde{X}}_i \tag{21}$$

$$\begin{bmatrix} \dot{\Omega}_i \\ \ddot{\theta}_i \end{bmatrix} = \underline{W}_i^{-1} \left(J_c^k \right)_i^T \underline{U}_i^{-1} \left\{ \ddot{\tilde{X}}_i - \left(J_c^k \right)_i \begin{bmatrix} \Omega_i \\ \dot{\theta}_i \end{bmatrix} \right\} \tag{22}$$

$$\underline{U}_i = \begin{bmatrix} J_c^k \\ \end{bmatrix}_i \underline{W}_i^{-1} \begin{bmatrix} J_c^k \\ \end{bmatrix}_i^T \tag{23}$$

where \underline{W}_i is a (n+3) by (n+3) positive definite weighting matrix. The sub-index i , can be replaced by L for the first rover and its associated arm and R for the second rover-arm, respectively.

Through the application of the inverse dynamics of the system as illustrated in the control block diagram, the joint torques, and the rover-wheels moments then can be calculated accordingly, as shown below

The second stage of the block diagram is original, pragmatic, and based on optimal control techniques. The formulation is presented below.

A cost function C is designed to minimize the joint torques τ_{θ_L} , τ_{θ_R} , the wheel moments M_L and M_R as well as the contact forces and moments \tilde{F}_i and \tilde{N}_i applied by the arms on the common payload

Thus, the cost function, C can be described as:

$$C = \frac{1}{2} \tilde{S}^T \underline{W} \tilde{S} + \tilde{\lambda}^T \tilde{E} \tag{24}$$

\underline{W} is a square (2n+18, 2n+18) positive definite weighting matrix, n is being the total number of joints for each arm while $\tilde{\lambda}$ is the ((2n+12), 1) Lagrangian multiplier and \tilde{E} is the constraints equations vector which is defined by (26).

The \tilde{S} vector consists of the contact forces and moments as well as the wheel moments and joint torques for the two rovers and arms, respectively as shown below:

$$\text{Let } \tilde{Q}_i = \begin{bmatrix} \tilde{N}_i \\ \tilde{F}_i \end{bmatrix}, \text{ then } \tilde{S}^T = [\tilde{Q}_L, \tilde{Q}_R, M_L, M_R, \tau_{\theta_L}, \tau_{\theta_R}] \tag{25}$$

where \vec{F}_i and \vec{N}_i are the contact forces and moments applied by the arm ($i = L$ and R for the first and the second arm respectively) on the common payload, respectively.

The \vec{E} vector can be written as follows:

$$\vec{E} = \begin{bmatrix} \vec{F}_L + \vec{F}_R - m_t \ddot{x}_t \\ \vec{N}_L + \vec{N}_R - d_L \times \vec{F}_L - d_R \times \vec{F}_R - [I_t \Omega_t + \Omega_t \times I_t \Omega_t] \\ G_{WL} & G_{WLR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_L \\ G_{WLR}^T & G_{WR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_R \\ G_{W\theta L}^T & G_{W\theta L}^T & G_{\theta L} & G_{\theta LR} & \ddot{\theta}_L \\ \llbracket G_{W\theta R}^T & G_{W\theta R}^T & G_{\theta LR}^T & G_{\theta R} \rrbracket \llbracket \ddot{\theta}_R \rrbracket \end{bmatrix} - \begin{bmatrix} c_L \\ c_R \\ c_{\theta L} \\ c_{\theta R} \end{bmatrix} - \begin{bmatrix} M_L \\ M_R \\ \tau_{\theta L} \\ \tau_{\theta R} \end{bmatrix} \quad (26)$$

where m_t and \ddot{x}_t are the mass and the translational acceleration of the payload and. $d_i^T = (x_i, y_i, z_i)$ is the displacement vector measured from the arm's attachment point to the centre of mass of the payload, I_t is the inertia matrix of the payload around its centre of mass and $\dot{\Omega}_t$ is the angular rate of the payload.

Minimizing the cost function C with respect to \vec{S} , and $\vec{\lambda}_i$ can be carried out by differentiating C with respect to $\vec{\lambda}_i$ and \vec{S} . One can calculate the minimum norm of joint torques, wheel moments and the end-effectors force and moments exerted on the common payload.

$$\frac{\partial C}{\partial \vec{S}} = \vec{0} \quad (27)$$

and

$$\frac{\partial C}{\partial \vec{\lambda}} = \vec{0} \quad (28)$$

After some algebraic manipulations and eliminating $\vec{\lambda}$ from the equations, one can write the minimum norm of \vec{S} vector which consists of joint torques, wheel moments and the contact force and moment vectors as follows:

$$\vec{S} = \underline{\Delta}^{-1} \vec{P} \quad (29)$$

where $\underline{\Delta}$ is a $((2n + 18), (2n + 18))$ square matrix while n is being the number of joints. $\underline{\Delta}$ and \vec{P} are provided below:

$$\underline{\Delta} = \begin{bmatrix} W_{NL} & \vec{0} & -W_{NR} & \vec{0} & -W_{FL} (J_C^k)_L^T & W_{FR} (J_C^k)_R^T \\ (D_L - D_R)W_{NL} & W_{FL} & \vec{0} & -W_{FR} & -(1 + D_L - D_R) (J_C^k)_L^T & W_{FR} (J_C^k)_R^T \\ \vec{0} & \frac{1}{D_L} & \vec{0} & \frac{1}{D_R} & \vec{0} & \vec{0} \\ \frac{1}{D_L} & -D_L & 1 & D_R & \vec{0} & \vec{0} \\ (J_C^k)_{L1}^T & (J_C^k)_{L2}^T & \vec{0} & \vec{0} & 1 & \vec{0} \\ \vec{0} & \vec{0} & (J_C^k)_{R1}^T & (J_C^k)_{R2}^T & \vec{0} & 1 \end{bmatrix} \quad (30)$$

$$\vec{P} = \begin{bmatrix} \vec{0} \\ \vec{0} \\ \vec{0} \\ \vec{0} \\ m_t \ddot{x}_t \\ [I_t \dot{\Omega}_t + \dot{\Omega}_t \times I_t \dot{\Omega}_t] \\ G_{WL} & G_{WLR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_L \\ G_{WLR}^T & G_{WR} & G_{W\theta L} & G_{W\theta R} & \dot{\Omega}_R \\ G_{W\theta L}^T & G_{W\theta L}^T & G_{\theta L} & G_{\theta LR} & \ddot{\theta}_L \\ \llbracket G_{W\theta R}^T & G_{W\theta R}^T & G_{\theta LR}^T & G_{\theta R} \rrbracket \llbracket \ddot{\theta}_R \rrbracket \end{bmatrix} + \begin{bmatrix} c_L \\ c_R \\ c_{\theta L} \\ c_{\theta R} \end{bmatrix} \quad (31)$$

$$D_i = \begin{bmatrix} 0 & -z_i & y_i \\ z_i & 0 & -x_i \\ -y_i & x_i & 0 \end{bmatrix} \quad (32)$$

where $d_i^T = (x_i, y_i, z_i)$ is the displacement vector measured from the arm's attachment point to the centre of mass of the payload and the matrix D_i is obtained from the components of the displacement vector.

III. RESULTS AND DISCUSSION

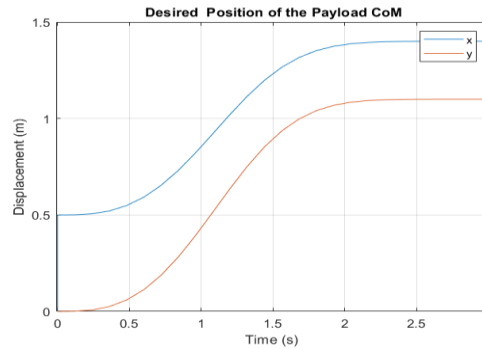
The computer simulation results are presented and discussed in this section.

In this simulation, the desired trajectory for the center of mass of the common payload was generated and in turn the trajectories for the two end-effectors were obtained while they were holding a common payload. The impedance control technique as illustrated in the first stage of the control block diagram was employed for the trajectory generation.

The main objective of this simulation was to obtain the minimum norm of the joint torques, the rover wheel moments and the contact forces and moments exerted on the payload by the two arms while tracking a desired end-effector pose.

Table 1 parameters were used for the rovers and the robotics arms in the computer simulations. Two scaled version of the 7 DOF Space Station Remote Manipulators were employed as part of the simulations.

The desired trajectories for translational as well as rotational motions of the payload are plotted with time in Fig. 3.



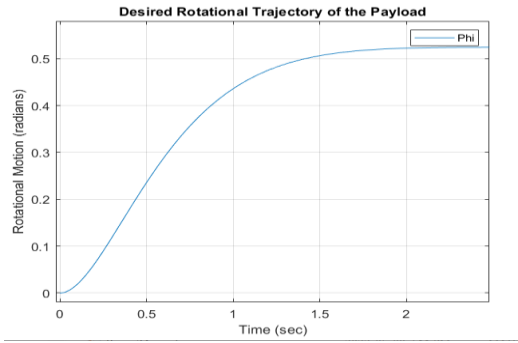


Fig. 3. Variation of the desired trajectory for payload

The minimum norm of the joint torques, and the contact forces and moments exerted on the payload by the two robot manipulators while carrying a common payload were plotted in Fig. 4a-i below. The minimum norm of joint torques (red lines) are plotted against the joint torques (blue lines) obtained by only minimizing the joint rates by the least square (i.e. only part 1 of the control block diagram is exercised). The comparison of the joint torques clearly demonstrate that the minimization scheme works efficiently.

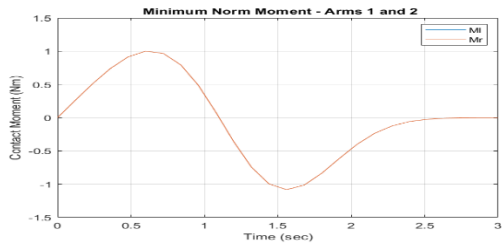
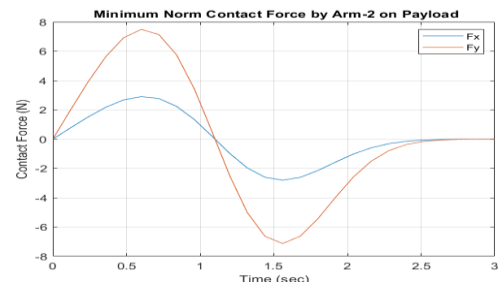
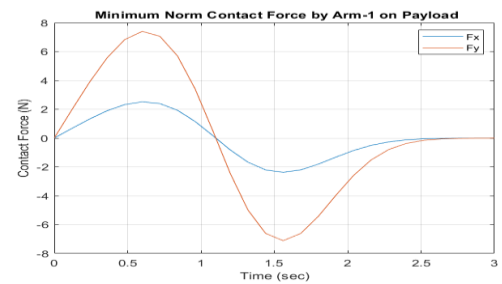
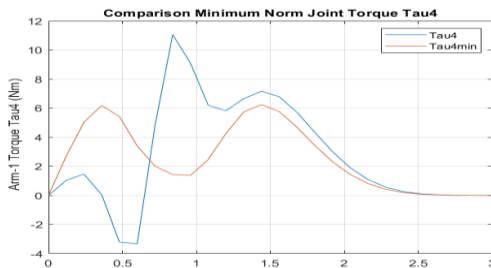
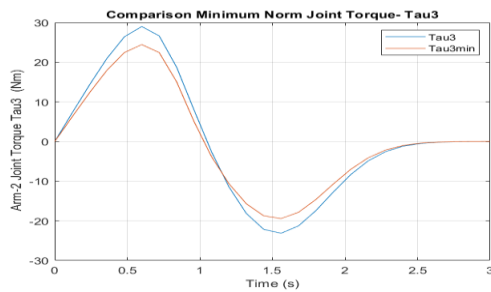
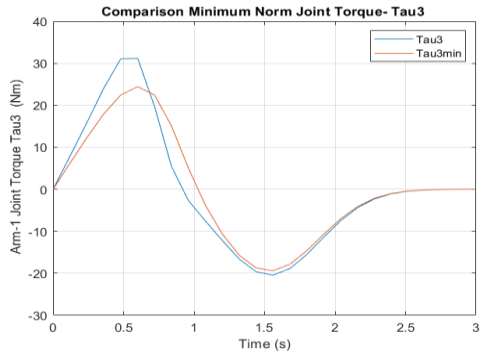
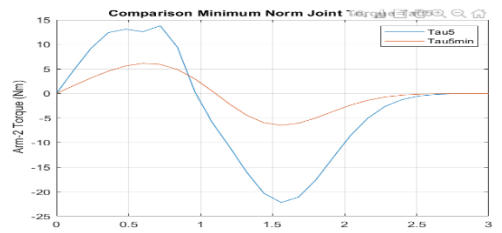
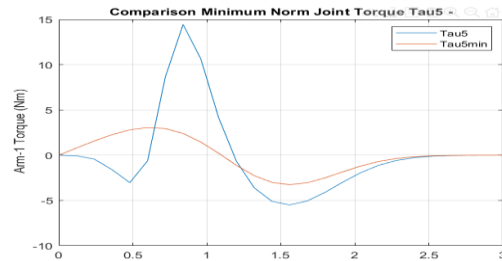
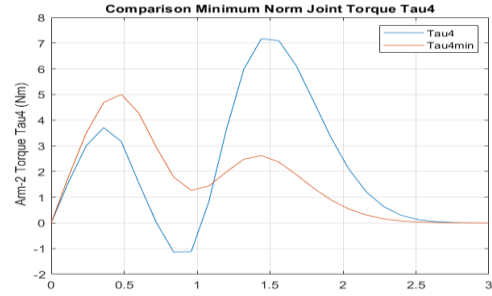


Fig. 4. Minimum norm of joint torques (first and second arm) and the contact forces and moments

The joint accelerations for each robot are also calculated using (22) and are plotted with time in Fig. 5.

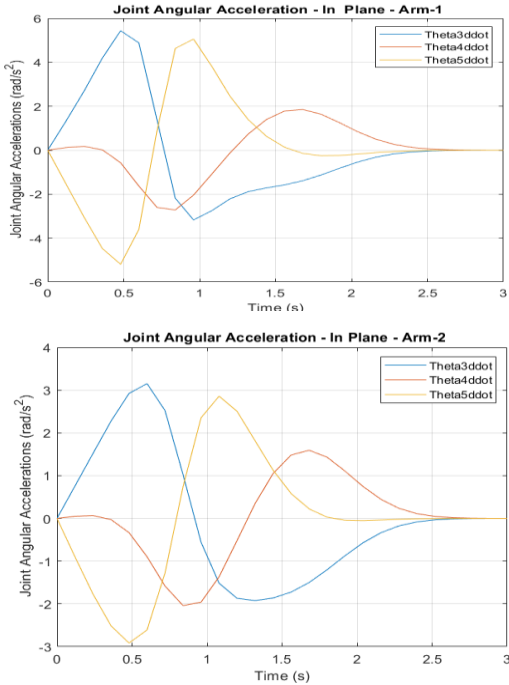


Fig. 5. Variation of joint angular accelerations for the first and second arm

The joint angular rates and angles were obtained by integrating the joint accelerations using (13) and were plotted in Fig. 6.

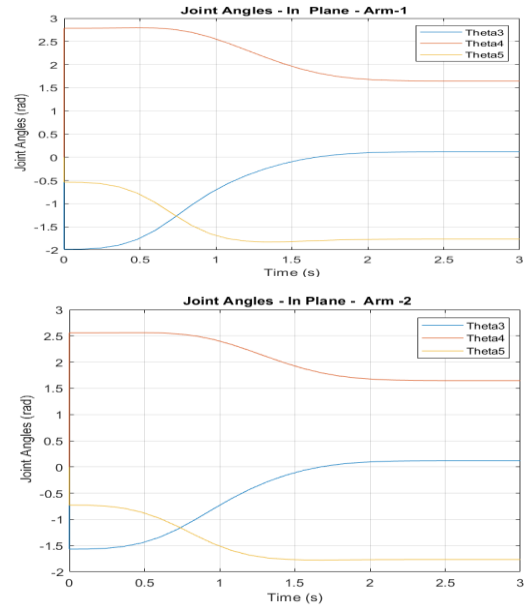
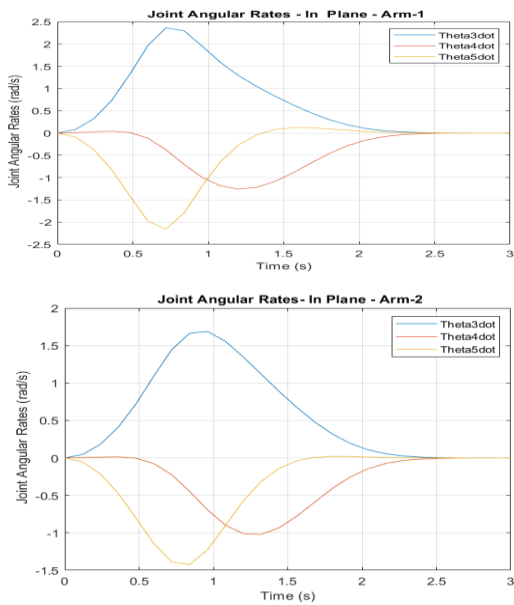
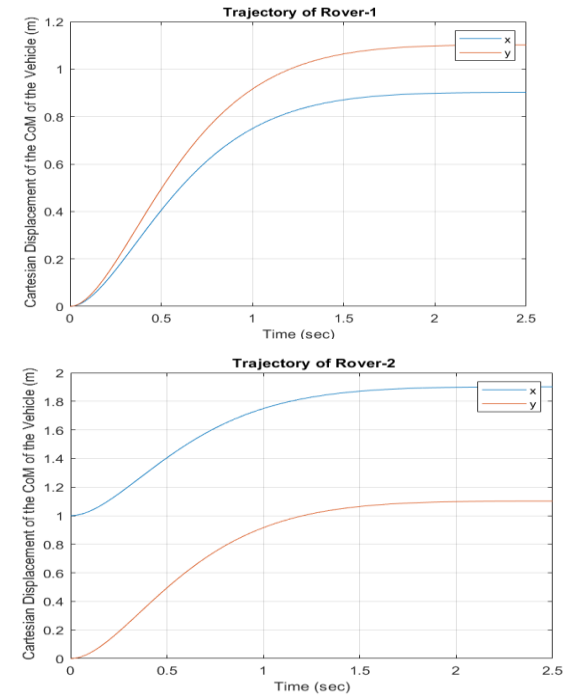


Fig. 6. Variation of joint angular rates and angles for the first and second arm

The trajectories of the centre of mass of the rovers 1 and 2 were also obtained by (22) and (8) and were plotted with time Fig. 7.



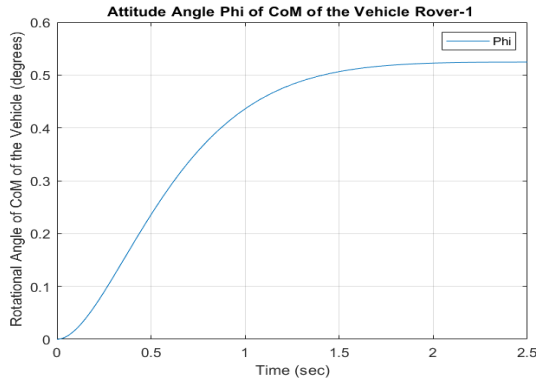


Fig. 7. Variation of Rover 1 and 2 positions and orientations with time

The angles of rotation for the rover wheels were also obtained by making use of (22) and were plotted in Fig. 8

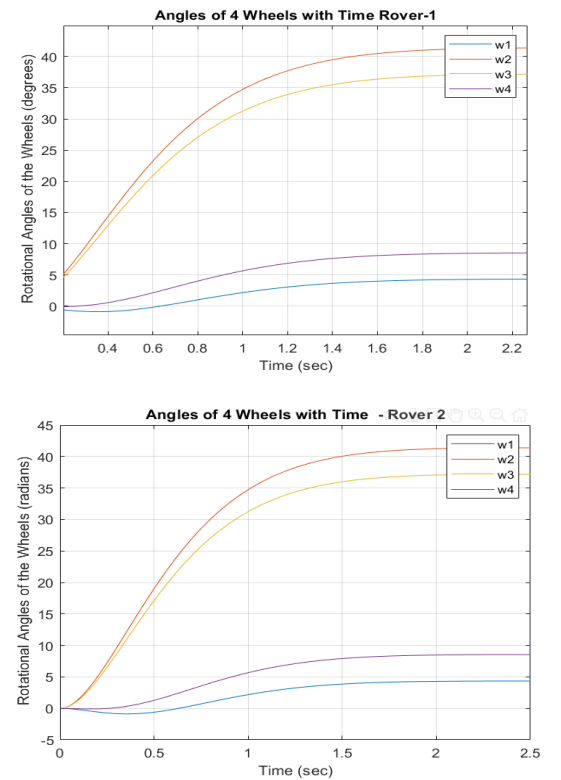


Fig. 8. Variation of angles of rotations for each wheel of rover 1 and 2

IV. CONCLUSION

This paper presented a novel optimal control algorithm for coordinated force and motion control of multi rover robotics system while manipulating a common payload. The proposed control algorithm focused on the minimization of joint torques, rover wheel moments as well as contact force / moments made with the payload. The norm of joint torques, wheel moments and the contact forces and moments were minimized to overcome the joint torque saturation problem commonly seen while

manipulating a common payload and also to provide an optimum solution for an underdetermined system with non-holonomic constraints. Furthermore, the proposed control algorithm provided an on-line trajectory generation capability while manipulating a common payload for both the rovers and the arms simultaneously.

The computer simulation results showed that the proposed control algorithm worked efficiently, and the minimum joint torques, and the contact forces and moments were obtained. Moreover, the optimal solution simultaneously satisfied the non-holonomic constraints while the end-effectors were carrying a common payload and tracking a desired payload trajectory.

REFERENCES

- [1] D. Neculescu D., B. Kim B. and S. Kalaycioglu, "Free and contact motion control for mobile robots", 8th International Conference on CAD/CAM, Robotics and Factories of the Future, Metz, France, Aug. 1992.
- [2] D. Neculescu D., B. Kim and S. Kalaycioglu, "Contact motion control for mobile robots ", 7th IFAC Symposium on Information Control Problems, Toronto, May 1992.
- [3] R. Fierro and F. Lewis, 1995 Control of a nonholonomic mobile robot: backstepping kinematics into dynamics. Decision and Control, Proceedings of the 34th IEEE Conference on. Vol. 4. .
- [4] T. Yu, N. Sidek and N. Sarkar, 2009 Modeling and control of a nonholonomic wheeled mobile robot with wheel slip dynamics. Computational Intelligence in Control and Automation, IEEE.
- [5] Y. H Amengonu and Y. P Kakad 2014 IOP Conf. Ser.: Mater. Sci. Eng. 65 012017
- [6] G. Campion , B. d'Andrea-Novel and G. Bastin, 1991 Controllability and state feedback stabilization of non holonomic mechanical systems. Lecture Notes in Control and Information Science (Berlin:Springer-Verlag) p 106-24
- [7] A. Bloch and N. McClamroch, 1989 Control of mechanical systems with classical nonholonomic constraints Proc. of 28th IEEE Conference on Decision and Control (Florida) p 201-05
- [8] S. Kalaycioglu, Control of multiple robot manipulators with optimal force distribution, Canadian Conference on Electrical and Computer Engineering, IEEE, September 1991.
- [9] F. G. Pin,S. M. Killough, A new family of omnidirectional and holonomic wheeled platforms for mobile robots, IEEE Trans. Robot. Autom.1994, 10, 480.
- [10] G. Campion,G. Basin,B. D'Andrea-Novel,Structural properties and classification of kinematic and dynamic models of wheeled mobile robots, IEEE Trans. Robot. Autom. 1996.
- [11] M. Wada,S. Mori,Proceedings of the IEEE International Conference on Robotics and Automation, Minneapolis, Minnesota, April, 1996, pp. 3671-3676.
- [12] J. Ostrowski, J. Burdick, The geometric mechanics of undulatory robotic locomotion, The International Journal of Robotic Research 1998, 17, 683.
- [13] C. Stoeger,A. Mueller, H. Gattringer,Parameter identification and model-based control of redundantly actuated, non-holonomic, omnidirectional vehicles, Informatics in Control, Automation and Robotics. Lecture Notes in Electrical Engineering 2018, 430, 207.
- [14] P. F. Muir,C. P. Neumann, Kinematic modeling of wheeled mobile robots, J. Robot. Syst.1987,4,281.
- [15] G. Wampfler,M. Salecker,J. Wittenburg, Kinematics, dynamics, and control of omnidirectional vehicles with mecanum wheels, Mechanics Based Design of Structures and Machines 1989.
- [16] A. Gfrerrer, "Geometry and kinematics of the Mecanum wheel," Computer Aided Geometric Design, vol. 25, no. 9, pp. 784-791, 2008.
- [17] L. Lin C and Y. Shih H, "Modeling and adaptive control of an omnimecanum-wheeled robot," Intelligent Control & Automation, vol. 2013, no. 2, pp. 166-179, 2013.

- [18] A. Shimada, S. Yajima, P. Viboonchaicheep, and K. Samura, Mecanum-wheel vehicle systems based on position corrective control, in Proceedings of the IECON 2005: 31st Annual Conference of IEEE Industrial Electronics Society, pp. 2077–2082, November 2005.
- [19] Y. Wang and D. Chang, Motion performance analysis and layout selection for motion system with four Mecanum wheels, Journal of Mechanical Engineering, vol. 45, no. 5, pp. 307–316, 2009.
- [20] M. O. Tatar, C. Popovici, D. Mandru, I. Ardelean, and A. Plesa, Design and development of an autonomous omni-directional mobile robot with Mecanum wheels, in Proceedings of the 2014 19th IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR 2014.

TABLE I. THE SYSTEM PARAMETERS USED IN THE SIMULATION

Hardware Configuration Item	Mass (kg)	Dimensions (m) (prism)
Rovers-1 and 2	40	0.5 x 0.5 x 0.3
Common Payload	10	0.4 x 1 x 0.4
Joint / Link 1	1	0.1 x 0.1 x 0.1
Joint / Link 2	1	0.1 x 0.1 x 0.1
Link 3	3	1 x 0.1 x 0.1
Link 4	5	1 x 0.1 x 0.1
Joint / Link 5	3	0.1 x 0.1 x 0.1
Joint / Link 6	1	0.1 x 0.1 x 0.1
Joint / Link 7	3	1 x 0.1 x 0.1

Analysis of the Influence of Factors on Flight Delays in the United States Using the Construction of a Mathematical Model and Regression Analysis

Timofey Kireev
Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
sofyachash@mail.ru

Vladislav Kukartsev
¹Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
²Department of Informatics
Siberian Federal University
³Digital Material Science: New
Materials and Technologie
Bauman Moscow State Technical
University
Krasnoyarsk, Russia
0000-0001-6382-1736

Alesya Pilipenko
Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
pilipenko.alesya@mail.ru

Anastasiya Rukosueva
Department of System Analysis
Reshetnev Siberian State University of Science and Technology
Krasnoyarsk, Russia
soboleanastasiya@mail.ru

Viktor Suetin
Department of System Analysis
Reshetnev Siberian State University of Science and Technology
Krasnoyarsk, Russia
suetin@sibsau.ru

Abstract—The paper compiled and structured data to construct a correlation model. The reasons for flight delays and cancellations between 2009 from 2019 are examined, based on the statistical electronic source “Bureau of Transportation Statistics”. The normalized data are calculated from the primary data. Based on the data collected, a primary histogram was constructed (outliers were not checked). In addition, significant and weak influencing factors have been identified for the constructing of a correlation table, this table helps to identify an additional factor influencing a flight delay or cancellation. Based on the data from the table, a new histogram was built and a correlation table was constructed, allowing regression analysis to be started. At the beginning of the regression analysis, regression statistics are used and an analysis of variance is performed, then the regression analysis is performed directly and the results of the regression are summed up. Based on the summed-up results, a histogram was built, which takes into account all the statistics and calculations, showing the predicted number of delays or cancellations of flights and their actual number.

Keywords—Regression analysis, mathematical model, climate, temperature sensation, temperature factors

I. INTRODUCTION

The solution of practical problems by mathematical methods is consistently carried out by formulating the problem (development of a mathematical model), choosing a method for studying the obtained mathematical model, analyzing the obtained mathematical result. The mathematical formulation of the problem is usually presented in the form of geometric images, functions, systems of equations, etc. The description of an object (phenomenon) can be represented using continuous or discrete, deterministic or stochastic and other mathematical forms. The theory of mathematical modeling ensures the identification of the patterns of the flow of various

phenomena of the surrounding world or the operation of systems and devices by their mathematical description and modeling without field tests [1].

Mathematical analysis is widely used in physics, computer science, statistics, engineering, economics, business, finance, medicine, demography and other areas where a mathematical model can be built to solve a problem and its optimal solution needs to be found.

Regression analysis can be used to estimate the degree of relationship between variables and to model future dependencies. In fact, regression methods show how changes in the “independent variables” can be used to fix the change in the “dependent variable”. This model can be used to detect trends and make forecasts. Suppose the company's sales have been growing for two years. By doing a linear analysis of the monthly sales data, the company could predict sales in the coming months. It is with the help of regression analysis that a mathematical model will be formulated showing the ratio of the predicted number of cancellations and delays of flights to the actual ones [2].

Air transport is gaining more and more popularity, more and more people prefer to use air transport to move from one locality to another, as it significantly reduces the travel time and allows you to get to the right place in the shortest possible time. Air transportation is becoming one of the most popular passenger transportation services at the present time, due to the speed, comfort, safety and availability of this type of transportation. This service reaches its maximum values in the field of international mobility, the purpose of which in most cases is tourism. Passenger air transportation is a service for the transportation of the population on an aircraft, provided on various conditions. In the Russian Federation, air transportation services for the population are among the types of activities carried out by civil aviation (in addition to

passenger air transportation, civil aviation carries out activities of a general social nature in the field of medicine, sports and culture, agriculture, etc.). Air transportation of the population is divided into: domestic and international; regular and charter; commercial and non-commercial. Among other things, air transport also has a significant important role in business (business trips, business meetings, etc.). But, despite all the comfort and accessibility, air transport also has a large number of nuances, which are mainly manifested in the loss of time, waiting for a plane that is delayed for one reason or another. It is also not uncommon for flights to be canceled, for example, due to weather conditions or other factors that are almost impossible to fully predict. We will stop on the cancellation factors. We will determine the possible reasons for the cancellation or delay of flights. Of course, one of the main reasons is the weather. Unfortunately, even with modern forecasting capabilities for positive weather conditions, not everything always goes according to plan, and in some cases it is necessary to delay or even cancel a flight due to the threat of a crash, respectively, a threat to the lives of passengers and crew members. Much less often, but still, the reason for canceling a flight can be a transport security check. The reason for the intervention of TB can be a complaint from crew members or directly from the passengers themselves about a suspicious person or object. In this case, the TB can land the plane for a forced check of the passenger, and the flight is delayed accordingly. Extremely rarely, but one of the reasons for the intervention of TB can be a terrorist threat or its prerequisites. Safety intervention is also necessary if there is a violent passenger in the cabin, disturbing the peace of the others or interfering with the work of the crew. A more frequent factor is the delay or cancellation of the flight by the national aviation system, the reason for the intervention may be the unstable political situation in the country that is part of the journey. In this case, a change of route or a flight cancellation may occur. Sometimes the flight can be canceled by the airline itself, the reasons for

this can be very different, often the company does not report the reason for canceling the flight. At present, all these factors have long been known and quite successfully predicted. Next, we will conduct a regression analysis and check which factors stand out from the rest and compare the actual number of events with the predicted ones [3], [4].

II. DATA COLLECTION AND SORTING

Using data taken from the Bureau of Transportation Statistics and other sources, we will compile a table from 2009-2019, in which the main indicators will be the number of delays and the factors affecting them [5]–[8]:

- The number of delays due to bad weather.
- Suppose for 2009 the number of delays was 41865, and for 2019 it is already 46303.
- The number of delays due to security checks.
- Delays for 2009 are 2328, and already for 2019 there were 2642.
- The number of delays due to the national aviation system.
- The number of delays in 2009 was 450,645, but in 2019 the number increased to 439,491.
- The number of delays due to the airline itself.
- Delays in 2009 were 323,743, and already in 2019 they were 387,853
- The number of delays due to other reasons.
- The number of delays in 2009 was 399709, but already in 2019 these delays increased to 512965.

Having collected the necessary data, we enter them into Excel and create a histogram (Figure 1).

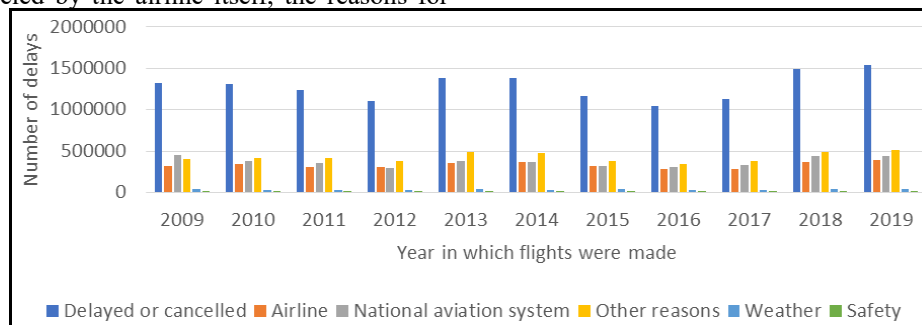


Fig. 1. Flight delays for one reason or another to the total number of delayed or canceled flights.

Having made a histogram, we carry out conditional formatting to exclude possible outliers. “Conditional formatting marks indicators with different colors, and if there is an indicator in a particular factor, the color of which is radically different from the rest, this is an outlier [9]. Outliers are values that are very different from the rest, and they are not desirable for use in constructing a regression equation, since the equation will be unstable in use, that is, when new data is added, the old indicators change greatly, which will lead to the function becoming completely different [10].

In this case, no such outliers were found, so it is possible to carry out a correlation analysis, as a result of which we

will get a table where you can see the factors that do not or weakly affect the number of delays.

We determine non-influencing factors according to the following principle, if the factor is close to 1, then it has an influencing value on the main factor, respectively, if it is far from 1, then it is not influencing.

III. BUILDING A CORRELATION TABLE OF LINKS AND WORKING WITH DATA

The correlation coefficient r_{jk} characterizes the relationship between two features x_j and x_k in the case of a linear correlation between them. For any signs and random variables, it is calculated by the formula (1):

$$r_{jk} = \frac{1}{N} \sum_{i=1}^N y_{ij}y_{ik} \quad (1)$$

where y_{ij} and y_{ik} are the normalized features x_j and x_k for the i -th dimension (object). As a result, we obtain the matrix of correlation coefficients R [11].

The matrix is calculated using normalized factors in Excel or Statistica. Also, we cannot use the initial data, so we calculate the normalized factors and use them to build a correlation table [12], [13].

Factors (normalized):

- Number of delays due to Bad weather (0.901).
- Number of delays due to Security Review (0.548).
- Number of delays due to the National Aviation System (0.968).

- Number of delays due to the Airline (0.97).
- Number of delays due to other reasons (0.972).

In our case, the weakly influencing factor was: Security check.

And also, the most influential:

- National aviation system.
- Airline.
- Other reasons.
- Bad weather.

The result of working with data can be seen in the table (Table I)

TABLE I. CORRELATION TABLE

	Y	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Y	1	-	-	-	-	-
Factor 1	0.970	1	-	-	-	-
Factor 2	0.968	0.893	1	-	-	-
Factor 3	0.972	0.947	0.912	1	-	-
Factor 4	0.901	0.898	0.909	0.822	1	-
Factor 5	0.548	0.482	0.636	0.417	0.564	1

Looking at the factors, we can calculate one of the unnecessary ones. This is a “Security Check”, this factor has little effect on delays and does not occur very often.

factors that affect the number of delays, we also perform conditional formatting to exclude outliers, in this situation no outliers were found [14].

After calculating unnecessary factors, we discard them and create a new more accurate table with the remaining

Detailed data see in the histogram (Figure 2).

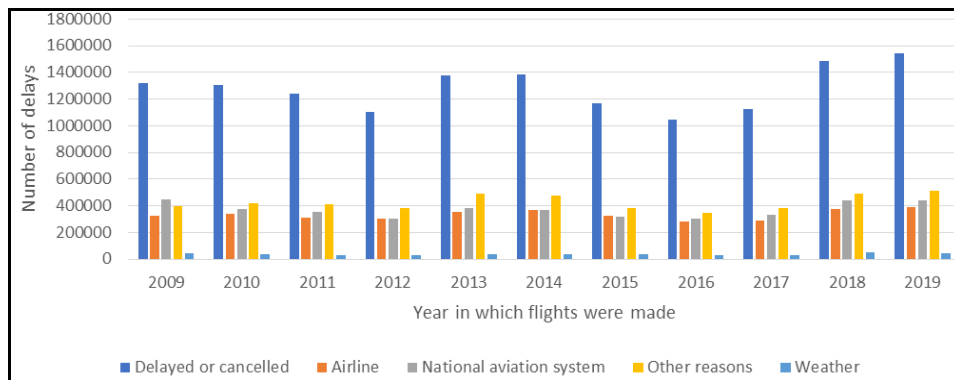


Fig. 2. Number of delays to total delays.

IV. REGRESSION DATA ANALYSIS

After we are convinced of the available data, we can proceed to regression analysis, as a result of which we obtain the following data.

The R-square describes how accurately the constructed model describes the reality of this function, the maximum value of the R-square is 1, that is, 100% correctly describes

reality, the model is considered good if this coefficient is higher than 0.8 [15]; in our case, the R-square is 0.996, which means the quality of the model is real; F-significance is how adequate the resulting equations are, the lower this value, the better [16]. Our value is 0.000004753.

See tables for details (Tables II-IV).

TABLE II. REGRESSION STATISTICS

Multiple R	0.99810816
R-square	0.9962199
Normalized R-square	0.9924398
Standard error	13987.8396

Observations	11.00
--------------	-------

TABLE III. ANALYSIS OF VARIANCE

	df	SS	MS	F	Significance (F)
Regression	5	257823908619.004	51564781723.801	263.543	0.00000475330752
remainder	5	978298283.905	195659656.781	-	-
Total	10	258802206902.909	-	-	-

TABLE IV. REGRESSION DATA MODEL

	coefficient	standard error	t-statistic	P-meaning	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Y-intersection	-109118.7	65080.44	-1.68	0.15	-276413.31	58175.90	-276413.31	58175.90
X1	1.78	0.58	3.07	0.03	0.29	3.27	0.29	3.27
X2	1.14	0.22	5.2	0	0.57	1.7	0.57	1.7
X3	0.91	0.3	3.06	0.03	0.15	1.67	0.15	1.67
X4	-1	2.22	-0.45	0.67	-6.71	4.71	-6.71	4.71
X5	12.82	13.3	0.96	0.38	-21.36	47.01	-21.36	47.01

Further, it will be useful for us to know what is the residual when predicting Y. The residual is the difference between the predicted Y and the real one. For example, in the 1st observation (2009), the predicted Y is 1.33 million, and the real one is 1.323 million, the remainder becomes -

7180, that is, this is the difference between the calculated Y and the real one. The residual reflects how accurately this model calculates delays, the smaller the residual, the more accurately the assembled model considers [17], [18]. Detailed data look at (Figure 4).

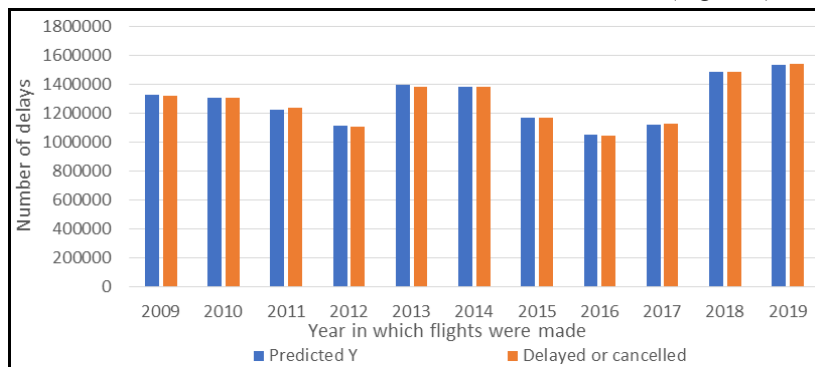


Fig. 3. Difference between predicted Y and total delays.

CONCLUSION

Using the method of regression analysis, a mathematical model was obtained, based on which, the authors from a large amount of data (Big Data) identified those that strongly affect the number of delays in the United States. Therefore, after receiving data from a mathematical model, the most influential factor can be determined. This factor includes many factors that are insignificant in their frequency and influence, which did not make sense to consider separately, since only in their totality they represent a significant unit in our study. It is also worth mentioning such factors as the delays of the National Aviation System and the delays of the Airline itself, they quite strongly influence the number of cancellations or flight delays, as well as other reasons for the delay. Other factors have less impact on flights than safety checks are not very frequent and therefore do not have a significant impact on overall statistics. Bad weather is also a fairly strong influence factor, but quite easily predictable, so it does not have much effect on statistics comparing the actual number of cancellations to the predicted one. This work demonstrates the significance of

regression analysis and its advantages. So far, regression analysis is one of the most popular and accurate. The precision of the regression equation is basis of the regression analysis. While all models will have some errors, but understanding these statistics will help determine if this model can be used for analysis purposes or if additional transformations are required [19]. The expected values are calculated based on the regression equation and the values for each independent variable. Ideally, the expected values should correspond to the observed (predicted) values [20]. In our case, these values are extremely close, which is very positive, since the predicted number of flight delays or cancellations almost completely corresponds to the real one, which means that airlines are ready to predict their financial losses in advance and minimize them, as well as create conditions for comfortable waiting for passengers.

REFERENCES

[1] "Bureau of Transportation Statistics," 2021. https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp?6B2r=F&20=E (accessed Apr. 30, 2022).

- [2] I. Alomar, M. Belitskaya, and A. Belitskaya, "Comparative Statistical Analysis of Airport Flight Delays for the Period 2019–2020. Almaty International Airport Case Study," *Lecture Notes in Networks and Systems*, vol. 410 LNNS, pp. 110–124, 2022, doi: 10.1007/978-3-030-96196-1_11.
- [3] N. Chakrabarty, "A data mining approach to flight arrival delay prediction for American airlines," in *IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference*, Mar. 2019, pp. 102–107. doi: 10.1109/IEMECONX.2019.8876970.
- [4] R. Nigam and K. Govinda, "Cloud based flight delay prediction using logistic regression," in *Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017*, Jun. 2018, pp. 662–667. doi: 10.1109/ISS1.2017.8389254.
- [5] N. M. Agarkov, A. M. Chukhraev, D. A. Konyaev, and E. V. Popova, "Diagnostics i prognozirovanie pervichnoi otkrytougol'noi glaukomy po urovnyu mestnykh tsitokinov," *Vestn Oftalmol*, vol. 136, no. 4, pp. 94–98, 2020, doi: 10.17116/oftalma202013604194.
- [6] V. A. Kukartsev, V. v. Kukartsev, and A. V. Kukartsev, "Effect of the Temperature Treatment of Quartzite on the Lining Resistance of Commercial-Frequency Induction Crucible Furnaces," *Refractories and Industrial Ceramics*, vol. 59, no. 3, pp. 252–256, Sep. 2018, doi: 10.1007/s11148-018-0216-2.
- [7] P. Monmousseau, D. Delahaye, A. Marzuoli, and E. Feron, "Predicting and analyzing US air traffic delays using passenger-centric data-sources," 2019.
- [8] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.
- [9] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, "Intellectual support system of administrative decisions in the big distributed geoinformation systems," *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.
- [10] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, S. G. Dokshanin, and V. V. Kukartsev, "Research of methods for design of regression models of oil and gas refinery technological units," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042078, Jun. 2019, doi: 10.1088/1757-899X/537/4/042078.
- [11] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, Y. V. Danilchenko, S. N. Ezhemanskaya, and N. V. Sokolovskiy, "Methodology for the formation of indicators balanced system for marketing activities of an industrial enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012084, Jan. 2020, doi: 10.1088/1757-899X/734/1/012084.
- [12] V. S. Tynchenko et al., "Software to Predict the Process Parameters of Electron Beam Welding," *IEEE Access*, vol. 9, pp. 92483–92499, 2021, doi: 10.1109/ACCESS.2021.3092221.
- [13] S. Berger, A. Kilchenmann, O. Lenz, and F. Schlöder, "Willingness-to-pay for carbon dioxide offsets: Field evidence on revealed preferences in the aviation industry," *Global Environmental Change*, vol. 73, p. 102470, Mar. 2022, doi: 10.1016/j.gloenvcha.2022.102470.
- [14] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 2, p. 022106, Aug. 2019, doi: 10.1088/1755-1315/315/2/022106.
- [15] H. Khaksar and A. Sheikholeslami, "Airline delay prediction by machine learning algorithms," *Scientia Iranica*, vol. 26, no. 5 A, pp. 2689–2702, 2019, doi: 10.24200/sci.2017.20020.
- [16] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.
- [17] A. A. Boyko, V. v. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, E. A. Chzhan, and A. S. Mikhalev, "Dynamic simulation of calculating the purchase of equipment on credit," *Journal of Physics: Conference Series*, vol. 1333, no. 3, p. 032009, Nov. 2019, doi: 10.1088/1742-6596/1333/3/032009.
- [18] S. Wang, "Area air traffic flow optimal scheduling under uncertain weather," in *Proceedings - 2009 International Conference on Computational Intelligence and Software Engineering, CiSE 2009*, 2009, p. 5362839. doi: 10.1109/CISE.2009.5362839.
- [19] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," *Journal of Physics: Conference Series*, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.
- [20] V. S. Tynchenko, A. V. Milov, V. V. Tynchenko, V. V. Bukhtoyarov, and V. V. Kukartsev, "Intellectualizing the process of waveguide tracks induction soldering for spacecrafts," *International Review of Aerospace Engineering*, vol. 12, no. 6, pp. 280–289, 2019, doi: 10.15866/irease.v12i6.16910.

Construction of a Factor Model Focused on the Reduction of Difficulties in Road Traffic

Natalia Fedorova

¹Department of management
Reshetnev Siberian State University of
Science and Technology

²Department of Advertising and Social
and Cultural Activities
Siberian Federal University
Krasnoyarsk, Russia
nvfed@mail.ru

Aleksander Myrugin

Information Control Systems
Department
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
0000-0003-2887-6162

Elena Filushina

Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
marbury@yandex.ru

Yuriy Seregin

Information Control Systems
Department
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
0000-0003-4309-8637

Elena Vaitekunene

Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
²Siberian Federal University
³Department of Civil Defence and
Emergency Management
Siberian Fire and Rescue Academy of
the Russian Ministry of Emergency
Situations
Krasnoyarsk, Russia
0000-0001-6839-6716

Yuriy Danilchenko

Department of management
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
ydanilchenko@sibsau.ru

Abstract—The article collected and structured data for regression analysis. The reasons for which cork is formed in the city of Krasnoyarsk were considered. Using the collected data, a primary histogram (not tested for emissions) was constructed. The following factors were identified as relevant and weak-influencing factors for the construction of the correlation table, this table helps to identify the superfluous influence on the formation of traffic jams. Taking into account the table, a new histogram was constructed and a correlation table was created, which allows to start regression analysis. At the beginning of the regression analysis regression statistics are used and the variance analysis is carried out, then the regression analysis is made directly and the regression results are summarized. Based on the summing up of the histogram, which takes into account all statistics and calculations, which reflects the predicted number of traffic jams and the real factors influencing the formation of traffic jams.

Keywords— *Regression analysis, mathematical model, climate, temperature factors, traffic jams*

I. INTRODUCTION

Mathematical models as tools allowing to study the complex processes of the real world, including transport infrastructure, without capital costs, are a popular tool for solving many problems in various spheres of the national economy. The development of information technology and computing tools allowed to expand the scope of tasks solved with the help of models [1], [2].

The rapid growth of the urban car fleet has stretched the capacity of the street and road networks. Therefore, the issue of assessing the efficiency of traffic management in a high-load environment has become more acute, especially for large and large cities [3]. When looking for the best strategies for managing In the search for better strategies for traffic management, in making optimal decisions in the design of new transport infrastructure, as well as in the

choice of rational traffic management, it is necessary to make the most effective transport decisions [4].

Regression analysis allows solving two problems. The first problem is based on the choice of independent variables that significantly influence the dependent quantity, and the definition of the form of the regression equation. This problem is solved by analyzing the studied relationship. Formal tools can help here only by some benchmarks. The second task is to evaluate the parameters. It is solved by different statistical methods of data processing and observation. Most often, regression parameters are estimated using the least squares method. The regression analysis uses a random variable as a function, with arguments based on non-random variables [5]. The field of application of regression analysis in the economy is the study of impacts on labor productivity and cost of various factors. Before building a regression model, the topics to be considered should be selected and quantified [6].

The choice of the type of regression dependency is based on the following position: the need to correlate the selected dependency with professional-logical assumptions about the nature and nature of the relationships under study. Regression dependencies often use simple dependencies that do not require complex calculations that can be easily interpreted. The practical application of regression analysis is based on the fact that the linear regression equation clearly expresses the relationship between indicators even when they are more complex and require significant calculations [7]-[9].

One of the key problems of any large city is the increase in the number of cars. For example, in Krasnoyarsk over the past 10 years, there have been almost twice as many cars. However, mass motorization has created traffic jams, which have long become a big problem primarily large cities all

over the world [10], [11]. Megacities work like cork factories. Every day and everywhere, traffic jams lead to huge economic, time, psychological, environmental and other costs that fall on the shoulders of drivers of cars and other vehicles, transport companies, city and national economies, as well as pedestrians and city residents [12]. For example, a significant part of the traffic is made up of various business entities, delivery services, delivery of retail goods. Timely delivery of goods is strategically important, especially for large shopping centers and hypermarkets. This ensures the rhythm of the trading process. Modern cities and megacities need a constant increase in the volume of transport communication, increasing its reliability, safety and quality. This requires an increase in the cost of improving the infrastructure of the transport network, turning it into a flexible, highly manageable logistics system. At the same time, the risk of investments increases significantly if the patterns of development of the transport network and the distribution of the load on its sections are not taken into account. Ignoring these patterns leads to frequent traffic jams, significant overload of individual network nodes, an increase in the accident rate, and environmental damage. To search for effective strategies for managing traffic flows, optimal solutions for designing a road network and organizing traffic, it is necessary to take into account a wide range of traffic flow characteristics, patterns of influence of

external and internal factors on the dynamic characteristics of a mixed traffic flow [13], [14]. The task of any city is to provide the ability to quickly move along the highways of the city, but this is not always possible, since many different factors affect the congestion of the city at a certain moment.

For example, the day of the week or the number of accidents on a certain day, time of day, various road works, broken traffic lights, etc. have a significant impact on city traffic. Using regression analysis, we identified the main factors that affect traffic congestion in the city of Krasnoyarsk and conducted an analysis with which, in the future, it is possible to analyze and improve the road capacity in the city of Krasnoyarsk [15].

II. SELECTION OF FACTORS FOR THE REGRESSION MODEL

We get input factors that affect the number of traffic jams in Krasnoyarsk. The first histogram (Figure 1) shows the factors morning, afternoon, evening, the number of cars, thousand. The second histogram (Figure 2) shows the evening factors, the number of cars in thousand.

The third histogram (Figure 3) shows the factors the number of accidents, faulty traffic lights, the price of gasoline per liter.

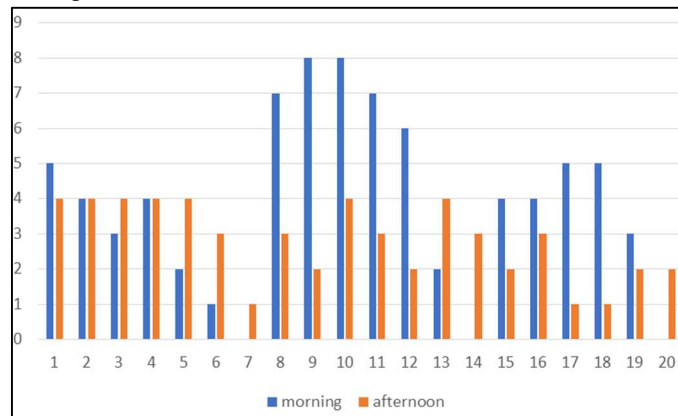


Fig. 1. Factors are shown: morning, afternoon.

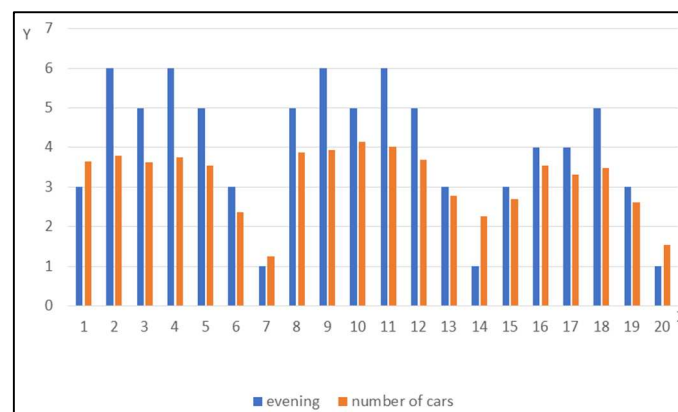


Fig. 2. Factors are shown: evening, number of cars in thousand.

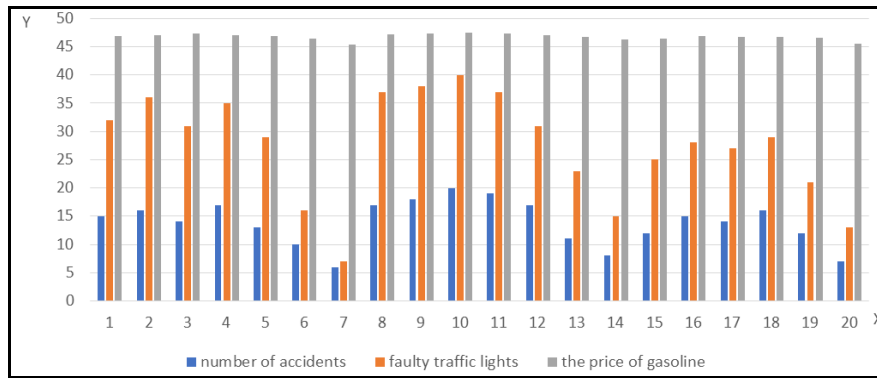


Fig. 3. Factors are shown: number of accidents, faulty traffic lights, price of gasoline per liter.

As a result, we got 7 factors that may affect traffic jams in the city. The main indicator for us is the number of traffic jams.

The next indicator that possibly affects the population is the time of day, morning, afternoon and evening. The

following are indicators of the number of cars, thousand, the number of faulty traffic lights, the price of gasoline per liter.

III. BUILDING CORRELATION TABLES

In Table I shows the correlation made on the output of the collected statistics.

TABLE I. CORRELATION TABLE OF VARIOUS FACTORS

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1							
X1	0.902	1						
X2	0.407	0.048	1					
X3	0.920	0.758	0.310	1				
X4	0.967	0.847	0.429	0.906	1			
X5	0.978	0.932	0.282	0.9	0.958	1		
X6	0.987	0.883	0.429	0.904	0.974	0.965	1	
X7	0.948	0.804	0.495	0.879	0.969	0.921	0.94	1

In the correlation table (Table I), all values and ranges are from -1 to 1, and the closer the coefficient is to 1 or (-1), the stronger the relationship between the factors. In the resulting table, you can clearly see that each factor affects traffic jams in different ways, some more than some less. The most influential factor is faulty traffic lights (X6), since its indicator of connection with traffic jams is as close as possible to one (0.987) the number of cars thousand. The next most important factor is the number of accidents (X5) (0.978) the next most important factor is the number of cars thousand. (X4) (0.967) the next most important factor is the price of gasoline per 1 liter (X7) (0.930) the next most important factor is evening (X3) (0.920) the next most

important factor is morning (X1) (0.902) the next most important factor is day (X2) (0.407).

Also, if the factors have too weak a connection with the key factor (traffic jams), they must also be excluded, since they will not have any effect in the constructed regression equation, this is a day. In the end, we are left with 6 factors, from which we collect a new table with factors that affect traffic jams for regression analysis.

The second table (Table II) shows statistics with factors after correlation. The third table (Table III) shows the correlation based on the new statistics.

TABLE II. STATISTICS AFTER CORRELATION

Traffic jams	Morning	Evening	Number of cars	Number of accidents	Faulty traffic lights	The price of gasoline per liter
12	5	3	3.648	15	32	46.9
14	4	6	3.784	16	36	47.1
12	3	5	3.62	14	31	47.3
14	4	6	3.754	17	35	47.1
11	2	5	3.543	13	29	46.9
7	1	3	2.356	10	16	46.4
2	0	1	1.254	6	7	45.4
15	7	5	3.876	17	37	47.2

16	8	6	3.947	18	38	47.3
17	8	5	4.143	20	40	47.5
16	7	6	4.023	19	37	47.3
13	6	5	3.698	17	31	47.1
9	2	3	2.785	11	23	46.7
4	0	1	2.267	8	15	46.3
9	4	3	2.696	12	25	46.5
11	4	4	3.542	15	28	46.9
10	5	4	3.312	14	27	46.7
11	5	5	3.475	16	29	46.8
8	3	3	2.621	12	21	46.6
3	0	1	1.543	7	13	45.5

TABLE III. CORRELATION TABLE OF THE RELATIONSHIP OF VARIOUS FACTORS

	Y	X1	X2	X3	X4	X5	X6
Y	1						
X1	0.902	1					
X2	0.92	0.758	1				
X3	0.967	0.847	0.906	1			
X4	0.978	0.932	0.9	0.958	1		
X5	0.987	0.883	0.904	0.974	0.965	1	
X6	0.948	0.804	0.879	0.967	0.921	0.94	1

The most influential factor is the number of faulty traffic lights (X5) as its relationship with traffic jams is as close as possible to one (0.987) the next most important factor is the number of accidents (X4) (0.978) the next most important factor is the number of cars thousand (X3) (0.967) the next most important factor is the price of gasoline per 1 liter (X6) (0.930) the next most important factor is the evening (X2) (0.920) the next most important factor is the morning (X1) (0.902).

Also, if the factors have too strong a relationship with each other, they must also be excluded, since they will not have any effect in the constructed regression equation, the number of accidents. In the end, we are left with 5 factors, from which we collect a new table with factors that affect traffic jams for regression analysis [16], [17].

The fourth table (Table IV) shows the final factors after the final correlation.

TABLE IV. FINAL TABLE BY FACTORS AFTER THE FINAL CORRELATION

Traffic jams	Morning	Evening	Number of cars	Faulty traffic lights	The price of gasoline per liter
12	5	3	3.648	32	46.9
14	4	6	3.784	36	47.1
12	3	5	3.62	31	47.3
14	4	6	3.754	35	47.1
11	2	5	3.543	29	46.9
7	1	3	2.356	16	46.4
2	0	1	1.254	7	45.4
15	7	5	3.876	37	47.2
16	8	6	3.947	38	47.3
17	8	5	4.143	40	47.5
16	7	6	4.023	37	47.3
13	6	5	3.698	31	47.1
9	2	3	2.785	23	46.7
4	0	1	2.267	15	46.3
9	4	3	2.696	25	46.5
11	4	4	3.542	28	46.9
10	5	4	3.312	27	46.7
11	5	5	3.475	29	46.8
8	3	3	2.621	21	46.6
3	0	1	1.543	13	45.5

IV. ANALYSIS OF VARIANCE

As a result of the analysis of variance, we get a table (5-7) in which the key factors for us are R-square, Significance F and all coefficients.

TABLE V. REGRESSION STATISTICS

Regression statistics	
Multiple R	0.99
R-square	0.99
Normalized R-square	0.99
Standard error	0.51
Observations	20

TABLE VI. ANALYSIS OF VARIANCE

	df	SS	MS	F	Significance F
Regression	5	348.54	69.71	266.67	0
Remains	14	3.66	0.26	-	-
Total	19	352.2	-	-	-

TABLE VII. ANALYSIS OF VARIANCE

	The coefficient	is the standard error	t-statistics	P-Value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Y-intersection	-96.91	38.68	-2.51	0.03	-179.87	-13.95	-179.87	-13.95
Variable X1	0.32	0.1	3.16	0.01	0.1	0.54	0.1	0.54
Variable X2	0.47	0.17	2.73	0.02	0.1	0.84	0.1	0.84
Variable X3	-1.16	0.89	-1.3	0.21	-3.08	0.75	-3.08	0.75
Variable X4	0.29	0.07	4.19	0	0.14	0.43	0.14	0.43
Variable X5	2.15	0.87	2.48	0.03	0.29	4	0.29	4

The R-square describes how accurately the constructed model describes the reality of this function, the maximum value of the R-square is 1, that is, 100% correctly describes the reality, the model is considered good if this coefficient is higher than 0.8; in our case, the R-square is 0.985, which means the quality of the model is close to reality; F-significance is how adequate the resulting equations are, the lower this value, the better, we have this value equal to 2.4656950212146E-12.

The resulting coefficients are the coefficients that are needed to build the regression model function. The function (1) itself looks like [18]-[20]:

$$Y=A_0+A_1*X_1+A_2*X_2+\dots+A_n*X_n \quad (1)$$

where A_n – regression coefficient, X_n – influencing factor, $n \in [0, \dots, i]$, i is number of variables (cases under consideration).

V. RESULTS AND DISCUSSION

Residuals affect the application of the model in real life and possible errors in predicting future data, the larger they are, the higher the probability of error. From the presented data, it can be concluded that the errors can be insignificant (Figure 4).

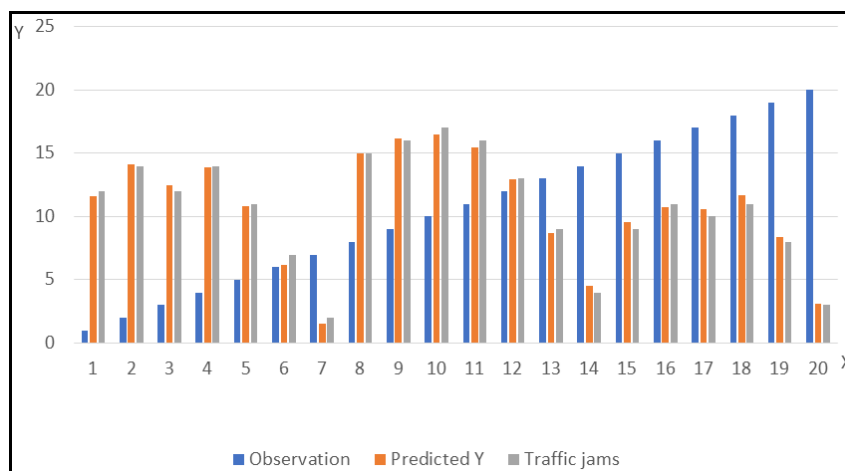


Fig. 4. Residuals: observation predicted by Y, traffic jams.

The morning factor, when it is changed by 1, the number of traffic jams will change by (0.274) this factor will also greatly affect the price of gasoline per liter (0.934).

CONCLUSION

In the course of the study, on the example of the city of millionaire, the possibility of using regression analysis in everyday life was illustrated and its effectiveness in solving a specific problem was shown - factors that influence the formation of traffic jams in the city of Krasnoyarsk. The regression analysis identified the main factors and the importance of each of them, obtaining an adequate model. The analysis revealed that the most important factor is the "number of cars", which significantly affects the factor "faulty traffic lights". A rather unexpected factor, which is in second place is the "price of gasoline". Also, with the help of regression analysis, little significant factors were found that were excluded from the final model. These were the "day" and the "number of accidents". Thus, with the help of the regression analysis, it is possible to predict and further optimize traffic flows in the city of Krasnoyarsk.

From the large amount of data, we found those that really affect the number of traffic jams getting a high level of the model. The most influential factor is the number of cars when changing it by 1 amount of traffic jams change to (0.685) as well as this factor will strongly affect the faulty traffic lights (0.974) and the price of gasoline per liter (0.954) the second most important factor is the price of gasoline per liter when an increase of 1 number of traffic jams change to (0.480) this factor does not affect other factors as it last in the 3rd place is the factor of the evening when changing it by 1 number of traffic jams change to (0.451) the same factor will strongly affect the number of cars (0.906) and faulty traffic lights (0.904) and the price of gasoline per liter (0.893) 4th place when changing it by 1 number of traffic jams change to (0.313) so this factor will greatly affect the evening (0.920) on the number of cars (0.967) for faulty traffic lights (0.987) per liter of petrol (0.930) 5th place is occupied by faulty traffic lights.

REFERENCES

[1] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.

[2] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Ereemeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.

[3] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, S. G. Dokshanin, and V. V. Kukartsev, "Research of methods for design of regression models of oil and gas refinery technological units," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042078, Jun. 2019, doi: 10.1088/1757-899X/537/4/042078.

[4] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, Y. V. Danilchenko, S. N. Ezhemanskaya, and N. V. Sokolovskiy, "Methodology for the formation of indicators balanced system for marketing activities of an industrial enterprise," IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012084, Jan. 2020, doi: 10.1088/1757-899X/734/1/012084.

[5] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.

[6] V. S. Tynchenko et al., "Software to Predict the Process Parameters of Electron Beam Welding," IEEE Access, vol. 9, pp. 92483–92499, 2021, doi: 10.1109/ACCESS.2021.3092221.

[7] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," Journal of Physics: Conference Series, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.

[8] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.

[9] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. v. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 2, p. 022106, Aug. 2019, doi: 10.1088/1755-1315/315/2/022106.

[10] A. A. Zhuravlev, O. P. Aksonova, O. I. Rubin, "Investigation of Route Completion Time in Dependence on Transport Network Parameters Using Computer Modelling," in Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBEREIT 2021, May 2021, pp. 375–378. <https://doi.org/10.1109/USBEREIT51232.2021.9455042>

[11] H. F. Yang, T. S. Dillon, Y. P. P. Chen, "Optimized structure of the traffic flow forecasting model with a deep learning approach," IEEE transactions on neural networks and learning systems, vol. 28, no. 10, pp. 2371-2381, 2016.

[12] Mochizuki, Y., & Sawada, K. (2022). An analysis of expansion and reduction speeds of traffic jams on graph exploration. *Artificial Life and Robotics*, 1-8.

[13] J. Fang, Y. Xiang, Y. Huang, Y. Cui, W. Wang, "A Vehicle Control Model to Alleviate Traffic Instability," IEEE Transactions on Vehicular Technology, vol. 70, no. 10, pp. 9863-9876, 2021.

[14] S. Fatimah, S. A. Matondang, "Simulation Model to Reduce the Traffic Jams with a Stochastic Program," WSEAS Transactions on Environment and Development, vol. 18, pp. 37-41, 2022.

[15] E. Shestеров, A. Mikhailov, "Method of evaluating transit hubs in Saint Petersburg," *Transportation Research Procedia*, vol. 50, pp. 654-661, 2020.

[16] D. Mandlik, "Pune Traffic Congestion: Reality, Cause and Regulation a Case Study," *International Journal of Management (IJM)*, vol. 11, no. 3, 2020.

[17] S. C., Lee, H. K. Kwon, "Computational algorithms and modeling for the traffic flow," *Indian Journal of Science and Technology*, vol. 9, no. 44, p. 105111, 2016. <https://doi.org/10.17485/ijst/2016/v9i44/105111>

[18] Deakin, E. (1988). Traffic Jams on Main Street. *Civil Engineering*, 58(4), 45.

[19] S. Kurashkin, V. Tynchenko, Y. Seregin, A. Murygin, V. Kukartsev, and V. Tynchenko, "Energy Distribution Modeling During the Electron Beam Welding Using Dynamically Changing Thermophysical Parameters of the Product," *Lect. Notes Networks Syst.*, vol. 230, pp. 47–58, Jul. 2021, doi: 10.1007/978-3-030-77442-4_3.

[20] V. Tynchenko et al., "Mathematical Modeling of Induction Heating of Waveguide Path Assemblies during Induction Soldering," *Metals (Basel)*, vol. 11, no. 5, p. 697, Apr. 2021, doi: 10.3390/MET11050697.

Theoretical Foundations of the Development Strategy of the Organization for the Production of the Refrigeration Equipment

Natalia Fedorova

¹Department of management
Reshetnev Siberian State University of
Science and Technology

²Department of Advertising and Social
and Cultural Activities
Siberian Federal University
Krasnoyarsk, Russia
nvfed@mail.ru

Aleksander Myrugin
Information Control Systems
Department

Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
0000-0003-2887-6162

Elena Filushina

Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
marbury@yandex.ru

Yuriy Seregin

Information Control Systems
Department
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
0000-0003-4309-8637

Dmitrij Eremeev

Department of Accounting, Finance
and Economic Security
Reshetnev Siberian State University of
Science and Technology
Krasnoyarsk, Russia
eremeev.dmitriy@gmail.com

Elena Vaitekunene

¹Department of Information Economic
Systems
Reshetnev Siberian State University of
Science and Technology
²Siberian Federal University
³Department of Civil Defence and
Emergency Management
Siberian Fire and Rescue Academy of
the Russian Ministry of Emergency
Situations
Krasnoyarsk, Russia
0000-0001-6839-6716

Abstract—This paper examines the theoretical basis for the elaboration of a development strategy for an organization producing household appliances, especially refrigeration equipment. We considered what the strategic management is and what it solves. And also studied the basic principles of strategic management. It has been found that there are two approaches to strategic management that differ in how organizations use their resources. Moreover, during the course of the work, issues related of development the strategic management of the organization were identified. It was found that it is economically beneficial for both the supplier and consumers to carry out domestic assembly of foreign equipment, as this allows manufacturers to significantly reduce customs duties and transportation costs. Thus, in order to develop an organization's development strategy, it is necessary to separate the tools to develop a strategy depending on the place of origin. It has proved necessary to apply a scientific and methodological approach, to identify and analyse the gaps and to develop patterns for selecting strategic alternatives.

Keywords— Costs, development strategy, household appliances, strategic management, theoretical foundations, control automation, optimization parameters process, technological parameters

I. INTRODUCTION

Over the past ten years, the Russian home appliances market has been formed and more clearly structured. In terms of growth rates, it occupies a leading position in the consumer goods market. To date, the household appliances market is already quite saturated, while growth rates are beginning to slow down, competition is becoming fiercer, and the entry of a new player into the market is associated with high risks and capital investments. The situation on the Russian refrigeration equipment market is characterized by

an excess of supply over effective demand, which stimulates increased competition among market players. Both “industrial refrigeration” and “commercial refrigeration” are characterized by oversupply and fierce competition between players. Under the current conditions, it is very difficult for domestic enterprises to compete with foreign manufacturers.

Under these conditions, the issues of developing an effective strategy for the development of an enterprise become relevant.

The efficiency of economic entities is determined by their strategy. Organizations that pay close attention to strategy are more competitive and resilient. The value of competitiveness, which allows the firm to survive in the competition, has recently increased dramatically. All companies in a highly competitive and rapidly changing environment must not only focus on the internal state of affairs in the company, but also develop a long-term strategy that would allow them to keep up with the changes taking place in their environment [1]-[3].

A modern tool for managing the development of an organization in the face of increasing changes in the external environment and the associated uncertainty is the methodology of strategic management.

Strategic management is an activity consisting in choosing the scope and system of actions to achieve the long-term goals of the organization in a constantly changing environment [4]-[6].

This is the area of activity of the top management of the company, whose main responsibility is to determine the preferred directions for the development of the organization, setting fundamental goals, optimal allocation of resources,

using everything that gives the organization a competitive advantage.

Strategic management acts as a process that allows the organization to interact with its environment. At the same time, strategic management is a field of knowledge about techniques, tools, methodology for making strategic decisions and methods for their practical implementation. Strategic management activities are associated with setting the goals and objectives of the organization, as well as maintaining relationships between the organization and the environment, which help it achieve its goals, correspond to its internal capabilities and allow it to remain susceptible to changes in the external environment.

II. MATERIALS AND METHODS

Strategic management solves problems:

- Overcoming the crisis state of the company, which is caused by the discrepancy between its capabilities and the requirements of the environment for occupying a leading position in the market (in the industry) in the future.
- Ensuring viability in any most unexpected situation.
- Creation of conditions for long-term development, taking into account external and internal opportunities.

The main principles of strategic management are:

- The overcoming of the unity of the company and the environment, used in setting the main goals and objectives, creating a program for their implementation.
- Orientation to the realization of the vision of the future, the mission of the company, its global quality goals, the achievement of competitiveness.
- Consideration in the formation and choice of strategies of the characteristics of the markets in which it operates, its strategic potential.

There are two approaches to strategic management: traditional and modern.

The traditional approach assumes that firms use their strengths for a strategic breakthrough in the existing competitive environment of the opportunities that open up before them [7]-[9].

The modern approach is that organizations, by manipulating their resources, form an external environment for themselves, the demands of which they can satisfy with the greatest benefit for themselves. For example, monopolies, by reducing the supply of their products and creating artificial shortages, are able to raise prices and extract excess profits.

In other words, the emphasis is gradually shifting from actions related to preparing for the future to actions aimed at its purposeful formation. At the same time, reliance is placed on personnel, as on the most valuable resource of the company, information systems and constant structural changes.

The subject of strategic management is a strategic process, the stages of which are shown in the Figure 1.

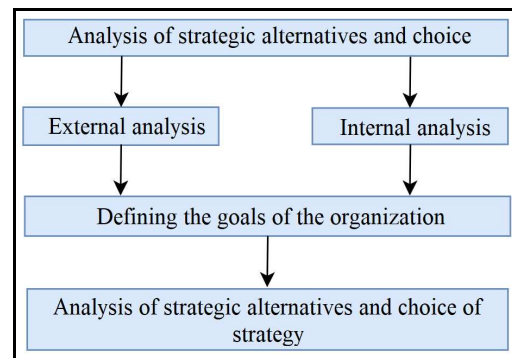


Fig. 1. Stages of the strategic process.

The main features of strategic management are:

- The mission of strategic management is aimed at the survival of the organization in the long term by establishing a dynamic balance with the environment, allowing to solve the problems of persons interested in the activities of the organization.
- The focus of strategic management is looking outside the organization, looking for new competitive opportunities, monitoring and adapting to changes in the environment.
- Strategic management is focused on the long term.
- The basis for building a management system are people, information support systems, the market.
- The personnel of the organization is its foundation, represents its main value and source of its well-being (with operational management, a view of employees as a resource of the organization, strategic management - as performers of individual works and functions).

The criterion for the effectiveness of strategic management is the timeliness and accuracy of the organization's response to new market demands and changes depending on the changing environment.

Strategic management is the management of an organization that relies on human potential as the basis of the organization, orients production activities to consumer needs, responds flexibly and makes timely changes in the organization that meet the challenge from the environment and allow achieving competitive advantages, which together makes it possible for the organization to survive in the long term, while achieving their goals [10]-[12].

III. PROBLEMS OF DEVELOPING THE STRATEGIC MANAGEMENT OF THE ORGANIZATION

Strategic problems are the discrepancy between the current state of the managed object and the set strategic goals.

The definition of strategic problems involves an assessment of the initial state of the managed object (organization), the timely determination of strategically targeted priorities, and then a comparison between them.

It is necessary to distinguish between the strategic problems of the organization and its weaknesses. The latter are determined by comparing the factors of the internal environment with those of competitors. Strategic problems

manifest themselves as a discrepancy between the strategic goals of the current state of the organization.

Modern economic realities (obsolescence of technologies, saturation of demand, rapid obsolescence of products, tougher competition) require a response, primarily to strategic problems.

Management should understand that all sources of production problems are not inside the enterprise, but in the external environment. It is noted that the balance of attention of managers to strategic and operational problems should, ultimately, be determined by the external environment in which the organization operates. If it is favorable, one can focus on operational issues, if the external environment is turbulent and volatile or demand is close to saturation, the focus should be on strategic issues.

An analysis of the activities of enterprises in various industries shows that, despite the abundance of published information and literature, almost no organization can demonstrate a more or less successful example of organizing strategic management.

The main problems of strategic management include the following:

A. The degree of validity of the choice of one or another direction of development

Despite the limited number of goals that an organization usually pursues, there are always alternatives in choosing ways to achieve these goals. The vast majority of well-known solutions in this area come down to a format convenient for reflection and discussion: all kinds of matrices, diagrams. However, they do not give the slightest idea about the real costs and income of the organization, about the value of the organization's business in certain areas.

Of course, the involvement, understanding and motivation of employees at all levels of company management is important, but this does not solve the problems of valuation of the strategies under consideration. And, besides, given the mindset of managers, it is hardly worth hoping for the correct choice of development strategy. Yes, and the economics of enterprises is not at the level to hope to solve problems with the help of conceptual "squares".

B. The problem lies in the one-sided approach to the development and implementation of the strategy

As you know, there are two levels of strategies: corporate strategy and business strategy. Corporate strategy is to determine the development path of the company as a whole.

C. With what product (service) and in what market the organization will work

Business strategy contains the entire set of actions to organize the production of a product or service with which the company has to work on the market. The connection between these two levels is obvious: the better the business strategy is implemented (production flexibility, high manufacturability, low costs, and so on), the more degrees of freedom in adopting a corporate strategy through effective diversification of production, product differentiation, and so on.

The relatively stable economy and the transparency of the activities of Western companies provide an opportunity for an approximate assessment of the quality of the implementation of a business strategy in a number of ways, and therefore the main emphasis is on the implementation of a corporate strategy. A similar approach is used by domestic enterprises. But the specifics of the connection between corporate strategy and business strategy for domestic enterprises is such that the unused, underestimated production potential of an enterprise can often turn all ideas about a seemingly successful corporate strategy adopted.

D. Banal confusion of long-term and strategic planning

The main difference between long-term (sometimes called corporate) and strategic planning lies in the interpretation of the future. The system of long-term planning assumes that the future can be predicted by extrapolation of existing growth trends, changes in the dynamics of the company's financial indicators.

In the system of long-term planning, the indicators to be achieved are, rather, the goal in the strategic management system, this is a guideline. In the first case, in the course of the enterprise's activity, the degree of achievement of the goal is controlled by comparing the actual and planned intermediate indicators. In the second case, it is the same, but each subsequent step is adjusted (through the system of management decisions) according to the results of the previous one, focusing on the final indicators. In the first case, the need to achieve the goal is unambiguous. In the second case, in the process of implementation, the landmarks can be replaced.

Thus, in the system of strategic management there is no assumption that the future must necessarily be better than the past, and it is not believed that the future can be studied by extrapolation.

The process of strategic management assumes that at each time interval alternatives will be selected to achieve the goals with maximum effect.

E. The problem arises at the stage of implementing the strategy

It lies in the "discontinuity" of the strategy along the vertical. Traditionally, the process of developing a strategy involves going through several stages. For example, how do most authors imagine the sequence of developing a strategic plan:

- Development of the mission and goals of the organization, assessment and analysis of the external environment.
- Management survey of strengths and weaknesses.
- Analysis of strategic alternatives, choice.
- Implementation and evaluation of the strategy.

But it would seem that a well-thought-out and understandable development strategy for management fails as soon as it comes to its implementation. And the point is not only that the appropriate functional strategies are poorly formulated and developed. In the vast majority of cases, having received a certain target setting (understandable at a high level), the head of department, manager, production engineer, who is used to dealing with very specific technical, technological parameters, does not understand

how to proceed with its implementation. If during the development of the strategy, tasks were discussed and solved at the level of financial and economic indicators, then the head of production should solve it at the "technical" level: what equipment should be changed first of all? what - in the second? And is it really worth changing?

F. "Instability" of adopted strategic plans

A well-designed strategic development plan falls apart at the first changes in transport tariffs, energy tariffs, violations of the structure of suppliers and their working conditions, and so on, not to mention the rapid assessment of emerging opportunities. To what extent are operational decisions taken in these cases consistent with the adopted development strategy? Or maybe, taking into account the influence of external factors, generally switch to a different strategy? To what extent, under the critical influence of external factors, should the adopted strategy be followed? And if circumstances force us to abandon the adopted strategy, then which one and how should we switch to? The dynamics of business development in modern conditions requires that the answers to these questions be given "yesterday".

G. The problem concerns the organizational side of strategic management

It can be viewed from two points of view: from the point of view of the organizational structure of the company in general and from the point of view of the strategic management unit.

1) *Aspect 1.* At present, new commercial structures are constantly being organized: a certain set of a number of enterprises (sometimes from different industries), united by one control superstructure. At the same time, management is carried out solely on the basis of financial and economic indicators. Naturally, the expediency of the participation of technologists in the management of large companies is being questioned. At first glance, it would seem that the right decision and quite a market one. If one of the enterprises is "fevery", then it must either be sold or tried to "treat" using drastic measures.

But an analysis of trends in the development of business management systems shows that more and more methods are emerging that take into account both financial and non-financial indicators.

When developing a strategy, we must answer a number of questions, namely: how can the market (and external environment) in which the enterprise operates potentially change? How should the company respond to this change? how should the product, its functionality and cost performance change in response to likely market changes? how to change the functional qualities of the product, the production technology, volumes and structure of resource support should change? how should business processes, management system and organizational structure at the enterprise change to ensure an adequate response to the market?

In this case, it is quite obvious that, by switching to managing an organization on the basis of only financial and economic indicators, the owners risk losing sight of one of the key components of the strategic plan - the use of new modern technologies, equipment and equipment. An example is the recent announcement that film cameras will

be discontinued. It is impossible to take into account such a factor when managing a strategy at the level of only financial and economic indicators.

This suggests that, along with financial and economic employees, employees with technical education should also participate in strategic management - specialists who are able to see and foresee (predict) the consequences of the emergence and development of new technologies, machinery, equipment, as well as specialists in the field of business management.

2) *Aspect 2.* Strategic plans are developed and adopted no more than once a year. As a rule, divisions of various levels are engaged in its development to varying degrees. A group of 3 to 5 people is involved in ensuring the development and preparation of the strategic plan. In this case, the final decision is made by the management, the owners of the company. But once the plan is developed, the group remains idle. Of course, one can object: "And control over the implementation of the plan?" For effective control, it is necessary to have a set of certain benchmarks by which you can assess the degree of deviation from the original plan. But monitoring is not enough, it is necessary to quickly find a solution that allows using the current situation with maximum benefit.

IV. RESULTS AND DISCUSSION

Home appliances are an integral part of people's lives. The transition to new business conditions, which in itself became very painful for many enterprises, coincided with an increase in the supply of foreign equipment and a fall in the purchasing power of the population.

To increase the competitiveness of products, it was necessary to modernize production in accordance with modern standards and launch new models, but most enterprises simply did not have the money for this. All this led to a reduction in unprofitable production, to the shutdown of enterprises and the almost complete disappearance of domestic electrical appliances from store shelves.

The devaluation of the ruble that took place had a positive effect: the consumer began to look closely at domestic products, which immediately turned out to be much cheaper than imported ones, and in terms of design and quality they were already close to foreign models. In addition, after the crisis, foreign investors moved to Russia, who decided that under these conditions it would be more profitable to open their production in Russia than to import equipment into the country. Thus, the growth in demand stimulated the growth of production.

We single out the typical features of household appliances [13]-[16]:

- Goods of mass and durable consumption.
- Variety of product brands on the market.
- Promotion of goods through the Internet (Internet sales).
- Modern and multifunctional equipment.
- Variety of built-in home appliances.

In this paper, refrigeration equipment is studied in more detail.

The capacity of the Russian market of commercial refrigeration equipment reaches 210-230 million rubles. Commercial refrigeration equipment occupies a significant segment of the entire refrigeration market and, according to most experts and company representatives, further dynamic development of this sector is expected. This is due to the influence of such a factor as the rapid development of the trade sector, including large retail chains, which accordingly increases the demand for trade equipment by an average of 25% per year. However, at the same time, the lack of specialized enterprises for the production of high-quality components and the generally undeveloped infrastructure of the domestic industry hinder the growth of production volumes and lead to an increase in imports of commercial refrigeration equipment (RTW) from abroad.

Considering the list of THW on the market, it can be classified as follows: freezing and refrigerating display cases, racks (hills, racks), cabinets, chambers, combined display cases, mobile refrigeration units, ice generators and other special equipment. All appliances differ in many ways, including size, price characteristics, manufacturer's brand, and much more [9].

Each category of products, depending on the modifications, is in greater or lesser demand among consumers. For example, according to statistics, the market demand for chest freezers with metal lids is about 18%, for products with straight glass lids - 45%, with curved glass doors - 15%, for chest freezers with inclined glass lids - 22%. As for islands, the situation is as follows: low-temperature islands with a mode from -25 to -18 °C account for about 65% of demand, products with a temperature regime of -25 ... + 8 °C - 25% and about 10% fall on other refrigeration equipment.

Depending on the temperature regime maintained in the refrigerated volume, medium-temperature, low-temperature and combined equipment is distinguished. With regard to refrigeration, the market offers installations with remote or built-in refrigeration, as well as equipment for connecting to central refrigeration. It is on the last point that it is worth dwelling in more detail.

Due to the changing working conditions of modern trade, the active introduction of modern technologies and a significant increase in competition, trade enterprises are faced with such tasks as reducing their own costs and meeting the growing needs of potential buyers. Therefore, many domestic and foreign manufacturers have developed commercial refrigeration equipment for central cooling, which successfully helps to solve the above problems.

Firstly, this equipment can be built in lines of any length with arbitrary bending angles of the lines, which allows designing sales areas according to individual projects and taking into account the specific features of the stores.

Secondly, the remoteness of the refrigeration units in the utility rooms significantly reduces the noise level in the trading floors.

And finally, the life of the central cooling compressors is 2.5 times longer than that of installations with built-in units.

With regard to new developments in this area, it became possible to computer monitor the central refrigeration system with automatic scheduling of temperature changes and monitoring the pressure level in the compressor.

Modern refrigerator factories are, in fact, assembly plants. A high level of mechanization and automation of technological processes, combined with an extensive unification of units and parts, ensures high quality goods at minimal manufacturing costs. Directly at the factories, cabinets, chests and doors are made, which are not economically feasible to transport over long distances. For deliveries over many kilometers, one would have to "carry air", since the cabinets take up a lot of space and weigh little.

There are two main product groups in the market under consideration:

1) *Group "Industrial refrigeration", which includes the following types of equipment:*

a) *Equipment for freezing and storage at low and medium temperatures of food in storage warehouses.*

b) *Equipment for technological conditioning of commercial real estate and production shops.*

c) *Refrigeration equipment to accompany all kinds of production processes.*

The main consumers of such equipment are: warehouses, food processing enterprises, industrial enterprises, commercial real estate objects, breweries and enterprises for the production of soft drinks.

2) *Group "Commercial refrigeration", includes the following types of equipment:*

a) *Commercial refrigeration equipment.*

b) *Refrigerators of small sizes.*

c) *Refrigeration equipment for technological processes in the catering system.*

d) *Commercial district cooling systems.*

The main consumers of this equipment are: shops, supermarkets, food markets, as well as catering establishments and small breweries.

Note that the predominant share of the Russian market is industrial refrigeration equipment - sixty percent, commercial refrigeration equipment accounts for forty percent of the market. The data is presented in Figure 2.

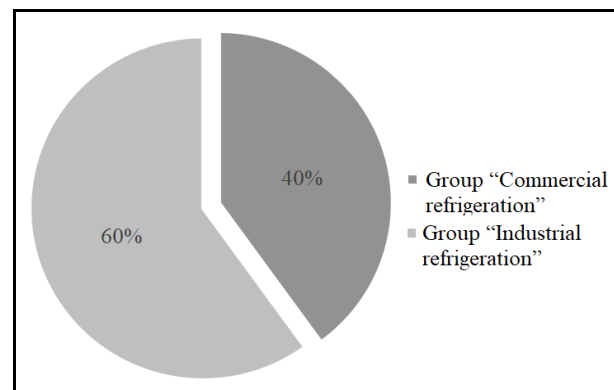


Fig. 2. Structure of the Russian market of refrigeration equipment.

The most cost-effective, both for suppliers of refrigeration equipment and for its consumers, is the

assembly of foreign equipment in Russia. The quality of the Russian assembly is quite satisfactory, at the same time, this allows manufacturers to significantly reduce customs duties and transportation costs, and consumers to purchase equipment at a lower price than imported equipment, in a shorter period of time.

The annual increase in sales of commercial refrigeration equipment is about thirty, thirty-five percent and is provided mainly by the active development of large retail chains.

At the same time, the growth in the production of commercial refrigeration equipment is constrained by the underdeveloped infrastructure of the domestic industry, which leads to an increase in imports.

Describing some of the main trends in demand for refrigeration equipment, it should be noted that the main consumers of refrigeration equipment are chain supermarkets, food plants, breweries and industrial refrigerators.

Thus, the prospects for the development of the refrigeration equipment market in the coming years are directly related to the prospects for the development of medium and large enterprises in the processing industry and trade [17]-[20].

Among the main buyers of units it is worth noting the retail trade, which is one of the fastest growing and profitable segments of the refrigeration equipment market. Today, most of the world's retail formats are represented in Russia - stalls, shopping pavilions "within walking distance", discounters, specialized stores, super- and hypermarkets.

With regard to suppliers, the main types of sales organization can be divided into three categories:

- Realization of equipment directly without intermediaries, which is typical for equipment manufacturers operating in the local market with the simultaneous organization of other promotion channels.
- The most common dealer organization of sales, which assumes that the manufacturer sells products to dealers at special prices with the provision of warranty service and a number of other benefits.
- Representative sales consist in the organization of regional and international representative offices through which goods are sold. Among the main advantages of this type of sales is central management and a single pricing policy.

It should be noted that the current situation in the Russian refrigeration equipment market is characterized by an excess of supply over effective demand, which stimulates increased competition among market players. Both "industrial refrigeration" and "commercial refrigeration" are characterized by an overabundance of supply, fierce competition between players.

CONCLUSION

Under the current conditions, it is very difficult for domestic enterprises to compete with foreign manufacturers. According to experts, only a significantly lower price in relation to imported analogues allows domestic equipment

to remain competitive to some extent, while the quality and level of service are significantly inferior.

Considering the general trend, it can be noted that the domestic industry is developing very rapidly and in record time, the only limitation is the financial capabilities of some enterprises. However, even in such conditions, domestic manufacturers went far ahead, new modifications of commercial refrigeration equipment were introduced to the market, such as Economy class with built-in refrigeration, with a smooth transition to the model range of more elite Business class units with an external cooling system, etc. Also, manufacturers are mastering a variety of related niches and the production of equipment of the "Lux" and "Centroholod" classes.

Thus, sectoral features of the activities of the organization for the production of household appliances were identified when developing a development strategy. These features require the organization of strategy development tools according to the places of occurrence, there is a need to apply a scientific and methodological approach, identify and analyze deviations and draw up a scheme for choosing strategic alternatives. To facilitate the choice of strategic development tools, it is recommended to find an approach to strategic development, and also, if it is appropriate, apply this method to a particular firm.

REFERENCES

- [1] R. J. Baumgartner, and R. Rauter, "Strategic perspectives of corporate sustainability management to develop a sustainable organization," *Journal of Cleaner Production*, vol. 140, pp. 81-92, Jan. 2017, doi.org/10.1016/j.jclepro.2016.04.146.
- [2] V. V. Velikorossov, et. al., *Strategic management*, 2020.
- [3] J. Bryson, and B. George, "Strategic management in public administration," *Oxford Research Encyclopedia of Politics*. – March 2020, doi.org/10.1093/acrefore/9780190228637.013.1396.
- [4] L. Á. Guerras-Martín, and A. Madhok, "Montoro-Sánchez Á. The evolution of strategic management research: Recent trends and current directions," *BRQ Business Research Quarterly*, vol. 17, no. 2, pp. 69-76, 2014.
- [5] J. Rosenberg Hansen, and E. Ferlie, "Applying strategic management theories in public sector organizations: Developing a typology," *Public Management Review*, vol. 18, no. 1, pp. 1-19, 2016.
- [6] V. S. Tynchenko, N. V. Fedorova, V. V. Kukartsev, A. A. Boyko, A. A. Stupina, and Y. V. Danilchenko, "Methods of developing a competitive strategy of the agricultural enterprise," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 2, p. 022105, Aug. 2019, doi: 10.1088/1755-1315/315/2/022105.
- [7] J. Lundén, V. Vanhanen, T. Myllymäki, E. Laamanen, K. Kotilainen, and K. Hemminki, "Temperature control efficacy of retail refrigeration equipment," *Food Control*, vol. 45, pp. 109-114, 2014.
- [8] A. A. Boyko, V. V. Kukartsev, K. Y. Lobkov, and A. A. Stupina, "Strategic planning toolset for reproduction of machinebuilding engines and equipment," *Journal of Physics: Conference Series*, vol. 1015, no. 4, p. 042006, May 2018, doi: 10.1088/1742-6596/1015/4/042006.
- [9] L. Liu et al., "An unsupervised model for classification and recognition of household appliances," *Journal of Computational Information Systems*, vol. 10, no. 1, pp. 403-410, 2014, doi: .
- [10] P. Augustyniak, "Sensorized elements of a typical household in behavioral studies and prediction of a health setback," in *2015 8th International Conference on Human System Interaction (HSI)*, pp. 254-259, July 2015, DOI: 10.1109/HSI.2015.7170676.
- [11] V. V. Shishov, and M. S. Talyzin, "Efficiency of Refrigeration Equipment on Natural Refrigerants," *Chemical and Petroleum Engineering*, vol. 56, no. 5. pp. 385-392, Sep. 2020, doi.org/10.1007/s10556-020-00785-w.
- [12] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," *IOP Conference*

Series: Earth and Environmental Science, vol. 315, no. 2, p. 022106, Aug. 2019, doi: 10.1088/1755-1315/315/2/022106.

- [13] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.
- [14] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, V. V. Kukartsev, and A. I. Kuklina, "Evolutionary method for automated design of models of vortex flowmeters transformation function," *Journal of Physics: Conference Series*, vol. 1118, no. 1, p. 012041, Dec. 2018, doi: 10.1088/1742-6596/1118/1/012041.
- [15] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.
- [16] N. V. Fedorova, N. N. Dzhioeva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogol, and V. S. Tynchenko, "Methods of assessing the efficiency of the foundry industrial marketing," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.
- [17] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.
- [18] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," *Journal of Physics: Conference Series*, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.
- [19] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Eremeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.
- [20] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.

Methods and Tools for Developing an Organization Development Strategy

Vladislav Kukartsev

¹Department of Information Economic Systems

Reshetnev Siberian State University of Science and Technology

²Department of Informatics Siberian Federal University

³Digital Material Science: New Materials and Technologies

Bauman Moscow State Technical University

Krasnoyarsk, Russia

0000-0001-6382-1736

Elizaveta Shutkina

Information Control Systems Department

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

elizaveta-shutkina@mail

Kristina Moiseeva

Information Control Systems Department

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

kristina2001irba@mail.ru

Larisa Korpacheva

Department of Digital Management Technologies

Siberian Federal University

Krasnoyarsk, Russia

corp_0777@mail.ru

Timofey Kireev

Department of Information Economic Systems

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

sofyachash@mail.ru

Abstract—The article considers the role of methods in the organization of management activities. The methods of formation of the organization's development strategy are singled out, how they are applied and the use of methods in the formation of the organization's development strategy is justified. Groups of tools for forming an organization's development strategy, which must be used to assess the factors of an organization's competitiveness, have also been studied, and models for using these tools have been considered. As a result, the factors that determine the competitiveness of the enterprise were identified. It was revealed that the main idea of the market competition comes from the struggle to obtain the greatest benefit through the best efficient use of economic resources with their limitations, and also that the competitiveness of a trading company in a market economy is the connecting part of the management of an economic entity, revealing the rate of efficiency of use of economic resources. economy of economic resources opposite to the efficiency of the use of economic resources by potential competitors. It was concluded that strategic development tools are a scientific and methodological approach related to the choice of the proposed strategic planning tools.

Keywords— Development strategy, company, methods and tools, organization, UML

I. INTRODUCTION

The role of methods in the organization of management activities is no less important than the role of technologies, since with the help of methods we can navigate through numerous strategies, identify, classify and group them, develop typical schemes for searching, selecting and implementing strategies [1]-[4].

In the practice of strategic management, a system of techniques has developed that allows planning the strategic position of an enterprise. These are the so-called strategic management models. They were developed at various enterprises in order to plan their future strategic position aimed at providing competitive advantages.

There are many strategic models that can be used in a wide variety of situations. Many of them cannot be applied

at other enterprises, since they are based on taking into account the specifics of a particular production system. Others, on the contrary, are universal [5]-[8].

II. METHODS FOR FORMING THE ORGANIZATION DEVELOPMENT STRATEGY

To date, the most developed for practical application in the enterprise are such methods of strategic management as:

A. Management Method by Ranking Strategic Objectives

This method consists in early detection of unexpected changes both inside and outside the enterprise and quick response to them.

As part of management using ranking, the following activities are performed [9]-[11]:

- Constant monitoring of trends in the external environment is carried out.
- An analysis of the identified trends in changes in the external environment is carried out and an assessment of the urgency of decision-making is carried out, which are brought to the attention of the top management of the organization.
- The top management and planning and economic service of the enterprise considers the results of the analysis of external and internal trends of the enterprise and, in turn, ranks them into four categories:
 - a. The most urgent and important tasks that require immediate consideration are sent to the study, during which acceptable decisions are developed and adopted by existing parts of the organization.
 - b. Important tasks of medium urgency that can be solved within the next planning period.
 - c. Important but non-urgent tasks that require constant monitoring.

- d. Tasks that are insignificant for the enterprise and do not deserve further consideration.
- The top management of the enterprise controls the decisions made by the company's divisions and evaluates them in terms of possible strategic and tactical consequences..
- Management should continually review and update the list of emerging issues and their prioritization.

From the point of view of practical use, the management method by ranking strategic tasks is a relatively simple system for tracking trends in changes in the external and internal environment of the organization.

B. Weak and Strong Signal Control Method

In the decision-making process in organizations, a large role is given to the stage of problem recognition. Is the firm able to identify the problem based on the available information, evaluate its significance and take appropriate measures to solve it? If "yes", then we are dealing with information of such quantity and quality that we define it as "strong signals". If "no", then the point is the "weakness" of the signals. We could identify many impending problems if we could learn to identify and account for the so-called "weak signals" - early inaccurate signs of upcoming important events [12]-[15].

The basis of the method of strategic management by weak signals is the development of strategies for "weak reactions" (cautious, preliminary) of the company in the external and internal environment.

Technological operations of the method are as follows:

- Establish warning-sensitive surveillance and identify "weak signals".
- Identification of problems and assessment of consequences.
- Development of alternative "weak reactions" and selection of the preferred reaction.
- Establishing possible responses and response dynamics, as well as diagnosing readiness to respond.

The strong signal control method is a very common method. Indeed, situations often arise in firms when individual specialists have been talking about impending threats for a long time, but due to various filters operating in the organization, the decision to react is made only when the situation becomes clear to everyone, including employees of the firm at all levels, consumers, suppliers and important to competitors. And when the decision is made, it turns out that there is no more time.

C. Method of Management in the Conditions of Strategic Surprises

In real life, some problems elude observers, no matter how hard they try to identify them, and turn into strategic surprises. It means that:

- Problem occurs suddenly and unexpectedly.
- It sets new challenges that do not match the past experience of the organization.

- An organization's failure to take adequate countermeasures results in either major financial damage or reduced profit opportunities.
- Countermeasures must be taken urgently, but the organization's normal operating procedures do not allow for this.

In such cases, as a rule, the old strategies are not suitable: the tasks are new, the information that needs to be comprehended is escalating, creating overload for decision makers. An initiative from below in the face of strategic surprises, numbing of a systemic nature, can only aggravate the situation.

The characteristic features of this system are as follows [16]-[18]:

- In times of strategic surprise, an emergency communication network is activated that operates across organizational boundaries, filters information, and quickly communicates it to all parts of the organization.
- During the state of emergency, management responsibilities are redistributed: one group devotes their attention to monitoring and maintaining a healthy morale in the organization, another to carry out normal work with a minimum level of disruption, and a third to take emergency measures.
- A network of task forces is put in place to develop contingency measures:
 - a. Heads and members of operational groups, despite the established channels of intra-organizational relationships, constitute units or groups of strategic action, not planning.
 - b. Communication between the operational teams and the senior management team is built directly.
 - c. A management group belonging to the top management formulates the overall strategy, allocates responsibility among the performers and coordinates management.
 - d. Grassroots task forces carry out work in their areas as part of an overall strategy.

Operational groups and links between them are formed in advance and are tested. To do this, several communication systems for various purposes can be organized in advance: one for solving unexpected sales problems, another for technology, a third for politics, etc. At the same time, operational teams learn to quickly respond to fundamentally new problems by combining precise methods of analysis with a creative approach.

The central objective of the method is to ensure the strategic flexibility of the organization.

D. Method of Management Through the Choice of Strategic Positions

One of the disadvantages of strategic planning is that this type of management does not take into account the capabilities of the company itself to implement certain strategies. As a result, in any case, the company's managers need to check (evaluate) the strategy for feasibility.

Management through the choice of strategic positions is a management in which the planning of the company's strategy is carried out simultaneously with the planning of the capabilities (resources) of the company. This allows us to weed out strategies that are not feasible in advance at an early stage and select those that are most likely to be implemented. In addition, while implementing the company's strategy, along with actions aimed at improving the competitive position of the company, its resource base is also developing. The resource base of the company forms its

functional potential, i.e. the potential of the firm's marketing, manufacturing, R&D, financial services, and general corporate management skills.

III. TOOLS FOR FORMING THE ORGANIZATION DEVELOPMENT STRATEGY

To assess the factors of the company's competitiveness, it is necessary to apply the tools of strategic analysis, which are shown in Figure 1.

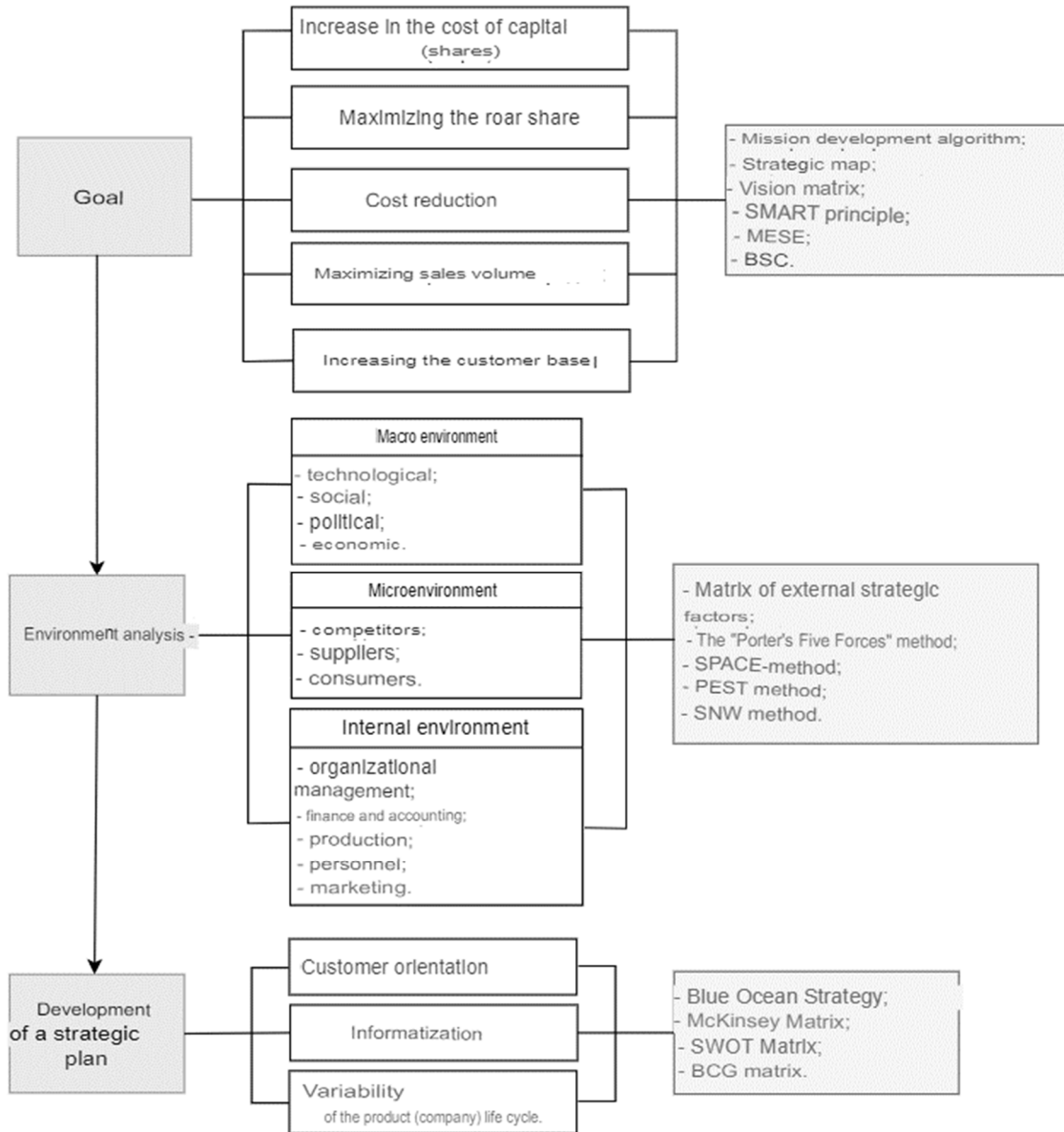


Fig. 1. Strategic Analysis Tools.

Goal setting and decomposition tools: the classic tree of goals, SMART - analysis, and management tools that have recently appeared in management theory: the SLRO paradigm (Socio-Labor Relations-Organization) and the model (methodology) for the development of VPM, based on the paradigms "as is" and "it should be". At the same time, when setting goals, no fundamental changes are required from what has been achieved, the well-known "tree of goals" works. If it is required to set high goals that are unattainable with the current organization, then it is

necessary to set goals for carrying out changes - a model (methodology) for the development of HRP is needed.

A. SMART Analysis

Currently, one of the most common methods for determining the goals of the organization is the SMART technology. Accordingly, the goal should be [19]-[21]:

- Specific.
- Measurable.
- Achievable.

- Realistic.
- Time-Bound.

B. Enterprise paradigm Social and Labor Relations - Organization

A paradigm is a necessary and sufficient clearly structured system of key provisions that determines the formal and informal organization of an enterprise, all its activities, and the results it achieves.

The need to form a paradigm is determined by the fact that the vision of an enterprise is a huge, unlimited array of heterogeneous subjective information. In fact, the paradigm is a structured vision of the enterprise in all major aspects of its activities. The SLRO paradigm has a basis - social and labor relations in the enterprise and its organization (formal and informal) determined by this basis. The practical use of the SLRO enterprise paradigm leads to the resolution of most of the problems of both management consulting and independent organizational change.

C. Model (Methodology) of the Development of HRP

(Vision-Paradigm-Model - 2016) - the most modern model, based on the SLRO paradigm, which allows you to solve both the development of the enterprise as a whole and the problems that arise in the course of its activities.

Model steps:

- Building on the basis of the current vision of the system of necessary and sufficient key provisions that determine the activities of the enterprise - the "as is" paradigm.
- Modeling, strategic analysis and obtaining a system of key provisions that provide the maximum achievable results of the enterprise - the paradigm "as it should be".

At the same time, the goals themselves are defined, the system of personnel motivation to achieve the accepted goals, the joint development of the organizational structure and the main provisions of the enterprise strategy, otherwise it is simply impossible to ensure the setting of maximum, but achievable goals of the enterprise.

- Strategic analysis of the key provisions of the "as is" and "as it should be" paradigms, obtaining a system of changes that need to be carried out in the enterprise in order to set and achieve the highest possible goals.
- Determination of a strategy for implementing the necessary changes necessary to achieve the accepted goals of implementing these changes.

Thus, this model is both a goal setting tool and a strategy development tool. It should be noted that the VPM methodology, including the SLRO paradigm, is effective for all types of development - both the integrated development of an enterprise and the development of individual critical areas: organizational structures, business processes, organizational documentation, remuneration and motivation systems, etc.

D. Tools for Developing Strategy and Strategic Management, Organizational Development in General

The second group includes well-developed and widely used methods and tools for developing a strategy and strategic management, as well as tools for developing an organization (carrying out changes), increasing its potential, transferring it to a new higher level of development, which themselves are still constantly evolving.

Kurt Lewin's (1947) developmental model includes three stages:

- Defrosting.
- Changes.
- Freeze.

At the first stage, the need for changes is recognized and a strategy for their implementation is developed. In the second stage, the changes themselves are carried out. At the third stage, a new organizational culture is formed, making the changes irreversible. These three generalized phases are used by all subsequent development models, but they are so general that they give only the first idea of how to make changes.

Larry Greiner's Model - Model of Successful Organizational Change (1967) includes 6 stages:

- Pressure and arousal.
- Intervention and reorientation.
- Diagnostics and recognition.
- Innovation and commitment.
- Experiment with a new solution and seek improvement.
- Reinforcement and agreement.

Here, the first four stages are the unfreezing stage of the Levin Map, followed by the change and freezing stages.

John Kotter's Development Model (1995) is the latest and best-known organizational development model. It has 8 steps, they are more detailed than Kurt Lewin's, but again they give only a general direction, they do not give specific methods.

- Create an atmosphere of urgency around a big opportunity.
- Formation of a powerful governing coalition.
- Create a vision and strategy for its implementation.
- Sharing the vision.
- Enable others to act on the vision.
- Planning and creating short term wins.
- Consolidate improvements and produce more further changes.
- Institutionalize new approaches.

In this case, the first four to five stages are Kurt Lewin's unfreezing stage, then there are the change and freeze stages.

E. Management Decision Support Tools

The third group includes methods for analyzing the internal and external environment, which allow setting adequate goals for strategic management and determining the need for organizational development: carrying out organizational changes to overcome emerging problems and crises, as well as the implementation of new ideas and opportunities.

a) GAP analysis is used to develop a strategy based on the concept of solving strategic problems. GAP analysis is understood as a set of activities that allow drawing conclusions about the discrepancy between the internal marketing environment and the external environment. Its purpose is to identify those market opportunities that can become effective market advantages for the company..

The essence of this model is to determine the strategies and processes that a firm can use to achieve excellence in customer service. A simple concept, however, turns out to be difficult to apply in practice.

b) PEST analysis (sometimes referred to as STEP) is a marketing tool designed to identify political (Political), economic (Economic), social (Social) and technological (Technological) aspects of the external environment (consumers, competitors, suppliers, political and economic situation and its trends, new product trends) that affect the company's activities.

To carry out a PEST analysis, each specific company must have its own list of key environmental factors that have a significant impact on its business, contain potential threats or new opportunities for the development of the organization.

The results of the PEST analysis make it possible to assess the external economic situation in the sphere of production and commercial activities, to study the market in which the company operates or is going to operate.

c) SNW analysis is an analysis of the strengths and weaknesses of an organization. The internal environment is evaluated by three values:

- *Strength* (strong suit).
- *Neutral* (neutral side).
- *Weakness* (weak side).

As practice has shown, in a situation of strategic analysis of the internal environment of an organization, it is best to fix the average market state for this particular situation as a neutral position.

To successfully analyze the organization's environment, it is necessary to identify threats and opportunities and evaluate them in terms of their importance and degree of influence on the organization's strategy.

Usually SNW-analysis is used for a deeper study of the internal environment of the organization after the SWOT-analysis.

d) SWOT analysis is applied to study the environment. He divides the significant factors of the external and internal environment into four categories:

- *Strength*.

- *Weakness*.
- *Opportunities*.
- *Threats*.

The SWOT methodology involves first identifying strengths and weaknesses, as well as threats and opportunities, and then establishing chains of links between them, which can later be used to formulate the organization's strategy.

e) The ADL matrix (ADL) was developed by the consulting company A. D. Little. Unlike the BCG matrix, this model is based on two variables that reflect the maturity of the sector (the life cycle of the industry) and the position in relation to competitors.

The life cycle phases are as follows:

- Origin.
- Growth acceleration.
- Growth slowdown.
- Maturity.
- Attenuation.

The main theoretical position of the A. D. Little model is that a single type of business of any corporation can be at one of the stages of the life cycle described above, and, therefore, it must be analyzed in accordance with this stage.

The type of business, at the same time, can occupy one of five competitive positions:

- Dominant.
- Strong.
- Favorable.
- Durable.
- Weak.

The combination of two parameters: four stages of the production life cycle and five competitive positions - make up the so-called ADL/LS matrix, which consists of 20 cells.

f) Model (matrix) BCG (*Boston Consulting Group - BCG*). The Boston Matrix is based on the product life cycle model, according to which a product goes through four stages in its development.:

- Entering the market (product "problem").
- Growth (product "star").
- Maturity (goods "cash cow").
- Recession (product "dog").

To assess the competitiveness of certain types of business, two criteria are used:

- Growth rate of the branch market.
- Relative market share.

Thus, the division of business types (individual products) into four different groups is carried out.

Ideally, a balanced nomenclature portfolio of an enterprise should be formed in such a way that the enterprise operates effectively both in the short and long term.

g) The “General Electric”-“McKinsey” Matrix was developed by the McKinsey Consulting Group in conjunction with the General Electric Corporation and is called the “Business Screen”. It includes nine squares, and the analysis in this matrix is carried out according to the following parameters:

- Attractiveness of SZH.
- Position in competition.

The indicator “attractiveness of SZH” is not controlled by the company, i.e. those that one or another economic entity can only fix and focus on them. The indicator “position in competition”, on the contrary, depends on the results of the activities of the business entity itself.

h) “The Five Forces of Competition model” by M. Porter is the most well-known and widely used model for assessing the attractiveness of an industry. Porter's five forces include:

- Threats of substitute products.
- Threats of new players.
- Bargaining power of suppliers.
- Bargaining power of consumers.
- Level of competition.

This model makes it possible to more purposefully assess the competitive situation in the market and, on this basis, to develop such a variant of the long-term strategy of the company, which will ensure its protection from the impact of competitive forces to the greatest extent and at the same time will contribute to the creation of additional competitive advantages.

i) I. Ansoff's model is used to develop strategic alternatives. It allows you to use several strategies at the same time. This model is based on the premise that the most appropriate strategy for strong sales growth can be determined by the decision to sell existing or new products in existing or new markets. I. Ansoff's matrix is intended to describe the possible strategies of an enterprise in a growing market.

On one axis in the matrix, the type of product is considered - old or new, on the other axis - the type of market, also old or new.

The advantages of using planning according to the I. Ansoff matrix are visibility and ease of use. The disadvantages of using planning according to the I. Ansoff matrix are a one-sided focus on growth and restrictions in the context of two characteristics (product - market).

F. Strategy Implementation Tools

The tools for implementing the organization's strategy include the following:

a) Leadership is a belief, a motivation, a culture. Leadership is key to the implementation of the strategy, because it ensures the assimilation of new behaviors that are

necessary for the implementation of the new strategy. Flexibility and openness in the context of globalization are becoming one of the main managerial qualities.

Managers create and maintain a culture that underpins the strategy. Leadership unites employees around a new vision for the future of the company and creates commitment to a new strategy. Culture is a link between the strategy and the results of its implementation, each strategic area of activity has its own structure.

b) Structure is the organizational structure, teams, degree of decentralization, distribution of production capacities. Usually, a new strategy requires adjustment of the organizational structure: positions are added and changed, teams are reorganized, etc. These changes should support the new strategy.

c) Information and control systems are a system of remuneration and incentives, resource allocation, information and technology systems, rules and procedures. These are the main tools for implementing the strategy. Employees are rewarded for implementing new ideas. Additional resources are received by units that make the greatest contribution to the achievement of strategic goals. New information technologies support ongoing changes. Firms seek to meet the individual needs of consumers using the Internet.

d) Human resources are the most valuable asset of an organization, its employees. Recruitment, training and promotion of employees is carried out in the interests of achieving the strategic goals of the company. Implementation of changes causes resistance, to overcome which special programs are formed.

The factors that ensure the competitiveness of a trading company directly depend on the goals of a trading company. This process is interconnected and is directly proportional. In order to strengthen the ability to compete, the firm needs to consolidate its position in the market.

The competitiveness of an enterprise is determined by the following factors:

- Quality of sold products and services.
- Having an effective marketing and sales strategy.
- Qualification level of personnel and management.
- Technological level of production.
- The tax environment in which the trading firm operates.
- Availability of funding sources and their attraction.

There are several ways to increase the level of competitiveness of firms.:

- Regular use of innovations.
- Search for new, higher quality products.
- Release of goods of such quality which would correspond to the state and world standards.
- Sales of goods to those segments of the industry, where the highest requirements for service and quality.

- The use of only high-performance material and raw materials.
- Regular retraining and training of personnel.
- Increasing the monetary interest of personnel and improving the level of working conditions.
- Implementation of marketing research of the industry, in order to fix the needs of consumers.
- Analysis of potential competitors to identify their strengths and weaknesses.
- Keeping contacts with research companies and investing in research and development that affect the level of product quality.
- Application of the most effective marketing research.
- Accounting for personal product superiority over potential competitors.

Using such paths, the firm will be able to increase both personal competitiveness and consolidate monetary stability.

IV. RESULTS AND DISCUSSION

The best way to generate income in a market economy is the sale of goods and the surplus value included in it. At the same time, the creation and sale of goods is carried out by

using economically limited materials. From this it follows that the extraction of income in a market economy is determined by the volume of applicable economic potentials or the ratio of the resulting conclusion and the costs made to obtain them. Thus, the main idea of market competition comes from the struggle to obtain the greatest benefit through the best efficient use of economic resources with their limitations.

The degree of effectiveness of the use of economic resources by a firm correlates with the relativity of the level of development of the expended forces received by the social creation and, as a result, production and other relationships regarding the efficiency of spending resources by potential competitors. Thus, the competitiveness of a trading company in a market economy is the connecting part of the management of an economic entity, which reveals the rate of efficiency in the use of economic resources by the economic economy, which is opposite to the efficiency of the use of economic resources by potential competitors.

The level of competitiveness of an enterprise depends on many factors that can be conditionally grouped; the scheme for presenting factors is presented in Table I.

TABLE I. FACTORS PROVIDING THE STRATEGIC COMPETITIVENESS OF A TRADING COMPANY

	Factors	External	Internal	Relevant	Irrelevant
Competition factors	1. Increase in the cost of capital (shares)	+	-	-	+
	2. Increasing the customer base	+	-	+	-
	3. Maximizing market share	-	+	+	-
	4. Cost reduction	-	+	-	+
	5. Maximizing sales volume	-	+	+	-
Consumer factors	6. Political	+	-	-	+
	7. Economic	+	-	-	+
	8. Social	+	-	+	-
	9. Technological	+	+	+	-
	10. Competitors	+	-	+	-
	11. Suppliers	+	-	-	+
	12. Consumers	+	-	+	-
	13. Organizational management	-	+	-	+
	14. Staff	-	+	-	+
	15. Production	-	+	-	+
	16. Marketing	-	+	-	+
	17. Finance and accounting	-	+	-	+
Factors of informatization	18. The level of informatization	-	+	+	-
	19. Variability of the product life cycle, the life cycle of the company	+	+	-	+
	20. Customer orientation	-	+	+	-

CONSLUSION

It is necessary to understand that the problem of increasing the competitiveness of a trading firm is to reduce the bad inclination of the competitive environment, To determine which competitive advantages should be issued and controlled for strategic development planning, as they alone reveal the market position of a trading firm in a competitive environment.

And since there is no common definition of «tools of strategic planning», its object and object, goals and objectives, then, based on what has been considered in the work, it is possible to give the following definition. Tools of

strategic development - a scientific and methodical approach, connected with the choice of the proposed tools of strategic planning.

REFERENCES

[1] R. J. Baumgartner, and R. Rauter, "Strategic perspectives of corporate sustainability management to develop a sustainable organization," Journal of Cleaner Production, vol. 140, pp. 81–92, 2017.

[2] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.

[3] D. S. Shalaeva, O. I. Kukartseva, V. S. Tynchenko, V. V. Kukartsev, S. V. Aponasenko, and E. V. Stepanova, "Analysis of the

- development of global energy production and consumption by fuel type in various regions of the world,” IOP Conference Series: Materials Science and Engineering, vol. 952, no. 1, p. 012025, Nov. 2020, doi: 10.1088/1757-899X/952/1/012025.
- [4] P. Jarzabkowski, and S. Kaplan, “Strategy tools-in-use: A framework for understanding “technologies of rationality” in practice”, Strategic management journal, vol 36, no. 4, pp. 537–558, 2015.
- [5] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, “The simulation model of fixed assets reproduction of mechanical engineering enterprises,” International Russian Automation Conference (RusAutoCon), IEEE, pp. 1-6, Sep. 2018.
- [6] N. V. Fedorova, N. N. Dzhoieva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogol, and V. S. Tynchenko, “Methods of assessing the efficiency of the foundry industrial marketing,” IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.
- [7] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Ereemeev, “Application of Kohonen self-organizing maps to the analysis of enterprises’ employees certification results,” IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.
- [8] N. O. Ejimabo, “The influence of decision making in organizational leadership and management activities,” Journal of Entrepreneurship and Organization Management, vol. 4, no. 2, pp. 2222–2839, 2015.
- [9] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, “Essence and classification of the agribusiness organizations competitive strategies,” IOP Conference Series: Earth and Environmental Science, vol. 315, no. 2, p. 022106, Aug. 2019.
- [10] V.S. Tynchenko, S.O. Kurashkin, A.V. Murygin, and Y.A. Tynchenko, “Energy distribution modelling in the weld zone for various electron beam current values in COMSOL Multiphysics,” Journal of Physics: Conference Series, vol. 1889, no. 4, p. 042058, May 2021, doi:10.1088/1742-6596/1889/4/042058.
- [11] J. R. Hansen, and E. Ferlie, “Applying strategic management theories in public sector organizations: Developing a typology,” Public Management Review, vol. 18, no. 1, pp. 1–19, 2016.
- [12] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, “Intellectual support system of administrative decisions in the big distributed geoinformation systems,” International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.
- [13] F. David, and F. R. David, “Strategic management: A competitive advantage approach, concepts and cases,” Florence : Pearson–Prentice Hall, 2016.
- [14] L. A. Guerras-Martín, A. Madhok, and A. Montoro-Sánchez, “The evolution of strategic management research: Recent trends and current directions,” BRQ Business Research Quarterly, vol. 17, no 2, pp. 69–76, 2014.
- [15] A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, E. A. Chzhan, and A. S. Mikhalev, “Dynamic simulation of calculating the purchase of equipment on credit,” Journal of Physics: Conference Series, vol. 1333, no. 3, p. 032009, Oct. 2019.
- [16] A. A. Boyko, V. V. Kukartsev, D. V. Ereemeev, V. S. Tynchenko, V. V. Bukhtoyarov, and A. A. Stupina, “Imitation-dynamic model for calculating the efficiency of the financial leverage,” Journal of Physics: Conference Series, vol. 1353, no. 1, p. 012123, Nov. 2019.
- [17] V. S. Tynchenko, A. V. Milov, S. O. Kurashkin, V. E. Petrenko, Ya A Tynchenko, and D. V. Rogova “Mathematical model of the waveguide pipe heating in the process of induction brazing,” IOP Conference Series: Materials Science and Engineering, vol. 1047, no. 1, p. 012112, Feb. 2021, doi:10.1088/1757-899X/1047/1/012112.
- [18] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, (2019, May). “Methods of business processes competitiveness increasing of the rocket and space industry enterprise,” IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042009, May 2019.
- [19] A. O. Stupin, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, A. I. Cherepanov, and A. V. Rozhkova, “Management modelling of the natural resources extraction station by agency modelling means,” Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012196, Nov. 2020.
- [20] V. V. Kukartsev, V. V. Khramkov, N. V. Fedorova, A. V. Rozhkova, V. S., Tynchenko, and K. A. Bashmur, “Features of evaluating the effectiveness of industrial enterprise marketing activities,” IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012081, 2020.
- [21] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, “Storage and analysis of natural resources information in various territories,” Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.

Practical-Oriented Method of Development of Strategy of Development of a Production Enterprise

Sergei Kurashkin

¹Information Control Systems
Department

Reshetnev Siberian State University of
Science and Technology

²Laboratory of Biofuel Compositions
Siberian Federal University

³Digital Material Science: New
Materials and Technologies

Bauman Moscow State Technical
University

Krasnoyarsk, Russia

0000-0002-4017-4369

Vladislav Dmitriev

Department of Information Economic
Systems

Reshetnev Siberian State University of
Science and Technology

Krasnoyarsk, Russia

pilipenko.alesya@mail.ru

Kristina Moiseeva

Information Control Systems
Department

Reshetnev Siberian State University of
Science and Technology

Krasnoyarsk, Russia

kristina2001irba@mail.ru

Alexander Korostelev

Department of System Analysis

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

korostelev@sibsau.ru

Alexander Stashkevich

Department of System Analysis

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

0000-0001-6052-3901

Abstract— The article considers what should be taken into account for the development of the strategic plan of the organization for the production of household appliances and how to develop scientific and methodical approaches to the strategic development of the organization. In this work developed its own approach to scientific and methodical development of strategic development, which includes all the principles of scientific approach. It was revealed that it is necessary to carry out significant analytical work to identify strengths and weaknesses of the organization. The main stages of strategic management were studied, which were applied on the example of the enterprise producing refrigeration equipment. Accordingly, existing types of development strategies of the organization were reviewed and compared to select a suitable strategy. The chosen strategy was tested in practice and it was concluded that the implementation of the chosen strategy will entail the achievement of most of the strategic objectives of the enterprise.

Keywords— Application, methodology for developing, refrigeration equipment, strategic management

I. INTRODUCTION

In order to develop a strategic plan for the development of the organization for the production of household appliances, it is necessary to take into account and develop scientific and methodological approaches to the strategic development of the organization. However, it is likely that there is a way for each organization to develop a methodological approach. The crux of the methodology is that the household product organization cannot develop such a modality on its own. In addition, the organization's staff may not always appreciate the methodical approach as a basis for developing strategic development.

The sequencing of the choice and application of development tools is often only recommendatory, usable and can be neglected. Different methods are used to assess the impact of different planning tools, as mentioned above. There are a number of measures to implement the strategic plan, but the scientific and methodological approach

provides an opportunity to understand the strategic development model.

In this research work developed its own approach to scientific and methodical development of strategic development. It incorporates all the principles of the scientific approach, however, contains only those functions and principles that are necessary for the organization of the production of household products.

The process of developing a strategy based on the definition of vision, mission and purpose is to define the very characteristics of the organization in the future; to conceptualize the developer to the state of the organization that corresponds to those characteristics, projection of the state to the real environment in order to determine actions leading to an ideal result. However, once the vision, mission and objectives have been formulated, it is premature to proceed to the development of the strategy. The strategy cannot be detached from the organization and its reality. Therefore, significant analytical work is needed to identify strengths and weaknesses of the organization, opportunities and threats posed by the external environment, to investigate the problem field and to analyse the strategy of the organization.

II. METHODS AND MATERIALS

The main stages of strategic management are:

- Analysis of the environment.
- Defining the mission and goals of the organization.
- Formation and choice of strategy.
- Implementation of the strategy.
- Evaluation and control of the implementation of the strategy.

The company selected as an example produces household refrigerators, freezers, commercial refrigeration equipment, and medical equipment.

For functioning, the management of the organization needs to plan its activities and develop a methodology for the development of the organization, taking into account the tools. To apply the most appropriate strategy, the firm needs to consider all existing strategies, and finally implement the chosen strategy into the organizational structure.

III. DEVELOPMENT OF LONG-TERM AND SHORT-TERM OBJECTIVES OF THE ORGANIZATION

Long-term goals of the organization are determined for 10-30 years or more, short-term goals for 3-5 years. Sometimes organizations do not distribute goals into short-term and long-term goals, but form the highest-level goals, called strategic ones.

To adopt any of the existing strategies, the enterprise must be considered in terms of assessing factor advantages. In Table I, one can trace the implementation of the strategic development methodology on the example of an enterprise producing refrigeration equipment.

TABLE I. APPROBATION OF THE METHODOLOGY ON THE EXAMPLE OF AN ENTERPRISE PRODUCING REFRIGERATION EQUIPMENT

Factor	Specific gravity (%)	Conclusion	Strategy
1. Consumer Orientation			Concentrated growth
Social	65	The product is social, therefore, is in great demand among consumers.	
Technological	35		
2. Informatization			
Informatization level	75	The level of informatization of a trading company is quite high, which means that it is necessary to provide the end user with accessibility and ease of use of the services of the organization.	
Consumer awareness level	25		
3. Long-term perspective			
Increase in client base	70	By increasing the customer base, the firm may have the opportunity to maximize market share, which in the long term will allow the organization to become a monopolist in the refrigerator market..	
Market share maximization	30		
4. Competition			
Consumers	43	Increasing sales volume will allow the organization to reach new levels of competitive advantage.	
Competitors	37		
Sales volume maximization	20		

The most common types of firm strategies tested by practice reflect four different approaches to the growth of the firm and are associated with a change in the state of one or more elements: product-market, industry, position of the

firm within the industry, technology. Each of the elements can be in one of two states - existing or new.

Figure 1 shows the main strategies for the development of the organization.

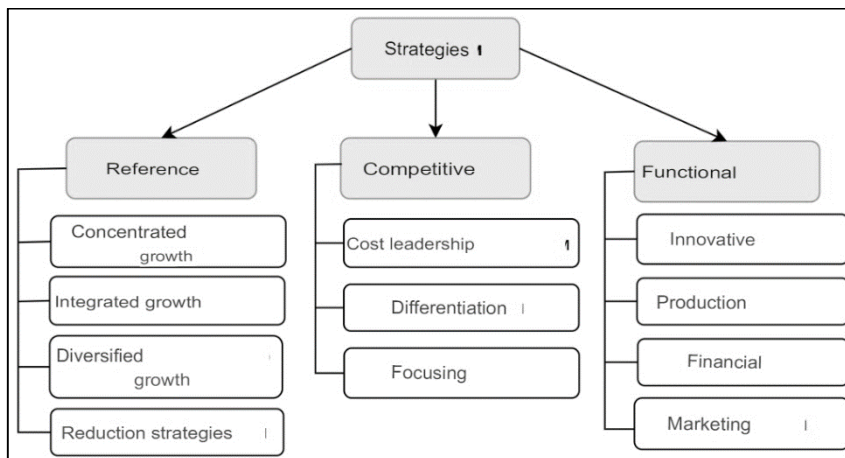


Fig. 1. Types of organization development strategy.

For the successful implementation of the chosen strategy, it is necessary that the goals and plans of the strategy be quickly communicated to the staff in order to achieve from the staff not only an understanding of what the company is doing, but also an informal introduction into the process of implementing strategies, in particular, identifying staff commitments to strategy implementation firm. It is also necessary that top management not only timely carry out the receipt of all important for the implementation of the strategy, but also draw up a plan for the implementation of the strategy in the form of specific standards and record the achievement of each of the goals.

When organizing a business in any field of activity, the essence of global competition strategies is clearly visible when the two components are combined.

First, the unit cost of manufactured products, which depends on:

- Production economies of scale (the cost of production decreases up to a certain volume, then increases).
- The production effect of development (the unit cost decreases as the volume of manufactured products increases).

Secondly, the market for a product with a specific purpose, which can be segmented. At the same time, the capacity of the segment is inversely proportional to the price of the segment's product.

The combination of these facts among themselves makes it possible to have various options for reducing the price of the manufactured product without loss of quality, which makes it possible to successfully compete in a particular sales market.

The most important source of cost minimization is the establishment of a specific value for the volume of production (the effect of scale from production), promotion and marketing (the effect of scale from the marketing strategy).

With a larger load, the wear resistance of the equipment increases, with a smaller load, the cost of the goods increases.

The disadvantage of the cost minimization strategy is the relative contradiction with the differentiation strategy. It partially eliminates the use of flexible commodity production, which requires:

- Low costs of reloading and readjustment of equipment during production.
- Maximum vision of economies of scale from production with small volumes of product creation.

The differentiation strategy is based on the production of a large range of products for one direct purpose. This allows the trading company to maximize the demand for its product.

Differentiation can be divided into two types:

- Horizontal (the price is approximately the same, the average level of income of buyers is similar).
- Vertical (prices and average income of buyers are different).

The application of a differentiation strategy gives success with the superiority of non-price competition, and the industry market has a complex organizational structure.

Disadvantages of Differentiation Strategy:

- Increase in the price of goods due to the need to take advantage of economies of scale.
- High costs of advertising the company's reputation.
- The appearance on the market of a lower-priced product of inadequate quality.
- Strengthening of the price factor in competition.

The focusing strategy contains the focus of the company's activities on the satisfaction of buyers of a smaller segment of buyers, characterized by special needs for this product. The value of this strategy depends on:

- The existence of the largest group of buyers whose requirements for products with special purposes differ from the average.
- The existence of a small group of consumers containing unusual needs that are not satisfied in the proper amount.

- The company's resources do not allow meeting the needs of huge groups of customers with the usual wishes.

Disadvantages of Focusing Strategy:

- Exclusion from a number of differences in the prevailing for buyers of the target segment in the general market.
- Strengthening of differential trends among firms operating in the market.

The innovation strategy involves the creation of new products or technological steps, or the satisfaction of existing tastes with a modern method.

Firms using this strategy create a competitive advantage and the ability to extract more profit by increasing the profitability of the implementation or creating a different circle of buyers. Another way is the implementation of new technologies to other firms in the industry.

The disadvantage of the innovation strategy is a huge part of the risk. A special option for minimizing losses is structured financing based on the results of previous stages, with each subsequent implementation usually higher than the previous one. Such a strategy is most often implemented by small firms that have only two alternatives: lose or win - which is why they are not particularly at risk.

Such a strategy determines the advantage by reacting sharply to changes in the external environment. To provide new buyers of the product market, it is necessary to use the time. Firms using this strategy produce habituation of a new product in a very short time. If, at the same time, this is done an order of magnitude faster than potential competitors, it is provided with an additional percentage of income.

Rapid response strategies are given first place most often by firms engaged in counterfeiting of branded goods from world famous manufacturers.

The portfolio of businesses as a whole is very different from the simple number of businesses that include it. By creating a portfolio of businesses, the company optimizes cash flows, profitability, risks, and more.

The strategy of related diversification is applied by concentrating firms of modern directions in the business portfolio.

The use of diversification in a single portfolio in a business environment can lead to synergistic effects, that is, an increase in the efficiency of the company's functioning as a result of combining, connecting each part into one common system. They affect the reduction of single joint costs and arising from the repeated use of resources. Synergies in strategic management are interpreted by strategic correspondences and are defined as the same cost parameters in the cost structures of other firms that have a single business portfolio that must flow into each other. In strategic management, there are the following strategic certainties:

- Marketing (single buyer, suppliers, same geographic boundaries, distribution channels, advertising capabilities, similar trade logos, after-sales service).
- Production (similar technologies, single production facilities, R&D).

- Managerial (one training and management services, managers).

The diversification strategy can be of two types: connected (assumes the existence of existing strategic alternatives between the business services included in it), unrelated.

Relatedly diversified firms are called concerns. In the usual form, with a general economic decline, it is the concerns that have the maximum chances for success. Reducing costs is the most important task that concerns can easily do better.

Firms that use unrelated diversification are referred to as conglomerates. There are weak strategic uniformities among themselves in their portfolio of business circles.

The advantage of an unrelated diversification strategy comes from risk reduction, i.e. industries of different types can equally appear in different phases of industry life cycles. A decline in some leads to an increase in others. There is a possibility of early reprioritization of services and products that are seasonal in nature as opposed to planned demand.

The well-being of the conglomerate directly depends on the early vision and ability of the management team to profitably command the components of the portfolio of business areas.

The negative qualities of a conglomerate, as mentioned above, are an order of magnitude less effective than reducing costs in an economic crisis or decline.

A change in the external environment or internal changes can lead to the fact that one of the business areas ceases to be profitable, and the best way out may be to stop its functioning. Let's analyze the options for the implementation of this plan:

- Implement the most unattractive organization (in this case, you need to find a company whose business areas contain the greatest strategic fit).
- Liquidation (with this outcome, it is recommended to take into account that the cost of the company's assets sold in parts may be less than their real value).
- Bankruptcy (a method in which it is necessary to come in the very last case, since the business reputation of the entire business may be mired).
- Restructuring and course transformation strategy.

The desire of the corporate management to transform the efficiency of the portfolio, or even to leave its existence unchanged, is a stable motive for the execution of the strategy of changing the course and restructuring.

The following course change strategy options are available:

- Focusing on extracting profitability from declining business areas.
- Measures to increase profitability in all areas of business.
- The introduction of a savings regime in all areas, the sale of weak areas and the acquisition of attractive ones.

- Change of individual corporate level managers.
- Withdrawing resources from weak areas and directing them to more promising ones (harvesting).

The restructuring strategy proposes the elimination of some areas from the portfolio and the inclusion of others.

The strategy is used in such cases:

- The board of directors decides to change the scope of the business.
- The corporation lacks long-term prospects due to the presence in the portfolio of a considerable number of modestly developing or non-competitive and fading business areas.
- Hard times have come for advanced business areas.
- The emergence of new technologies require taking a position in a new promising industry.
- Sale of several business areas in order to acquire a new promising industry.
- Core business areas lose their attractiveness.

The strategy offers the operation of a diversified portfolio in a variety of Russian markets. The strategy received a resonant spread in the mid-70s and 80s, when it became known that "international diversified firms will concentrate an advantage over simple international firms". Such organizations "are able to expose new markets and compete successfully by selling products during a specific period of time at the lowest prices, covering losses with income extracted from already conquered markets".

An *offensive strategy* includes a number of interrelated actions to retain and acquire competitive advantages that dictate the conditions of character. Defensive strategies talk about actions that are in the nature of a response to a challenge.

The main directions of the implementation of the offensive strategy:

Attacking competitor strengths:

- Conquering a part of the market by achieving dominance over the strengths of the most vulnerable competitors.
- Negate the advantages of stronger competitors by holding or reducing prices, using relative advertising within the existing one, as well as endowing the product with qualities that are a priority for competitors' buyers.

It should be noted that the adequacy of the actions used in the company should be distinguished by the margin of stability of the upcoming company.

Attacking competitor's weaknesses:

- Development of areas that the competitor neglects or cannot cope with, concentration of efforts on those products where the competitor's analogues are of low quality business development in geographic entities that are not mastered by competitors.

If an organization has been attacked from outside, it needs to apply a number of measures of a defensive strategy:

- Launch of goods with parameters similar to those of competitors' products.
- Free or inexpensive training of the personnel of the consumer organization in the operation and promotion of the company's goods.
- Active participation in the development and development of new technologies, patenting promising technologies.
- Maintaining low prices.
- Conclusion of exclusive contracts with dealers and distributors.
- Spreading rumors about upcoming product price cuts or new models, which will help reduce the likelihood of consumers switching to competitors' products.
- Extended warranty periods.

The strategy associated with the consolidation of a part of the length of development from the delivery of raw materials to the sale of the finished product to the final buyer is determined by the strategy of vertical integration. Reverse vertical integration refers to the movement back to raw materials, while direct vertical integration is directed forward, that is, to delivery to the buyer of products.

The main method for implementing a vertical integration strategy is to establish a firm's competitive advantage by losing ground to the competitive power of buyers and suppliers.

The current industry advantage is characterized by the volume of the market share. The goal of the winner strategy is to maintain existing positions and assign the status of a leading dominant (significant leadership over other firms).

Three basic strategies need to be analyzed:

- Attack strategy according to the principle "there are only two types of promotion: back and forth" (maintaining innovation, launching new product technologies, high maximization of the increase in the market share);
- A strategy of protection and strengthening, which implies the creation of such factors under which it would be difficult for competitors to maximize the market share at the expense of the leader company (maximization of entry barriers for new companies through the introduction of advertising innovations, maximization of the quality of service, maximization of research costs, access to a new market for products under a private label, etc.);
- A strategy of democratization of force, involving the introduction of the company's business reputation, taking into account the attempts of competitors to restructure the alignment of the parties (when competitors try to enter the market share - the greatest price minimization, hints about punishing any attacking actions, increasing dealer discounts, etc.).

The most common strategies analyzed are ways of the overall strategy of the firm, containing at the moment of delivery a specific advantage.

The choice of strategy is the main point of such strategic management. The choice of strategy is carried out after an in-depth analysis of common new opportunities and external dangers, taking into account them, the entire internal structure of the organization is brought back to normal.

Determining a strategy means taking into account the decision about what to do with a particular business or product, what direction to choose, what niche to occupy in the market. "Social, civic and regulatory, political norms hinder the firm's strategic actions that the organization should or could have foreseen".

Dependence on the external environment, in many cases, can play the biggest role in choosing a company's strategy, compared to other factors. Rigid dependence on the external environment can be determined by legal norms, the state of the economy in the country and abroad, the conditions of interaction with the natural environment, and the like.

IV. BUSINESS PHILOSOPHY, PERSONAL AMBITION, AND ETHICAL VISIONS OF TOP MANAGEMENT

Management decisions are often influenced by personal views on how to compete, what niche the organization should occupy, and what image it should apply. It is rare that such an influence on strategy occurs simply subconsciously. Strengths and weaknesses of the company, the internal structure of the company and its competitive capabilities.

The way in which the proposed strategies are chosen should be what the firm is particularly comfortable doing. It is not recommended to create strategies based on little known or unfamiliar activities. The firm's strategy must be based on what it does with great success, that is, on competitive capabilities and organizational strength. A huge danger to strategy leadership is to focus on what the firm does not do very well, i.e. competitive and organizational weakness.

V. RESULTS AND DISCUSSION

When creating a strategy, it is recommended to take into account both negative and positive results of previous strategies. In addition, in connection with the transition to new strategies, it is impossible to completely renounce all previous commitments. Therefore, it is recommended to take into account the fact that for some time there will be obligations of previous years, which will naturally correct or constrain the ability of implementation strategies.

The firm will not implement the strategy at every moment and not at any time, but only then and at such time as it becomes possible to do so. The company that is better at keeping track of time and, accordingly, more intelligently able to manage processes over time, always succeeds in implementing the strategy and, as a rule, leadership in the competition.

The initial process in developing strategic alternatives is strategic allocation. In this application, analysis is carried out, involving the study of the external environment of the organization for the most precise dangers and the emergence of new opportunities. The object of such an analysis is a strategic business zone - a separate segment of the external environment in which the company works or wants to get out.

The strategic choice should be based on a clear concept of the company's development, since the chosen strategy for a long period of time limits the freedom of action of management and affects deep management decisions.

The relationship of strategic goals forms a strategic map. Figure 2 shows the cause-and-effect relationships that lead to the implementation of higher strategic goals.

It can be concluded that the implementation of the chosen strategy will lead to the achievement of most of the strategic goals of the enterprise.

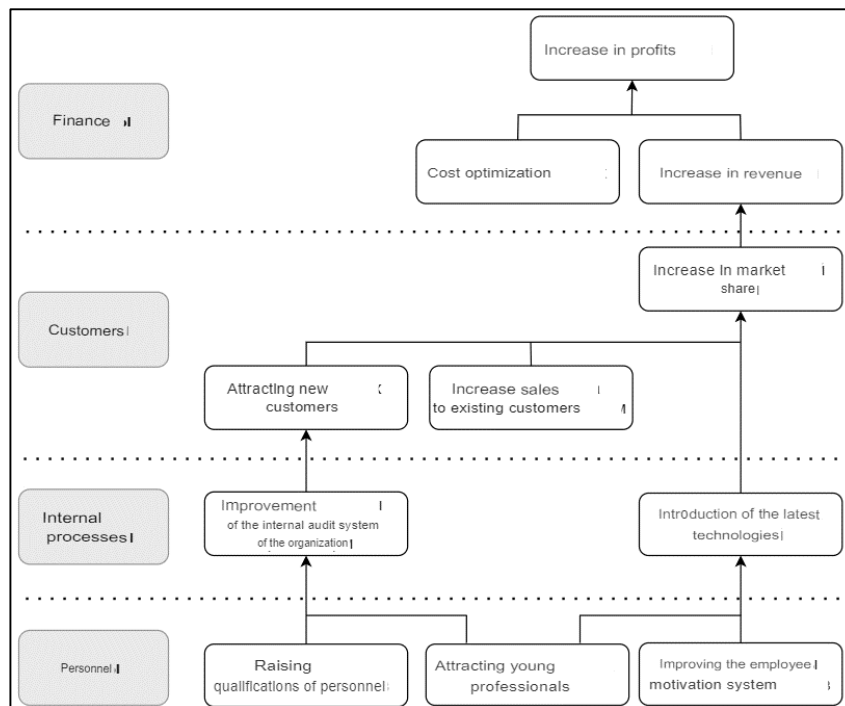


Fig. 2. Enterprise strategic map.

CONCLUSION

The developed method of the development strategy of the organization was applied in practice in the organization for the manufacture of refrigeration equipment and can be applied in the enterprises of the industry, which will allow their managers to apply the developed methodology and plan further development of the organization, bypassing the sharp corners. The developed methodology will help to increase the efficiency of activities, including a more rational approach to strategic development and promotion in this market, taking a competitive advantage.

REFERENCES

- [1] D. S. Shalaeva, O. I. Kukartseva, V. S. Tynchenko, V. V. Kukartsev, S. V. Aponasenko, and E. V. Stepanova, "Analysis of the development of global energy production and consumption by fuel type in various regions of the world," IOP Conference Series: Materials Science and Engineering, vol. 952, no. 1, p. 012025, Nov. 2020, doi: 10.1088/1757-899X/952/1/012025.
- [2] R. J. Baumgartner, and R. Rauter, "Strategic perspectives of corporate sustainability management to develop a sustainable organization," Journal of Cleaner Production, vol. 140, pp. 81–92, 2017.
- [3] P. Jarzabkowski, and S. Kaplan, "Strategy tools-in-use: A framework for understanding "technologies of rationality" in practice", Strategic management journal, vol 36, no. 4, pp. 537–558, 2015.
- [4] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.
- [5] N. V. Fedorova, N. N. Dzhiyeva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogol, and V. S. Tynchenko, "Methods of assessing the efficiency of the foundry industrial marketing," IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.
- [6] V. S. Tynchenko, A. V. Milov, S. O. Kurashkin, V. E. Petrenko, Ya A Tynchenko, and D. V. Rogova "Mathematical model of the waveguide pipe heating in the process of induction brazing," IOP Conference Series: Materials Science and Engineering, vol. 1047, no. 1, p. 012112, Feb. 2021, doi:10.1088/1757-899X/1047/1/012112.
- [7] N. O. Ejimabo, "The influence of decision making in organizational leadership and management activities," Journal of Entrepreneurship and Organization Management, vol. 4, no. 2, pp. 2222–2839, 2015.
- [8] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 2, p. 022106, Aug. 2019.
- [9] V.S. Tynchenko, S.O. Kurashkin, A.V. Murygin, and Y.A. Tynchenko, "Energy distribution modelling in the weld zone for various electron beam current values in COMSOL Multiphysics," Journal of Physics: Conference Series, vol. 1889, no. 4, p. 042058, May 2021, doi:10.1088/1742-6596/1889/4/042058.
- [10] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Ereemeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.
- [11] J. R. Hansen, and E. Ferlie, "Applying strategic management theories in public sector organizations: Developing a typology," Public Management Review, vol. 18, no. 1, pp. 1–19, 2016.
- [12] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, "Intellectual support system of administrative decisions in the big distributed geoinformation systems," International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.

- [13] F. David, and F. R. David, "Strategic management: A competitive advantage approach, concepts and cases," Florence : Pearson–Prentice Hall, 2016.
- [14] L. A. Guerras-Martín, A. Madhok, and A. Montoro-Sánchez, "The evolution of strategic management research: Recent trends and current directions," BRQ Business Research Quarterly, vol. 17, no 2, pp. 69–76, 2014.
- [15] A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, E. A. Chzhan, and A. S. Mikhalev, "Dynamic simulation of calculating the purchase of equipment on credit," Journal of Physics: Conference Series, vol. 1333, no. 3, p. 032009, Oct. 2019.
- [16] A. A. Boyko, V. V. Kukartsev, D. V. Ereemeev, V. S. Tynchenko, V. V. Bukhtoyarov, and A. A. Stupina, "Imitation-dynamic model for calculating the efficiency of the financial leverage," Journal of Physics: Conference Series, vol. 1353, no. 1, p. 012123, Nov. 2019.
- [17] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, (2019, May). "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042009, May 2019.
- [18] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," International Russian Automation Conference (RusAutoCon), IEEE, pp. 1-6, Sep. 2018.
- [19] A. O. Stupin, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, A. I. Cherepanov, and A. V. Rozhkova, "Management modelling of the natural resources extraction station by agency modelling means," Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012196, Nov. 2020.
- [20] V. V. Kukartsev, V. V. Khramkov, N. V. Fedorova, A. V. Rozhkova, V. S., Tynchenko, and K. A. Bashmur, "Features of evaluating the effectiveness of industrial enterprise marketing activities," IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012081, 2020.
- [21] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, "Storage and analysis of natural resources information in various territories," Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.

Road Map “TechNet” National Technological Platform

Alena Stupina

¹Department of System Analysis
Reshetnev Siberian State University of
Science and Technology

²Department of Digital Management
Technologies
Siberian Federal University

³Department of Civil Defence and
Emergency Management
Siberian Fire and Rescue Academy of
the Russian Ministry of Emergency
Situations

Krasnoyarsk, Russian
0000-0003-2557-6316

Natalia Fedorova

¹Department of management
Reshetnev Siberian State University of
Science and Technology

²Department of Advertising and Social
and Cultural Activities
Siberian Federal University

Krasnoyarsk, Russia
nvfed@mail.ru

Yuriy Danilchenko

Department of management
Reshetnev Siberian State University of
Science and Technology

Krasnoyarsk, Russia
ydanilchenko@sibsau.ru

Dmitriy Ereemeev

Department of Accounting, Finance
and Economic Security

Reshetnev Siberian State University of
Science and Technology

Krasnoyarsk, Russia
eremeev.dmitriy@gmail.com

Elena Vaitekunene

¹Department of Information Economic
Systems

Reshetnev Siberian State University of
Science and Technology

²Department of Digital Management
Technologies

Siberian Federal University

³Department of Civil Defence and
Emergency Management
Siberian Fire and Rescue Academy of
the Russian Ministry of Emergency
Situations

Krasnoyarsk, Russia
0000-0001-6839-6716

Yuriy Seregin

Information Control Systems
Department

Reshetnev Siberian State University of
Science and Technology

Krasnoyarsk, Russia
0000-0003-4309-8637

Abstract—In the article, for a visual representation of the scientific and methodological approach, as a process of strategic development, a model is used that together can solve all the tasks of an organization for the production of household appliances at the stage of strategic development. The roadmap “TechNet” of the national technology platform, the purpose of its creation and the tasks that it performs have been studied. From the whole variety of strategic alternatives, the choice of the most preferred option was made. It was found that the choice of strategy is influenced by many factors that were considered in the work. It was concluded that the information contained in the roadmap is of a reference nature, cannot be considered as a description of directly implemented activities and is not the basis for the allocation of state support funds in any form.

Keywords—component, formatting, style, styling, insert

I. INTRODUCTION

The central place in the economy is occupied by the sphere of material production - a high-tech industry that must meet the requirements of global competitiveness, efficiency and high labor productivity. To meet these requirements, total digitalization, automation and intellectualization of industry are rapidly developing in the world, a transition to cyber-physical systems is underway, and the material and digital (virtual) worlds are merging. These global changes are accompanied by the development of fundamentally new business processes at all levels [1]-

[4].

With the aim of developing Russia in accordance with global technological trends and for the most complete realization of the opening opportunities for increasing the competitiveness of the Russian industry and the final products being created, the “TechNet” roadmap of the National Technology Initiative (NTI) was developed [5]-[8].

II. METHODS AND MATERIALS

For a visual representation of the scientific and methodological approach, as a process of strategic development, a model was used that together can solve all the tasks of an organization for the production of household appliances at the stage of strategic development. In addition, the specificity of the scientific and methodological approach lies in the unified process of applying the strategic development tools necessary for the successful functioning and development of an organization for the production of household appliances. For a detailed analysis shown in Figure 1, it is necessary to understand that this model contains only advisory character [9]-[12].

The figure shows that the use of one or another factor involves the use of a set of tools offered for selection. For each stage of the formation of a strategic plan, it is necessary to take into account the degree of influence of the factor on the environment of the trading company and its future prospects.

From the whole variety of strategic alternatives, the most preferred option is directly selected. The very process of

The work was carried out within the framework of the state support program for leading scientific schools (grant of the President of the Russian Federation NSH-421.2022.4).

strategic choice takes place not only on a rational level, but also on an irrational one - according to intuition, experience, and the ability to foresee the situation. It has been established that the choice of strategy is influenced by many factors. The most important of them [13]-[16]:

- Type of activity and features of the industry in which the trading company operates.
- The nature of the goals that the company sets for itself.
- Values that guide decision-making by top managers.

- Financial resources and obligations of the trading company for already made decisions.
- The state of the external environment.
- Types and degree of risks.
- Entrepreneurial abilities.
- Degree of dependence on the environment.
- Time factor.

Figure 2 shows the priorities of the scientific and technological development of Russia until 2035.

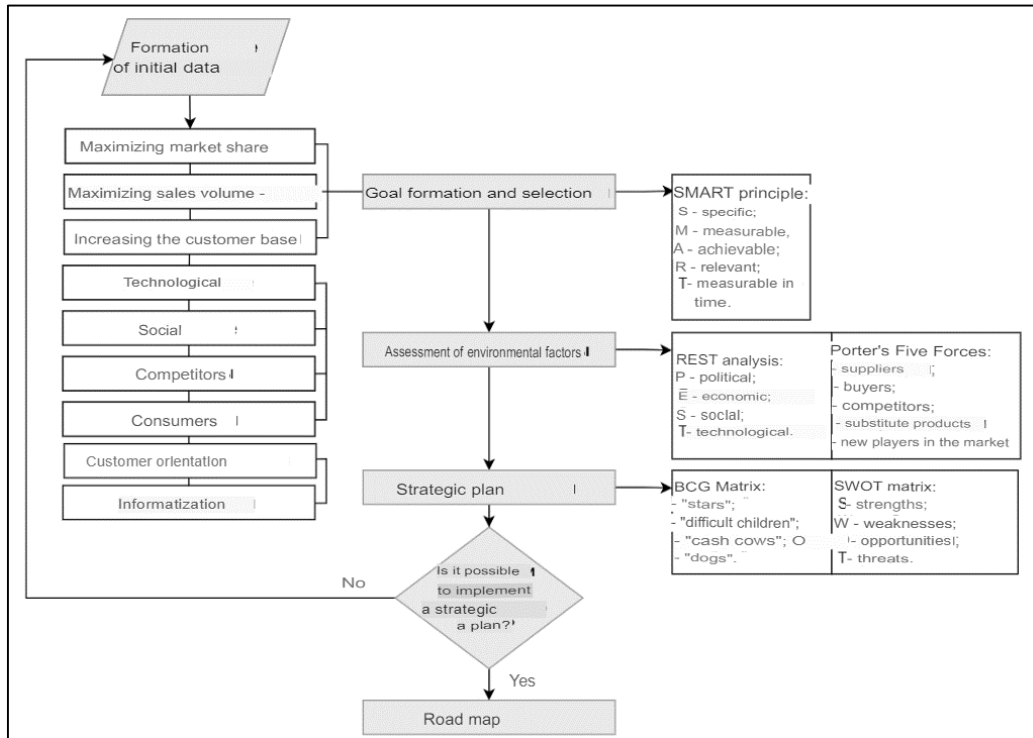


Fig. 1. Strategy Development Algorithm.

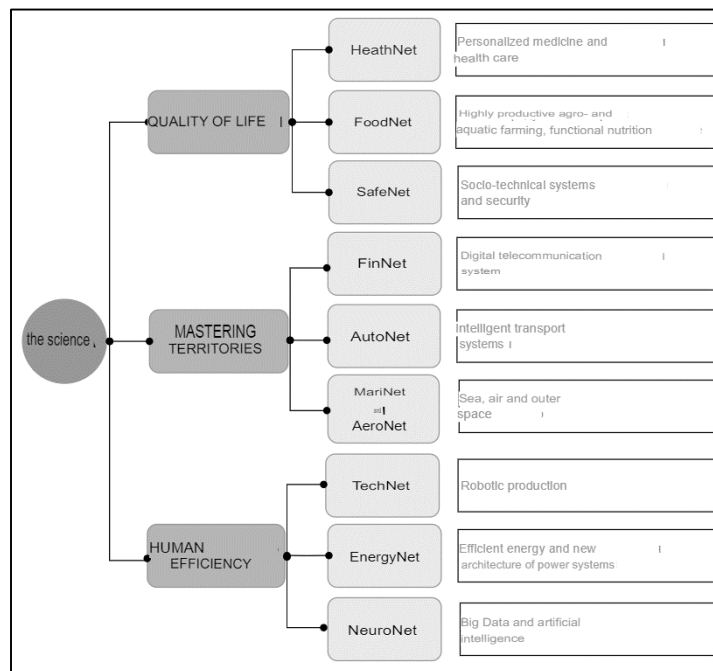


Fig. 2. Strategy Development Algorithm.

Formed strategies are evaluated according to the level of suitability for achieving the main goals of the enterprise and their compliance with the requirements of the environment, as well as the possibilities for the development of the organization.

“TechNet” is an NTI action plan for the development of a cross-market, cross-industry direction “Advanced Manufacturing Technologies” (AMT), which will ensure the competitiveness of domestic companies in the NTI markets and in high-tech industries.

“TechNet” is the first NTI roadmap developed for the development and effective application of “end-to-end technologies”, first of all, new production technologies.

The “TechNet” roadmap covers such advanced manufacturing technologies as digital design and modeling,

new materials, additive and hybrid technologies, robotics, industrial sensors, industrial Internet, big data, information systems for production and enterprise management, virtual and augmented reality technologies, artificial intelligence.

The key importance in the “TechNet” roadmap is given to the formation of “Factories of the Future” (digital, “smart”, virtual factories), which are systems of integrated technological solutions that provide the design and production of globally competitive next-generation products in the shortest possible time.

With successful formation, a strategic plan follows, in this case a strategic map of the development of the *AutoNet* market.

The roadmap is shown in Figure 3, which clearly depicts the essence of market development.

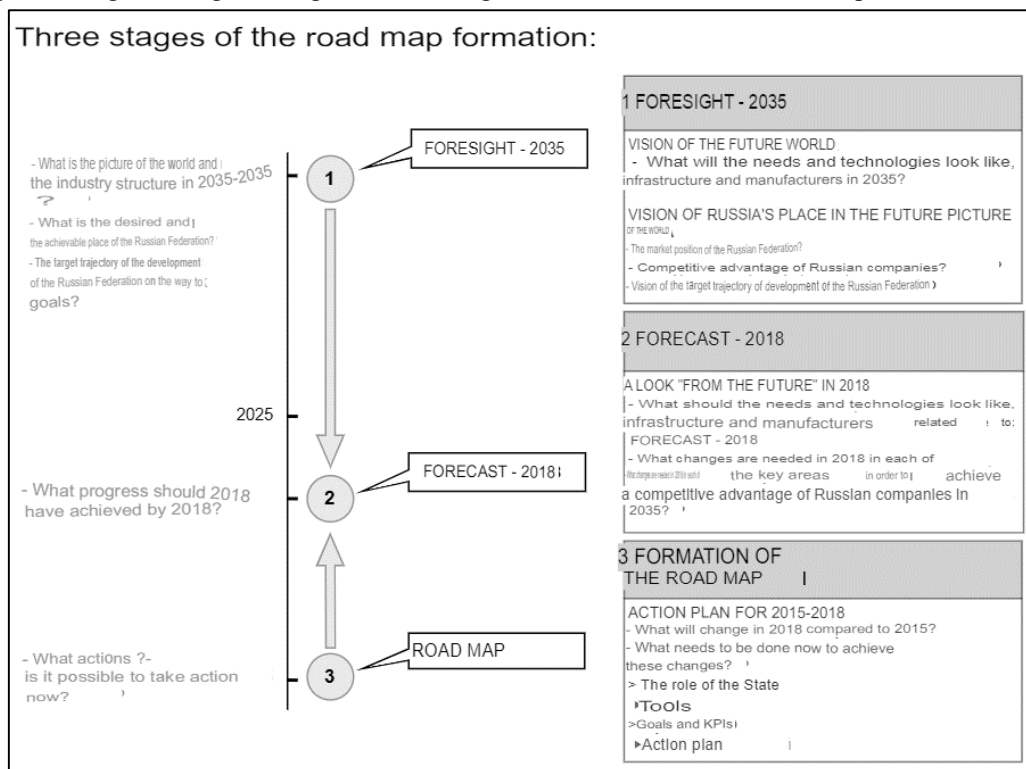


Fig. 3. *AutoNet* Market Development Roadmap.

Thus, the scientific and methodological approach is a kind of sequence of applying strategic tools in strategic development.

When choosing, you must adhere to certain criteria:

- 1) Compliance with the opportunities and threats of the external environment.
- 2) Compliance with the goals of the enterprise and compatibility with its mission.
- 3) Achieving competitive advantages:
 - a) Exploiting the strengths of the enterprise and the weaknesses of competitors.
 - b) Neutralization or compensation of the weaknesses of the trading company and the strengths (advantages) of competitors.
- 4) Strategy feasibility:
 - a) Availability of necessary resources.

- b) Compatibility of strategy with internal organization.
- c) Possible consequences.

III. METHODOLOGY FOR DEVELOPING A DEVELOPMENT STRATEGY FOR THE ORGANIZATION OF HOUSEHOLD APPLIANCES IN THE “TECHNET” MARKET

The implementation of the adaptation of a trading company to market conditions as a management function is necessary in order to bring the processes of the internal environment of the enterprise into line with the target plans or action programs of the organization in relation to the external environment, i.e., adapt the production and marketing activities of the organization to market conditions.

Applying the theory, the management of a trading company should build its activities in accordance with its key principle: "Produce what is sold, and not sell what is produced." The basis for making a decision on expanding or

reducing production volumes, upgrading products or removing them from production is marketing data obtained as a result of market research and the internal potential of the enterprise.

An analysis of the development of the Russian household appliances market made it possible to identify a number of trends in its development [17]:

- Dependence on import supplies.
- Increase in the number of Western enterprises operating in Russia.
- Slowdown in sales growth in the household appliances and electronics market.
- The economic downturn and the decrease in consumer activity will affect the dynamics of the home appliances market, and a slowdown in sales growth in the Russian home appliances market as a whole is predicted.
- Further decrease in profit growth rates of trading companies in this market.
- Switching consumer preferences to cheaper models in all categories of household appliances.

- Further strengthening of competition in all segments of the household appliances market.
- Transition from price competition to non-price.

To further change the situation, major players have developed and are implementing anti-crisis measures [18]:

- Optimization of the assortment, increase in the assortment of the share of the mid-price segment, etc.
- There is an increase in sales of household appliances through the Internet – trade.
- Further promotion of large retail chains of household appliances to regional markets.
- Further growth of basic and additional services.
- Growth in sales of household appliances through mobile applications.
- Further development of multi-channel sales.

The sequence of interrelated work on strategic analysis, selection and implementation of the strategy is the process of strategic management, which is shown in Figure 4.

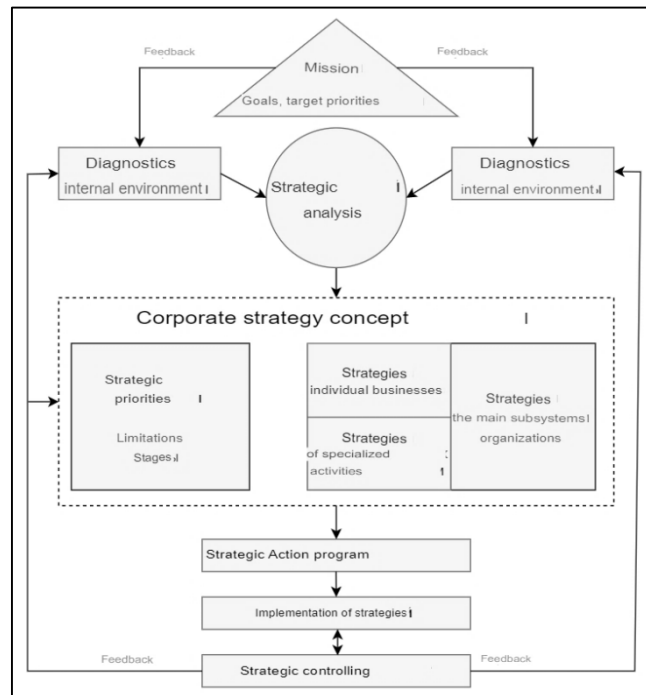


Fig. 4. AutoNet Market Development Roadmap.

Forecast for the development of the Russian household appliances market for 2019 - 2022:

- The economic downturn and the decrease in consumer activity will affect the dynamics of the home appliances market, and a slowdown in sales growth in the Russian home appliances market as a whole is predicted.
- Further decrease in profit growth rates of trading companies in this market.
- Switching consumer preferences to cheaper models in all categories of household appliances.

- Further strengthening of competition in all segments of the household appliances market.

As can be seen from the diagram, the strategy development process is iterative (cyclic). Thus, the definition and selection of a strategy can take place at the stage of analysis of the external environment, and the evaluation of the strategy will require additional external analysis. In addition, a change in strategy leads to the need to monitor and annually adjust strategic decisions and plans.

In the light of the new conditions for the functioning of a production organization, determined by the "rules of the game" in the system of market relations, the organizational

principles and functions of the activities of the upper and middle levels of the management structure should be revised and refined.

Each production organization must develop and consistently implement a program of measures to restructure its capabilities in order to most fully adapt to the new business conditions.

The TechNet roadmap has goals such as [19]:

- Formation of a set of key competencies in Russia that ensure the integration of advanced production technologies and new business models for their distribution as “Digital Factories”.
- Creation of globally competitive next-generation personalized products for NTI markets and high-tech industries.

The purpose of this map is to:

- Creation of infrastructure for the development of a set of key competencies for Digital Factories.
- Implementation of a set of key competencies by creating globally competitive companies in the NTI markets and in high-tech industries.
- Long-term planning for the development of advanced manufacturing technologies and related business models.
- Formation of an ecosystem for creating, attracting, developing and transferring best-in-class technologies.
- Creation of legislative and institutional conditions for the development of advanced production technologies.

IV. RESULTS AND DISCUSSION

Advanced production technologies in accordance with the NTI concept include: digital design and modeling, including supercomputer engineering, new materials (composite materials, metamaterials, metal powders), additive and hybrid technologies, flexible production cells (robotic complexes), various sensors, industrial Internet, big data, virtual and augmented reality technologies, expert systems and artificial intelligence.

A key role in the digital economy will be played by digital factories - this is a certain type of business process system, a way of combining business processes, which has the following characteristics:

- Creation of digital platforms, original ecosystems of advanced digital technologies.
- Development of a system of digital models of both new designed products and production processes. Digital models should have a high level of adequacy to real objects and real processes.
- Digitalization of the entire life cycle of products, everyone understands that the cost of changes is

greater the later we make these changes, and therefore world practice shows that the “center of gravity” is shifting towards design processes, within which the characteristics of global competitiveness or high consumer demands.

At the stage of formation of a digital factory, new key competencies are formed, for example:

- Quick personalization of response to market or customer requests.
- Use of system approaches (“system engineering”).
- Formation of a multi-level matrix of targets and restrictions as the basis of a new design, which significantly reduces the risks, volumes of full-scale tests and the amount of work associated with “finishing products and products based on tests”.
- Development and validation (proof that the requirements of a particular user, product, service or system are satisfied) of mathematical models with a high level of adequacy to real objects and real processes.
- Change management throughout the life cycle.
- “Digital certification” based on thousands of virtual tests of both individual components and the entire system as a whole.

The main directions of the “TechNet” roadmap, in terms of creating advanced Russian technologies for industry [20]:

- Creation and implementation of “virtual machines” (a full-fledged 3D model of the machine, which makes it possible to conduct virtual tests of future products, including those with the ability to connect to real controllers) and real-time collaboration functionality.
- Creation of a single information space of the enterprise, which will allow resolving the contradiction between the unity of the production and administrative activities of the enterprise and the autonomy of management at the level of individual systems.
- Release of the Industrial Internet. The strategy includes cross-platform standardization of open systems for secure data exchange, interoperability, as well as optimization of user experience for customers, implemented in the form of embedded intelligent models. This approach also provides comprehensive automation and informatization of customer production based on a centralized approach and a single integrated architecture.
- Integrated automation of the entire chain of interacting production enterprises, the exclusion of a person and risks associated with the human factor from the industrial production process.

Figure 5 presents the overall picture of the “factory of the future” markets for 2015-2035 years.

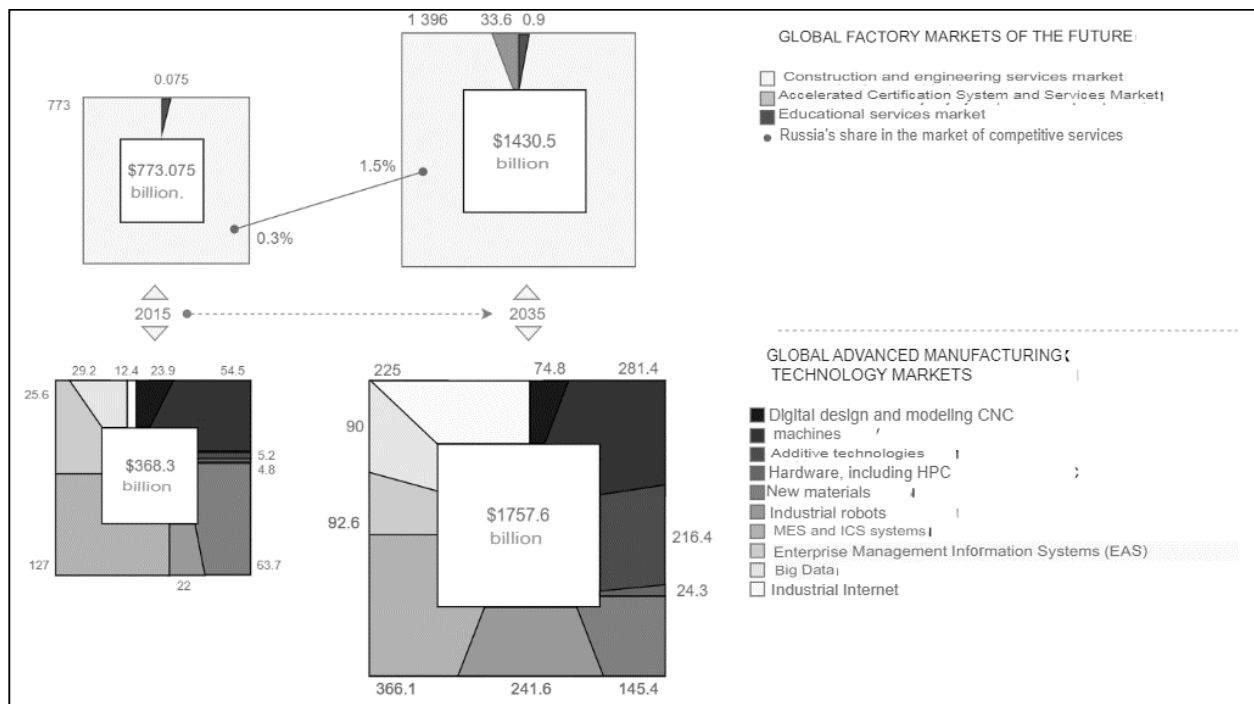


Fig. 5. "Future Factory" Markets.

The diagrams show the growing role of industry, but on fundamentally new solutions, which are now united by the general term "smart factory". Behind this is the end-to-end integration of processes, the total digitalization of workflow and management, the distribution of design functions in space, production quality control, streaming individualization.

The distinguishing features of "smart" industries are the following [21]:

- The ability to act smart and respond smartly, maximizing technical efficiency, cost effectiveness and benefit through planning, continuous monitoring of operations and continuous learning.
- "Operational assets" - workers, plant, equipment, operating models and databases - are integrated and aware of their status through a system of sensors. Each device is able to determine its state and report it to all devices associated with it.
- Smart manufacturing equipment is able to detect abnormal situations and adapt to them. Through constant monitoring and application of the acquired knowledge, the system has the ability to function adequately depending on changing circumstances such as sudden interruption of work processes, changes in the properties of the resulting raw materials, etc.
- The equipment has full access to the necessary information at any time of operation.
- To prevent accidents, smart manufacturing collects real-time information.
- The system has the ability to quickly respond to changes in the process and failures.
- Smart manufacturing is sustainable, recyclable and has minimal environmental impact.

- A necessary feature of smart manufacturing is a highly skilled workforce.
- The system has an understanding of the limits of automatic action and supplies all the necessary information to operators and managers to make the necessary decisions.

The main tasks for the enterprise for the production of household products [22], [23]:

- Flexibility of production: the possibility of restructuring production processes for new types of products;
- Decentralization of management as a mechanism to ensure the required flexibility of production: distribution of decision-making functionality throughout the production system;
- Evolutionary gradual transition to the enterprise of the future: no need to stop production to make changes;
- Service approach as a model of inter-corporate interactions that determines organizational and financial relationships;
- Economic analysis of the possibilities of manufacturing new products in real time with setting the task for partner enterprises;
- Custom computer design, modeling, testing, certification of products in digital format in real time.

CONCLUSION

Thus, it is obvious that the use of marketing information contributes to the development and implementation of enterprise development plans, as evidenced by the experience of effectively operating companies, including some Russian ones. However, the problems of collecting and analyzing marketing data, and most importantly, their

productive application are far from simple for most Russian managers and require professional approaches to their solution.

The implementation of the roadmap is carried out in the form of projects that have been selected in the prescribed manner. The information contained in the roadmap is for reference only, cannot be considered as a description of directly implemented activities and is not the basis for the allocation of state support funds in any form.

REFERENCES

- [1] V. V. Makarov, Y. B. Frolov, I. S. Parshina, and M. V. Ushakova, "MES Systems as an Integral Part of Digital Production. 13th International Conference "Management of large-scale system development"(MLSD), pp. 1–5, Sep. 2020, IEEE.
- [2] J. Baumgartner, and R. Rauter, "Strategic perspectives of corporate sustainability management to develop a sustainable organization," *Journal of Cleaner Production*, vol. 140, pp. 81-92, Jan. 2017.
- [3] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.
- [4] V. Serebrenny, D. Lapin, A. Lapina, "The Concept of Perspective Flexible Manufacturing System for a Collaborative Technological Cells," *Transactions on Engineering Technologies*, Springer, Singapore, pp. 233–244, 2021.
- [5] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, Y. V. Danilchenko, S. N. Ezhemanskaya, and N. V. Sokolovskiy, "Methodology for the formation of indicators balanced system for marketing activities of an industrial enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012084, Jan. 2020.
- [6] A. A. Dorofeeva, L. B. Nyurenberger, "Trends in digitalization of education and training for industry 4.0 in the Russian Federation," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042070, 2019.
- [7] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.
- [8] P. Jarzabkowski, S. Kaplan, "Strategy tools-in-use: A framework for understanding "technologies of rationality" in practice," *Strategic management journal*, vol. 36, no. 4, pp. 537–558, 2015.
- [9] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 2, p. 022106, Aug. 2019.
- [10] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.
- [11] N. O. Ejimabo, "The influence of decision making in organizational leadership and management activities," *Journal of Entrepreneurship and Organization Management*, vol. 4, no. 2, pp. 2222–2839, 2015.
- [12] S. Kurashkin, D. Rogova, V. Tynchenko, V. Petrenko, and A. Milov, (2020). "Modeling of Product Heating at the Stage of Beam Input in the Process of Electron Beam Welding Using the COMSOL Multiphysics System," in: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) *Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing*, vol 1294, https://doi.org/10.1007/978-3-030-63322-6_77.
- [13] A. A. Boyko, V. V. Kukartsev, K. Y. Lobkov, and A. A. Stupina, "Strategic planning toolset for reproduction of machinebuilding engines and equipment" *Journal of Physics: Conference Series*, vol. 1015, no. 4, p. 042006, May 2018.
- [14] J. Rosenberg Hansen, and E. Ferlie, "Applying strategic management theories in public sector organizations: Developing a typology," *Public Management Review*, vol. 18, no. 1, pp. 1-19, 2016.
- [15] V. S. Tynchenko, N. V. Fedorova, V. V. Kukartsev, A. A. Boyko, A. A. Stupina, and Y. V. Danilchenko, "Methods of developing a competitive strategy of the agricultural enterprise," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 2, p. 022105, Aug. 2019.
- [16] N. V. Fedorova, N. N. Dzhioeva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogol, and V. S. Tynchenko, "Methods of assessing the efficiency of the foundry industrial marketing," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.
- [17] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," *Journal of Physics: Conference Series*, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.
- [18] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Ereemeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.
- [19] O. Antamoshkin, V. Kukarev, A. Pupkov, and R. Tsarev, "Intellectual support system of administrative decisions in the big distributed geoinformation systems," *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.
- [20] A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, E. A. Chzhan, and A. S. Mikhalev, "Dynamic simulation of calculating the purchase of equipment on credit," *Journal of Physics: Conference Series*, vol. 1333, no. 3, p. 032009, Oct. 2019.
- [21] D. S. Shalaeva, O. I. Kukartseva, V. S. Tynchenko, V. V. Kukartsev, S. V. Aponasenko, and E. V. Stepanova, "Analysis of the development of global energy production and consumption by fuel type in various regions of the world," *IOP Conference Series: Materials Science and Engineering*, vol. 952, no. 1, p. 012025, Nov. 2020, doi: 10.1088/1757-899X/952/1/012025.
- [22] V. V. Kukartsev, V. V. Khramkov, N. V. Fedorova, A. V. Rozhkova, V. S., Tynchenko, and K. A. Bashmur, "Features of evaluating the effectiveness of industrial enterprise marketing activities," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012081, 2020.
- [23] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, "Storage and analysis of natural resources information in various territories," *Journal of Physics: Conference Series*, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.

Analysis of Data in Solving the Problem of Reducing the Accident Rate Through the Use of Special Means on Public Roads

Vladislav Kukartsev

¹Department of Information Economic Systems

Reshetnev Siberian State University of Science and Technology

²Department of Informatics Siberian Federal University

³Digital Material Science: New Materials and Technologie

Bauman Moscow State Technical University

Krasnoyarsk, Russia
0000-0001-6382-1736

Anton Mikhalev

Department of Informatics Siberian Federal University

Krasnoyarsk, Russia
0000-0002-8986-5953

Alexander Stashkevich

Department of System Analysis Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia
0000-0001-6052-3901

Kristina Moiseeva

Information Control Systems Department

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

kristina2001irba@mail.ru

Igor Kauts

Department of System Analysis

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

edu@kernitskiy.ru

Abstract—The purpose of this study is to conduct a statistical analysis of road accidents in the Russian Federation. The article considers ways to reduce accidents and are effective and efficient, as well as statistics given in the period from 1991 to 2019. The main causes of accidents have been identified, which is an essential aspect for reducing the accident rate. The analysis of statistical data on the number of accidents, population, length of roads, number of vehicles other than motorcycles and the general road situation available at the moment and several years ago, as well as the main causes of accidents and measures to reduce accidents. This analysis has led to correct and effective results and solutions to improve the situation on the roads. At the same time, the integrated measures implemented and being implemented at the moment, such as federal targeted, regional and local programs, as well as road safety strategies, are more likely to reduce road accidents. Statistical data and methods used are taken into account, the effectiveness of which is decreasing. Development of methods appropriate to the current situation is a necessity.

Keywords— Public roads, road accidents, statistical data, traffic

I. INTRODUCTION

The topic of accidents has always been a topic deserving special attention in the field of transport, and the accident itself has been and is one of the most important social, economic and dangerous problems in most countries of the world. The Federal Road Agency of the Russian Federation studies sources, main types of road accidents, considers and takes measures to prevent and reduce the number of road accidents [1]-[3].

The total number of road accidents has now declined slightly compared to the late 20th century, as the population has fallen. Thus, in 1992 the number of road accidents in Russia was 185 thousand, with a population of 148 million, in 2007 - 233 thousand, accidents with a population of 143 million, and in 2019 - 164 thousand, accidents with 145 million people [4]-[6].

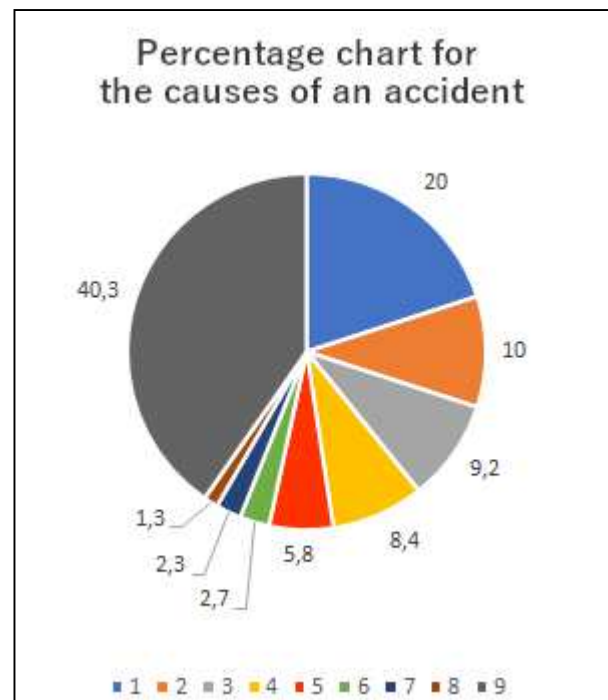


Fig. 1. The main causes of road accidents in the Russian Federation [1], where: 1 - non-observance of the traffic sequence at the intersection; 2 - incorrect distance and non-compliance with the norms of the distance between vehicles; 3 - traffic violation at a pedestrian crossing; 4 - exit to the oncoming lane; 5 - violation of the speed limit; 6 - movement is not in accordance with the traffic lights; 7 - excess of the set speed; 8 - incorrect overtaking; 9 - other minor factors.

The existing and applied sets of measures are for the most part effective, but it is not possible to assess their economic and non-economic effectiveness at this time. Also, the global complexity in the development and application of emergency measures is that all of the factors of accidents are consequences of wrong decisions of a person - i.e. a human

factor. While an undeniably smaller proportion of the causes of accidents, but still important, are amenable to control by the Federal Road Agency and local municipal authorities [10]-[12].

II. ANALYSIS OF TRAFFIC SAFETY STATISTICAL DATA

To assess the level of road safety in Russia, one should consider the accumulated statistical data for a certain number of years. At the same time, there is a need to study not only accident rates, but also information about the size of the vehicle fleet, the length of the road network, as well as the population (Figures 2-4).

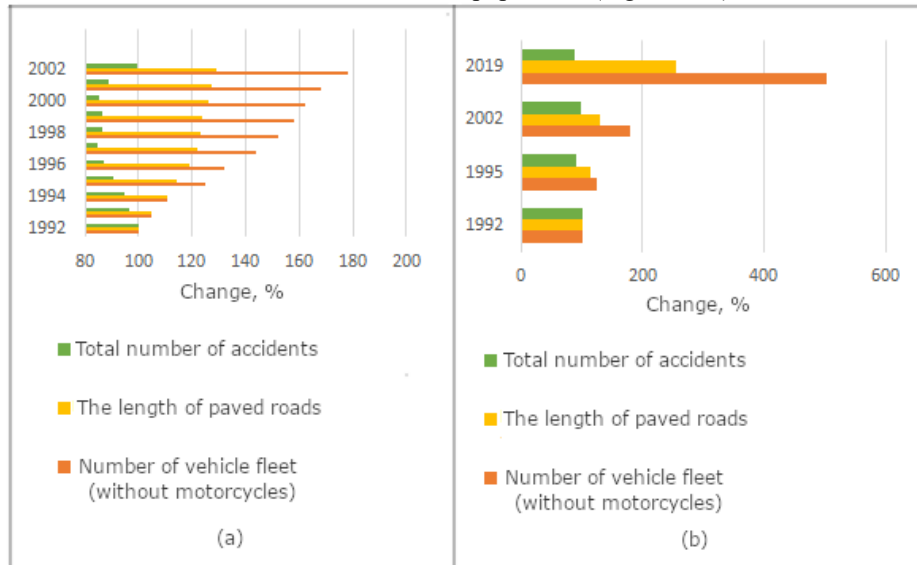


Fig. 2. Traffic safety data analysis graphs, where: a - Dynamics of the relative change in the number of motor vehicles, excluding motorcycles, the length of paved public roads and the total number of accidents in Russia from 1992 to 2002; b - Comparative chart 1992 - 2019.

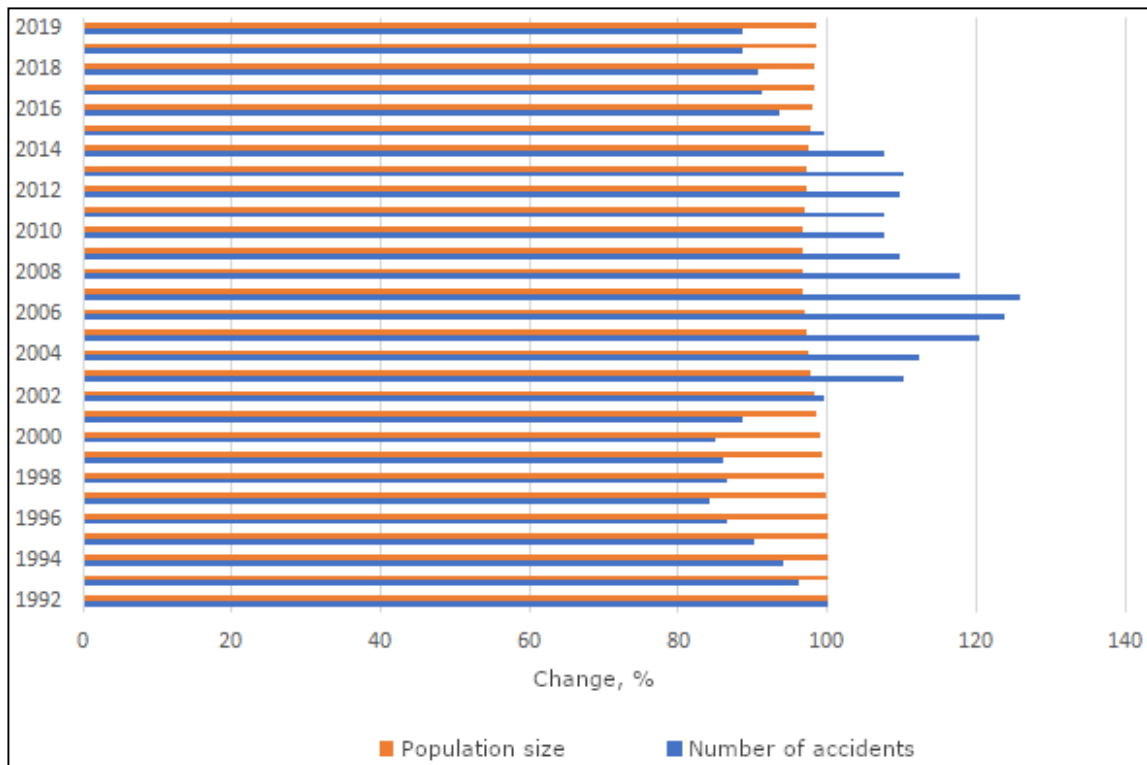


Fig. 3. Comparison of the population and the number of road traffic accidents in Russia for 1991-2019 *Indicators in Figures 2 and 3 in 1992 are taken as 100%. Further, the percentage change is calculated relative to each indicator.

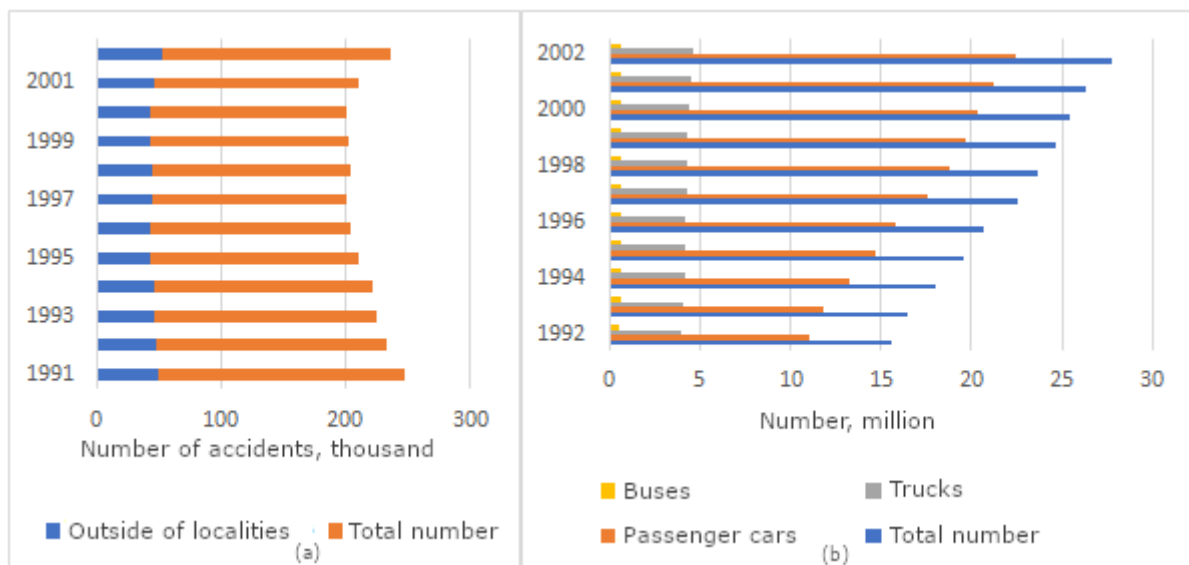


Fig. 4. Traffic safety data analysis graphs, where: a - Change in the number of road accidents in Russia for 1991-2002; b - Change in the number of vehicles in the Russian Federation by years (excluding motorcycles) [6].

After considering the statistics since 1992, it can be said that the population in Russia has changed insignificantly over the given period of time, the number of car parks in Russia has increased 3 times compared to 1992 and 2 times compared to 2002; the length of hard-surface public roads increased by 2.5 times in comparison with 1992 and 2 times in comparison with 2000; the number of accidents decreased by 1.2 times compared to 1992 and 2002.

III. OVERVIEW OF USED ROAD SAFETY PROGRAMS

Based on the analysis of the above statistical data, the Government of the Russian Federation is developing various federal programs and strategies to ensure and improve road safety.

In order to overcome the current negative situation in 2002, a long-term program for the maintenance of the road network was developed, introduced and launched. The adopted program of measures is consistent with the sub-program “Roads” approved by the Government of the Russian Federation as part of the Federal Target Program “Modernization of the Transport System of Russia (2002-2010)”, according to which the reduction of the road component in the total accident rate and the severity of the consequences of road accidents will be achieved mainly, due to a general increase in the transport and operational condition of public roads [2], [13]-[15].

In addition, this program includes the Informatization subprogram, which contains certain projects for the development of information and telecommunication technologies, communication systems on highways. Providing medical, rescue services, as well as road users with operational communication services will significantly reduce the time of accident detection, rescue and evacuation of victims to the nearest medical facilities, and thereby significantly reduce the severity of the consequences of accidents [3], [16]-[18].

In the Russian Federation, the development of similar systems is provided for by the Federal Target Program “Modernization of the Transport System of Russia (2002-2010)”, which includes subprograms and “Roads”,

Currently, the Government has developed a road safety strategy until 2024 [4], [19]-[21]. The state focuses on the road and transport situation in European countries and assumes the achievement of a zero death rate from injuries in road accidents. At the same time, the emphasis is on the irresponsibility of traffic participants, namely on the human factor - the main reason that affects the level of injuries due to road traffic accidents.

IV. RESULTS AND DISCUSSION

The regulation under consideration and the further increase in accident rates over the past four years (Fig. 4a) shows that the usual generally accepted measures related to the control of the traffic situation, the establishment of measures to influence road users for violating traffic rules and the like, in have lost their relevance to a certain extent due to the development and expansion of the road transport network. Therefore, new, more advanced methods and measures are needed that take into account the specifics of the causes of road traffic accidents. Due to the serious growth in the number of vehicles, it is very difficult to change the situation related to solving traffic safety problems in the country (over the past 10 years, the number of vehicles has increased by almost 1.8 times, and the number of cars - more than 2 times) (Figure 2b) with an increase in the length of the road network by only 1.3 times.

The comparative statistics considered above makes it possible to adjust, if necessary, the goals and objectives of the road transport policy defined in the national programs for improving road safety, taking into account the experience of their implementation in the past years. Also, the results of our statistics indicate the need to improve control over the traffic situation. While the results of the study of the causes of road accidents determine the need for the modernization of the road sector.

As the main ways to improve the level of road safety in Russia, it is necessary to consider the use of effective measures to reduce injuries in road accidents by improving the quality of vehicles, using the latest systems of road engineering equipment, information technology and communication systems to the greatest extent on busy road

sections, improving the discipline of drivers to reduce traffic accidents. The set of ongoing programs to modernize road safety at the regional and local levels is particularly effective.

CONCLUSION

The topic of accidents is really significant, because road accidents happen almost daily. At the same time, the above statistics makes us think about how to correct the current situation on the roads. It is a comprehensive analysis that contributes to the development of modern effective measures in the field of road policy aimed at reducing the number of accidents. Without a doubt, it can be argued that there is no specific answer to this question, but increasing the responsibility of each road user for their actions while participating in it can fully modify the current situation.

Future studies plan to collect and analyse data from other countries, also update data and identify other factors influencing road accidents.

REFERENCES

[1] "Accident statistics in Russia: from 2000 to 2019 and the 1st quarter of 2020," 2022. <https://rosinfostat.ru/dtp/#i> (accessed Jan. 17, 2022).

[2] "Overview information "The impact of the development and condition of the road network on the level of traffic safety on the roads of Russia. Overview information"," 2022. <https://files.stroyinf.ru/Data1/56/56230/index.htm#i27197> (accessed Jan. 17, 2022).

[3] V. V. Chvanov, "Transport corridors - reliable communication," Russian Federation today, no. 4, pp. 57-59, 2001.

[4] "Road Safety Strategy until 2024 in the Russian Federation - tasks for road safety," 2022. <https://avtobddinfo.ru/bdd/proekt-strategii-bezopasnosti-dorozhnogo-dvizheniya-do-2024> (accessed Jan. 21, 2022).

[5] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 2, p. 022106, Aug. 2019, doi: 10.1088/1755-1315/315/2/022106.

[6] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.

[7] N. V. Fedorova, N. N. Dzhioeva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogor, and V. S. Tynchenko, "Methods of assessing the efficiency of the foundry industrial marketing," IOP Conference Series: Materials Science and Engineering, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.

[8] Y. Wu et al., "Influence of thermal and lighting factors on human perception and work performance in simulated underground environment," Science of the Total Environment, vol. 828, p. 154455, Jul. 2022, doi: 10.1016/j.scitotenv.2022.154455.

[9] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1-6. doi: 10.1109/RUSAUTOCON.2018.8501776.

[10] S. Kurashkin, D. Rogova, V. Tynchenko, V. Petrenko, and A. Milov, (2020). "Modeling of Product Heating at the Stage of Beam Input in the Process of Electron Beam Welding Using the COMSOL Multiphysics System," in: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing, vol 1294, https://doi.org/10.1007/978-3-030-63322-6_77.

[11] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," Journal of Physics: Conference Series, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.

[12] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. V. Ereemeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.

[13] V. G. Loginov, M. N. Ignatyeva, and V. V. Balashenko, "Consistent approach to assess the comfort of living in the northern and arctic areas," Economy of Region, vol. 14, no. 4, pp. 1399-1410, 2018, doi: 10.17059/2018-4-26.

[14] V. S. Tynchenko, A. v. Milov, V. V. Tynchenko, V. V. Bukhtoyarov, and V. V. Kukartsev, "Intellectualizing the process of waveguide tracks induction soldering for spacecrafts," International Review of Aerospace Engineering, vol. 12, no. 6, pp. 280-289, 2019, doi: 10.15866/irease.v12i6.16910.

[15] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, S. G. Dokshanin, and V. V. Kukartsev, "Research of methods for design of regression models of oil and gas refinery technological units," IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042078, Jun. 2019, doi: 10.1088/1757-899X/537/4/042078.

[16] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, "Intellectual support system of administrative decisions in the big distributed geoinformation systems," International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, vol. 1, no. 2, pp. 227-232, 2014, doi: 10.5593/sgem2014/b21/s7.029.

[17] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," IOP Conference Series: Earth and Environmental Science, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.

[18] A.V. Murygin, S.O. Kurashkin, V.S. Tynchenko, and D.V. Rogova, "The use of ANSYS for modelling the energy distribution in steady mode with electron beam welding," Journal of Physics: Conference Series, vol. 1889, no. 4, pp. 042061, May. 2021, doi: 10.1088/1742-6596/1889/4/042061.

[19] D. S. Shalaeva, O. I. Kukartseva, V. S. Tynchenko, V. V. Kukartsev, S. V. Aponasenko, and E. V. Stepanova, "Analysis of the development of global energy production and consumption by fuel type in various regions of the world," IOP Conference Series: Materials Science and Engineering, vol. 952, no. 1, p. 012025, Nov. 2020, doi: 10.1088/1757-899X/952/1/012025.

[20] A. A. Boyko, V. V. Kukartsev, K. Y. Lobkov, and A. A. Stupina, "Strategic planning toolset for reproduction of machinebuilding engines and equipment," Journal of Physics: Conference Series, vol. 1015, no. 4, p. 042006, May 2018, doi: 10.1088/1742-6596/1015/4/042006.

[21] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, "Storage and analysis of natural resources information in various territories," Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.

Establishment of a Model for Managing Organizational Attendance Based on Data Analysis

Sergei Kurashkin

¹Information Control Systems

Department

Reshetnev Siberian State University of Science and Technology

²Laboratory of Biofuel Compositions

Siberian Federal University

³Digital Material Science: New

Materials and Technologie

Bauman Moscow State Technical

University

Krasnoyarsk, Russia

0000-0002-4017-4369

Viktor Suetin

Department of System Analysis

Reshetnev Siberian State University of

Science and Technology

Krasnoyarsk, Russia

suetin@sibsau.ru

Anton Mikhalev

Department of Informatics

Siberian Federal University

Krasnoyarsk, Russia

0000-0002-8986-5953

Alexander Korostelev

Department of System Analysis

Reshetnev Siberian State University of Science and Technology

Krasnoyarsk, Russia

korostelev@sibsau.ru

Vladimir Grishko

Department of Informatics

Siberian Federal University

Krasnoyarsk, Russia

0000-0001-5541-5471

Abstract—The article studies the influence of external factors on the attendance of the shopping center. It is necessary to isolate the essential traits-factors that determine the attendance of the shopping center (SC) and establish their mathematical dependence with the resulting indicator. Thus, in the course of the study, it was made possible to find various data for the construction of a regression model, namely the number of visits to the shopping center «Planet» for October 2017 and various factors such as: whether it was rain or not, the index of weather sensitivity, temperature, wind speed and day of the week, which can affect the visit. The data were formatted to exclude emissions to avoid model inadequacy. Next, a linear regression model was built, on the basis of which it is possible to determine the factors that influence the visits to shopping centers more. If the constructed model is adequate, then it will be possible to draw conclusions on what factors are more worth paying attention to, to determine the peaks of attendance at the SC.

Keywords— Correlation, data, excel, model, regression analysis

I. INTRODUCTION

Today the SC is no longer something new and unknown [1]-[3]. The SC is a set of trade and/or service enterprises that sell a universal or specialized range of goods and services located in a specific area in buildings, Planned, constructed and managed as a single unit and provided car parking within their territory [4]-[6].

The SC is also a large number of additional jobs and the provision of development for small businesses. We are

talking not only about individual stores or retail chains within the SC, but also about shopping “islands” that stand separately in the middle of the pavilion. This type of activity is an example of the effective use of the territory and the expansion of the number of tenants, and hence the increase in profits [7]-[9]. Such a system helps a business that does not have enough funds to rent a separate room or is at the initial stage of development [10]-[12].

However, the main indicator of the effective operation of the shopping center is attendance. After all, visitors are the basis of any trading activity, which means that their increase helps to improve the quantity and quality of sales. But what exactly attracts visitors to shopping malls? What factors influence attendance? [13]-[16]

Visiting the shopping center depends on some of the factors described in this paper. It is necessary to find out on which factors attendance depends, and which factors are interconnected. Using correlation-regression analysis of multivariate data, it is possible to determine what factors affect the number of people in the shopping center and whether they are interconnected [17]-[19].

II. DATA AND RESEARCH METHODS

All collected data for October 2017 are presented in Table I and the explanation in Table II.

TABLE I. INITIAL DATA

Date	Attendance	Was it raining or not	Weather sensitivity index	Temperature	Wind(m/s)	Day of the week
01.10.2017	55300	1	4	6	2.3	7
02.10.2017	36380	2	3	2	3.1	1
03.10.2017	35928	2	3	8	3.4	2
04.10.2017	37140	2	1	3	1.2	3

05.10.2017	38480	2	1	4	2	4
06.10.2017	42580	2	1	4	1.5	5
07.10.2017	61680	2	3	6	3.5	6
08.10.2017	56380	1	3	5	3	7
09.10.2017	37020	1	3	5	9.1	1
10.10.2017	34430	2	1	6	4.2	2
11.10.2017	36480	1	3	5	2.8	3
12.10.2017	40360	2	4	0	2.8	4
13.10.2017	41100	2	3	2	2.5	5
14.10.2017	59300	2	1	0	0.6	6
15.10.2017	55830	2	1	3	0.3	7
16.10.2017	36440	2	3	9	0.5	1
17.10.2017	35560	2	1	4	1.5	2
18.10.2017	37730	2	3	4	2.7	3
19.10.2017	37380	2	3	4	3	4
20.10.2017	41770	2	3	5	1.4	5
21.10.2017	62960	2	3	8	2.7	6
22.10.2017	60320	2	3	11	2.7	7
23.10.2017	33960	2	1	2	7.9	1
24.10.2017	35150	1	1	3	7.2	2
25.10.2017	35230	1	1	3	6	3
26.10.2017	38180	2	2	1	1.7	4
27.10.2017	41700	2	2	-1	6.6	5
28.10.2017	61680	2	1	5	4.1	6
29.10.2017	57120	2	1	5	3.5	7
30.10.2017	36460	2	4	-1	6.3	1
31.10.2017	35630	2	4	-3	3.7	2

TABLE II. EXPLANATION OF TABLE I.

Was it raining or not	
1	It was raining
2	There was no rain
Weather sensitivity index	
1	Comfortable weather conditions for weather-dependent people.
2	Weather conditions may affect some people with increased weather sensitivity.
3	There is a high probability of the influence of weather conditions on the well-being of weather-dependent people.
4	The meteorological situation can cause a sharp deterioration in the well-being of weather-sensitive people.

Based on the collected data, 5 factors were identified that may affect the number of people. The main indicator is the number of visits. The next indicator that may affect visits is the presence of rain on a given day, weather sensitivity index, temperature on a given day, wind speed and day of the week [20].

The Bravais-Pearson correlation coefficient ($r_{p_{xy}}$) is -0.262, -0.008, 0.338, -0.283, 0.864. On the Chaddock scale: $0.1 < r_{xy} < 0.3$, $0.3 < r_{xy} < 0.5$, $0.7 < r_{xy} < 0.9$, therefore, in the example under consideration, the correlation between feature Y and factor X in the first, second and fourth cases will be weak, in the third case moderate, and in the third fifth high (Table III).

TABLE III. TABLE FRAGMENT

The Bravais-Pearson correlation coefficient	0,055	-0,008	0,337	-0,282	0,864
Least squares method	42593,333	43598,620	46217,005	45109,820	56548,939
Fisher 's Criterion	0,055	-0,008	0,351	-0,290	1,309
Student's Criterion	2,426E-21	2,427E-21	2,422E-21	2,433E-21	2,417E-21

Let's check the adequacy of the constructed model to the object using the Fisher criteria.

The idea behind the adequacy test is to compare the prediction variance based on the regression model under study with the noise variance. The data are presented in table IV.

TABLE IV. FISHER CRITERIA

	Initial data	Fisher 's Criterion
Factor 1	-0.262	-0.268

Factor 2	-0.009	-0.008
Factor 3	0.338	0.352
Factor 4	-0.283	-0.290
Factor 5	0.864	1.309

Conclusion: In the first, second, third and fourth cases, F_{tab} is almost equal to F, but still a little more, so we can conclude that the model is adequate, and in the fifth case $F > F_{tab}$, because of this, it is concluded that the model is inadequate and it is necessary improve the model.

We carry out regression analysis to check the quality of the constructed models. The data is presented in Tables V-

TABLE V. THE RESULT OF THE REGRESSION ANALYSIS OF THE FIRST FACTOR

OUTPUT OF RESULTS	-	-	-	-	-	-
-	-	-	-	-	-	-
Regression statistics	-	-	-	-	-	-
<i>Multiple R</i>	0.055448366	-	-	-	-	-
<i>R-square</i>	0.003074521	-	-	-	-	-
<i>Normalized R-square</i>	0.031302219	-	-	-	-	-
<i>Standard error</i>	0.407846868	-	-	-	-	-
<i>Observations</i>	31	-	-	-	-	-
-	-	-	-	-	-	-
Analysis of variance	-	-	-	-	-	-
-	df	SS	MS	F	Significance F	-
<i>Regression</i>	1	0.014876716	0.014876716	0.08943609	0.767028004	-
<i>Remainder</i>	29	4.823832962	0.166339068	-	-	-
<i>Total</i>	30	4.838709677	-	-	-	-
-	-	-	-	-	-	-
-	Coefficient	Standard error	t-statistic	P-Value	Lower 95%	Upper 95%
<i>Y-intersection</i>	1.700258619	0.362567616	4.689493882	6.00855E-05	0.958724583	2.441792656
<i>Variable X1</i>	2.17825E-06	7.28367E-06	0.299058673	0.767028004	-1.27185E-05	1.7075E-05

TABLE VI. THE RESULT OF THE REGRESSION ANALYSIS OF THE SECOND FACTOR

OUTPUT OF RESULTS	-	-	-	-	-	-
-	-	-	-	-	-	-
Regression statistics	-	-	-	-	-	-
<i>Multiple R</i>	0.008631522	-	-	-	-	-
<i>R-square</i>	7.45032E-05	-	-	-	-	-
<i>Normalized R-square</i>	-0.034405686	-	-	-	-	-
<i>Standard error</i>	1.150475716	-	-	-	-	-
<i>Observations</i>	31	-	-	-	-	-
-	-	-	-	-	-	-
Analysis of variance	-	-	-	-	-	-
-	df	SS	MS	F	Significance F	-
<i>Regression</i>	1	0.00285996	0.00285996	0.002160753	0.963243203	-
<i>Remainder</i>	29	38.38423681	1.323594373	-	-	-
<i>Total</i>	30	38.38709677	-	-	-	-
-	-	-	-	-	-	-
-	Coefficient	Standard error	t-statistic	P-Value	Lower 95%	Upper 95%
<i>Y-intersection</i>	2.336883584	1.022749642	2.284902862	0.029819325	0.245125699	4.428641468
<i>Variable X1</i>	-9.55066E-07	2.05462E-05	-0.046483898	0.963243203	-4.29767E-05	4.10666E-05

TABLE VII. THE RESULT OF THE REGRESSION ANALYSIS OF THE THIRD FACTOR

OUTPUT OF RESULTS						
-						
Regression statistics						

<i>Multiple R</i>	0.338001338					
<i>R-square</i>	0.114244904					
<i>Normalized R-square</i>	0.083701625					
<i>Standard error</i>	2.918329753					
<i>Observations</i>	31					
-						
Analysis of variance						
-	df	SS	MS	F	Significance F	
<i>Regression</i>	1	31.85590172	31.85590172	3.740426946	0.062922974	
<i>Remainder</i>	29	246.982808	8.51664855			
<i>Total</i>	30	278.8387097				
-						
-	Coefficient	Standard error	t-statistic	P-Value	Lower 95%	Upper 95%
<i>Y-intersection</i>	-1.107572481	2.594336125	-0.426919423	0.67259093	-6.413585626	4.198440664
<i>Variable X1</i>	0.000100797	5.2118E-05	1.934018342	0.062922974	-5.79611E-06	0.00020739

TABLE VIII. THE RESULT OF THE REGRESSION ANALYSIS OF THE FOURTH FACTOR

OUTPUT OF RESULTS						
-						
Regression statistics						
<i>Multiple R</i>	0.283037069					
<i>R-square</i>	0.080109983					
<i>Normalized R-square</i>	0.048389637					
<i>Standard error</i>	2.145648871					
<i>Observations</i>	31					
-						
Analysis of variance						
-	df	SS	MS	F	Significance F	
<i>Regression</i>	1	11.62695615	11.62695615	2.52550789	0.122862984	
<i>Remainder</i>	29	133.5104632	4.603809076			
<i>Total</i>	30	145.1374194				
-						
-	Coefficient	Standard error	t-statistic	P-Value	Lower 95%	Upper 95%
<i>Y-intersection</i>	6.317148857	1.907438448	3.311849388	0.002489552	2.415999202	10.21829851
<i>Variable X1</i>	-6.08957E-05	3.83188E-05	-1.589184662	0.122862984	-0.000139266	1.74751E-05

TABLE IX. THE RESULT OF THE REGRESSION ANALYSIS OF THE FIFTH FACTOR

OUTPUT OF RESULTS						
-						
Regression statistics						
<i>Multiple R</i>	0.86					
<i>R-square</i>	0.75					
<i>Normalized R-square</i>	0.74					
<i>Standard error</i>	1.08					
<i>Observations</i>	31					
-						

Analysis of variance						
-	df	SS	MS	F	Significance F	
<i>Regression</i>	1	99.98	99.98	85.57	3.751E-10	
<i>Remainder</i>	29	33.89	1.1685			
<i>Total</i>	30	133.87				
-						
-	Coefficient	Standard error	t-statistic	P-Value	Lower 95%	Upper 95%
<i>Y-intersection</i>	-4.77	0.96	-4.96	2.8E-05	-6.74	-2.80
<i>Variable X1</i>	0.0001	1.93E-05	9.25	3.8E-10	0.0001	0.0002

One of the main indicators is the R-square, which indicates the quality of the model. In our case, this coefficient is 0.262, 0.008, 0.338, 0.283, 0.864, which corresponds to a low, medium and high quality level of the model.

The higher the value of the coefficient of determination, the more applicable the chosen model for a particular task. It is believed that it correctly describes the real situation when the value of R-square is higher than 0.7.

Another important indicator is located in the cell at the intersection of the “Y-intersection” line and the “Coefficients” column. Here it is indicated what value Y will have, and in our case, this is the visit rate, with all other factors equal to zero. In this table, this value is 0.75, 2.33, -1.1, 6.317, -4.77.

The value at the intersection of the column “Variable X1” and “Coefficients” shows the level of dependence of Y on X. In our case, this is the level of dependence of the visit

indicator on other factors. Coefficient -4.22, 9.5, 0.0001, -6.08, 0.00017.

III. RESULTS AND DISCUSSION.

As a result of the study, it was revealed that factors such as the meteosenstivity index affect shopping center visits to a lesser extent. Temperature, the presence of rain on a certain day and wind speed has a medium effect, and the day of the week has a greater effect. Based on this, all 5 factors can be divided into 3 groups: group 1 - little influencing factors (yellow color), group 2 - medium influencing factors (green color), group 3 - strongly influencing factors (red color). Looking at this table, you can understand that in order to determine the peaks of visiting a shopping center, first of all, you need to look at the day of the week, then at temperature, wind speed and the presence of rain. The data are presented in Table X.

TABLE X. DEGREE OF INFLUENCE OF FACTORS

Date	Attendance	Was it raining or not	Weather sensitivity index	Temperature	Wind (m/s)	Day of the week
01.10.2017	55300	1	4	6	2.3	7
02.10.2017	36380	2	3	2	3.1	1
03.10.2017	35928	2	3	8	3.4	2
04.10.2017	37140	2	1	3	1.2	3
05.10.2017	38480	2	1	4	2	4
06.10.2017	42580	2	1	4	1.5	5
07.10.2017	61680	2	3	6	3.5	6
08.10.2017	56380	1	3	5	3	7
09.10.2017	37020	1	3	5	9.1	1
10.10.2017	34430	2	1	6	4.2	2
11.10.2017	36480	1	3	5	2.8	3
12.10.2017	40360	2	4	0	2.8	4
13.10.2017	41100	2	3	2	2.5	5
14.10.2017	59300	2	1	0	0.6	6
15.10.2017	55830	2	1	3	0.3	7
16.10.2017	36440	2	3	9	0.5	1
17.10.2017	35560	2	1	4	1.5	2
18.10.2017	37730	2	3	4	2.7	3
19.10.2017	37380	2	3	4	3	4
20.10.2017	41770	2	3	5	1.4	5
21.10.2017	62960	2	3	8	2.7	6
22.10.2017	60320	2	3	11	2.7	7
23.10.2017	33960	2	1	2	7.9	1
24.10.2017	35150	1	1	3	7.2	2
25.10.2017	35230	1	1	3	6	3
26.10.2017	38180	2	2	1	1.7	4
27.10.2017	41700	2	2	-1	6.6	5
28.10.2017	61680	2	1	5	4.1	6
29.10.2017	57120	2	1	5	3.5	7
30.10.2017	36460	2	4	-1	6.3	1
31.10.2017	35630	2	4	-3	3.7	2

CONCLUSION

Analysis of traffic peaks by day helps to:

- Choose the optimal place and time for promotions: where it is better to hold the promotions, at what time, how long.
- To understand what time and days of the week are key to work. Accordingly, when it is necessary to withdraw the best staff, be present management, strengthen the replacement by additional staff.

The simulation of attendance can be used to solve the problems of assessing the quality of incoming data on the attendance of groups of objects in the construction of market indicators, as well as to solve a wide range of tasks related to the work of the mall.

REFERENCES

- [1] M. Karimimoshaver, M. S. Shahrak, "The effect of height and orientation of buildings on thermal comfort," *Sustainable Cities and Society*, vol. 79, p. 103720, 2022.
- [2] "Official website of the Planeta Shopping Center," 2022. <https://krs.planeta-mall.ru/> (accessed May 07, 2022).
- [3] E. A. Chzhan, V. S. Tynchenko, V. V. Kukartsev, N. V. Fedorova, A. S. Yamshchikov, and D. A. Krivov, "Essence and classification of the agribusiness organizations competitive strategies," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 2, p. 022106, Aug. 2019, doi: 10.1088/1755-1315/315/2/022106.
- [4] A. V. Kukartsev, A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. V. Bukhtoyarov, and S. V. Tynchenko, "Methods of business processes competitiveness increasing of the rocket and space industry enterprise," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042009, Jun. 2019, doi: 10.1088/1757-899X/537/4/042009.
- [5] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, "The simulation model of fixed assets reproduction of mechanical engineering enterprises," in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.
- [6] Y. Sun, X. Zhang, A. T. K. Wan, and S. Wang, "Model averaging for interval-valued data," *European Journal of Operational Research*, vol. 301, no. 2, pp. 772–784, Sep. 2022, doi: 10.1016/j.ejor.2021.11.015.
- [7] N. V. Fedorova, N. N. Dzhioeva, V. V. Kukartsev, N. A. Dalisova, A. R. Ogol, and V. S. Tynchenko, "Methods of assessing the efficiency of the foundry industrial marketing," *IOP Conference Series: Materials Science and Engineering*, vol. 734, no. 1, p. 012083, Jan. 2020, doi: 10.1088/1757-899X/734/1/012083.
- [8] N. V. Fedorova, V. V. Kukartsev, V. S. Tynchenko, S. M. Atluhanov, D. K. Gek, and E. A. Zagudaylova, "Problems of the digital economy development in the transport industry," *IOP Conference Series: Earth and Environmental Science*, vol. 315, no. 3, p. 032047, Aug. 2019, doi: 10.1088/1755-1315/315/3/032047.
- [9] D. S. Shalaeva, O. I. Kukartseva, V. S. Tynchenko, V. V. Kukartsev, S. V. Aponasenko, and E. V. Stepanova, "Analysis of the development of global energy production and consumption by fuel type in various regions of the world," *IOP Conference Series: Materials Science and Engineering*, vol. 952, no. 1, p. 012025, Nov. 2020, doi: 10.1088/1757-899X/952/1/012025.
- [10] A. A. Boyko, V. V. Kukartsev, K. Y. Lobkov, and A. A. Stupina, "Strategic planning toolset for reproduction of machinebuilding engines and equipment," *Journal of Physics: Conference Series*, vol. 1015, no. 4, p. 042006, May 2018, doi: 10.1088/1742-6596/1015/4/042006.
- [11] A. A. Boyko, V. v. Kukartsev, D. V. Ereemeev, V. S. Tynchenko, V. V. Bukhtoyarov, and A. A. Stupina, "Imitation-dynamic model for calculating the efficiency of the financial leverage," *Journal of Physics: Conference Series*, vol. 1353, no. 1, p. 12123, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012123.
- [12] V. S. Tynchenko, A. v. Milov, V. V. Tynchenko, V. V. Bukhtoyarov, and V. V. Kukartsev, "Intellectualizing the process of waveguide tracks induction soldering for spacecrafts," *International Review of Aerospace Engineering*, vol. 12, no. 6, pp. 280–289, 2019, doi: 10.15866/irease.v12i6.16910.
- [13] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, "Intellectual support system of administrative decisions in the big distributed geoinformation systems," *International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM*, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.
- [14] A.V. Murygin, S.O. Kurashkin, V.S. Tynchenko, and D.V. Rogova, "The use of ANSYS for modelling the energy distribution in steady mode with electron beam welding," *Journal of Physics: Conference Series*, vol. 1889, no. 4, pp. 042061, May. 2021, doi: 10.1088/1742-6596/1889/4/042061.
- [15] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, "Storage and analysis of natural resources information in various territories," *Journal of Physics: Conference Series*, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.
- [16] V. G. Loginov, M. N. Ignatyeva, and V. V. Balashenko, "Consistent approach to assess the comfort of living in the northern and arctic areas," *Economy of Region*, vol. 14, no. 4, pp. 1399–1410, 2018, doi: 10.17059/2018-4-26.
- [17] S. Kurashkin, D. Rogova, V. Tynchenko, V. Petrenko, and A. Milov, (2020). "Modeling of Product Heating at the Stage of Beam Input in the Process of Electron Beam Welding Using the COMSOL Multiphysics System," in: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) *Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing*, vol 1294, https://doi.org/10.1007/978-3-030-63322-6_77.
- [18] A. A. Boyko, V. V. Kukartsev, E. S. Smolina, V. S. Tynchenko, Y. I. Shamlitskiy, and N. V. Fedorova, "Imitation-dynamic model of amortization of reproductive effect with different methods of calculation," *Journal of Physics: Conference Series*, vol. 1353, no. 1, p. 012124, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012124.
- [19] V. S. Tynchenko, V. V. Tynchenko, V. V. Bukhtoyarov, V. V. Kukartsev, V. A. Kukartsev, and D. v. Ereemeev, "Application of Kohonen self-organizing maps to the analysis of enterprises' employees certification results," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042010, Jun. 2019, doi: 10.1088/1757-899X/537/4/042010.
- [20] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, S. G. Dokshanin, and V. V. Kukartsev, "Research of methods for design of regression models of oil and gas refinery technological units," *IOP Conference Series: Materials Science and Engineering*, vol. 537, no. 4, p. 042078, Jun. 2019, doi: 10.1088/1757-899X/537/4/042078.

Software Products Using an Object Approach

Evgeniya Semenova
 Department of System Analysis
 Reshetnev Siberian State University of
 Science and Technology
 Krasnoyarsk, Russia
 cancsl.v@yandex.ru

Vadim Tynchenko
¹Information Control Systems
 Department
 Reshetnev Siberian State University of
 Science and Technology
²Department of Technological
 Machines and Equipment
 for Oil and Gas Complex
 Siberian Federal University
³Digital Material Science: New
 Materials and Technologie
 Bauman Moscow State Technical
 University
 Krasnoyarsk, Russia
 0000-0002-3959-2969

Sofya Chashchina
 Department of Information Economic
 Systems
 Reshetnev Siberian State University of
 Science and Technology
 Krasnoyarsk, Russia
 sofyachash@mail.ru

Viktor Suetin
 Department of System Analysis
 Reshetnev Siberian State University of Science and Technology
 Krasnoyarsk, Russia
 suetin@sibsau.ru

Alexander Stashkevich
 Department of System Analysis
 Reshetnev Siberian State University of Science and Technology
 Krasnoyarsk, Russia
 0000-0001-6052-3901

Abstract—One of the most popular languages of graphic modeling is Unified Modeling Language (UML). This abstract presents an overview of the language UML. A wide range of diagrams were considered, together with a brief description of their application, the main advantages and disadvantages of the UML language, its object-oriented approach. Strengths (Strengths), Weaknesses (Weaknesses), Capabilities (Opportunities) and Threats (Threats) have been formulated for SWOT analysis. The SWOT-matrix is based on them. A detailed structure of diagram types was presented, dividing the above-mentioned diagrams into structural and behavioral diagrams. Among all the applications of the UML language in more detail considered the possibilities of its use in software engineering, and specifically - in software development. According to the study, it was found that the UML diagram language is one of the best tools.

Keywords— Graphic modeling, software, SWOT matrix, UML

I. INTRODUCTION

UML is a graphical description language that has been around for over 20 years, consisting of an integrated set of diagrams and used for object modeling in software development. UML is a generalist language and is suitable for graphical description of abstract system models because it is an open standard. UML was created to design, visualize, document and define, as a rule, software systems. UML is by no means a programming language; however, UML diagrams can be used to generate code [1]-[3].

II. UML DIAGRAM

Since diagrams are an integral part of the UML, it is a good idea to consider the types of these diagrams.

A list of most of the diagrams grouped is shown in Figure 1 below.

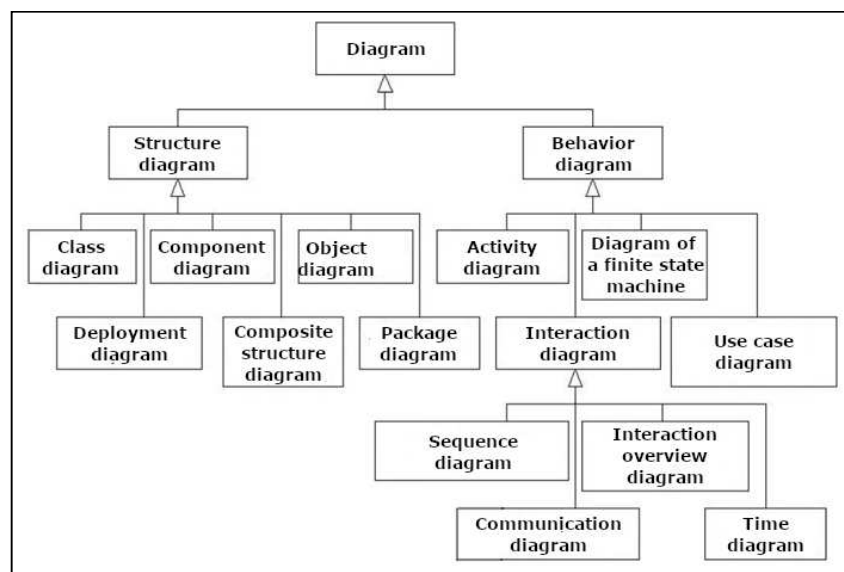


Fig. 1. Types of charts.

III. STRUCTURAL DIAGRAM

A. Class Diagram

A class diagram is used to illustrate the relationship between classes, objects, and other entities of the OOP paradigm. In addition, the class diagram displays in detail the relationship between the entities of a particular subject area. As an example, objects and subsystems [4]-[6].

B. Deployment Diagram

A deployment diagram shows a set of nodes and their relationships. Describes a static representation of an architecture. Associated with a component diagram (each node usually hosts one or more components) [7]-[9].

C. Component Diagram

This diagram illustrates the physical state of the system. It describes each of its files and how it functions with other files.

On it you can see both databases and code files, hardware components or business logic.

D. Composite structure diagram

A composite structure diagram represents the internal structure of a class or collaboration. There is little difference between a component diagram and a composite structure diagram, which is why they are treated as component diagrams in this book [10]-[12].

E. Object Diagram

This type of diagrams will inherit from class diagrams, as a result, they have a one-way dependency.

If you follow the OOP paradigm, then an object diagram is an instance of a class diagram. Their main meaning is the same, the only difference is that the object diagram is a "snapshot" of a certain moment during system modeling [13]-[15].

F. Package Diagram

Package diagrams show the dependencies between the packages that make up the model.

IV. BEHAVIOR DIAGRAM

A. Activity Diagram

The activity diagram shows the flow of calculations step by step. An activity describes a set of activities, their sequencing or branching flow, and the values generated or used by the activities. Activity diagrams illustrate the dynamic view of a system. They are especially important when modeling system functions. Focus on the flow of control when implementing some behavior [16]-[18].

B. State Machine Diagram

A diagram that shows a state machine with its associated states and transitions. The state machine is a linear display of state transitions of a certain element of the system. A CA always has an initial and final state. In other words, with the help of such a diagram, you can see what decisions any object makes, based on the current situation.

C. Interaction Diagram

Interaction diagram is a common name for sequence and communication diagrams. All sequence and communication

diagrams are interaction diagrams, and all interaction diagrams are either sequence diagrams or communication diagrams. They are based on the same model, although in practice they include different entities.

D. Use Case Diagram

This diagram allows you to see the relationship between the system actant and use cases, and is used to ensure that the customer, end user and developer can interpret the functionality and requirements of the system in the same way.

V. USING UML

The UML language is used in business process modeling, system design, displaying the structure of an organization, describing technical processes, and so on [19]-[22].

UML allows software developers to use one standard for graphic representations of classes, roles, generalizations, components, which allows them not to waste time on the agreement of the general form of the diagram, but to concentrate more on design and architecture.

A. Practical Application In Software Engineering

Quite often, when developing software, there is a need for a graphical, more accessible to many, way of representing abstract system models. The most common in UML are class diagrams, use case diagrams, and sequence diagrams.

A class diagram is necessary to illustrate the structure of classes in a program, class attributes, and class interactions.

The Use Case Diagram is necessary to illustrate the possible interactions of the actors with the program and the outcomes for each of these interactions.

A sequence diagram is needed to illustrate the life cycle of an object and the interaction of actors within a given use case.

VI. SWOT-ANALYSIS

A SWOT analysis of the UML methodology was carried out, the result of which was the SWOT matrix shown in Table I.

A. Strengths

- The UML is object-oriented, which allows it to function very well with the OOP paradigm that is popular these days.
- UML allows you to build diagrams with such detail that all the details of the system will be shown in an understandable way.
- UML methodology is easy enough for people with different levels of education to understand.

B. Weaknesses

- Newer versions of UML introduce new diagrams that are hardly ever used.
- UML uses only two classes of models, the diagrams of which were described above: structural and behavioral. Thus, the functional model remains

unaffected, and for its implementation, if required, it will be necessary to use other methodologies, for example, IDEF0 or DFD.

C. Opportunities

- UML is easily applied in the construction of various technical processes or in business processes.
- Due to its popularity, the UML is still getting updates, including new diagrams.

D. Threats

- Due to the widespread use of UML, it tries to become as versatile as possible, but for some projects where details and formalities are important, UML diagram engineers have to limit its functionality.
- Not all programming languages in which projects are implemented support an object-oriented approach. For such projects, this methodology is not suitable.

E. SWOT Matrix

TABLE I. SWOT MATRIX

	Opportunities	Threats
Strengths	Significant potential of using UML diagrams in technical and business design	Within the framework of an object – oriented approach , it provides the widest possibilities for modeling with varying degrees of detail
Weaknesses	Application of rarely used diagrams in particular cases in technical and business design	Using new diagrams to solve new problems

CONCLUSION

UML is a good tool for creating diagrams of any system. Its functionality allows you to clearly demonstrate almost any process during project development. In addition to IT - the UML sphere has established itself in other professional fields, where it is also necessary to demonstrate the details of the system.

UML is versatile, convenient and understandable for any person, and its object-oriented approach allows you to integrate into most modern programming languages.

Almost all its flaws are special cases, in general, it is one of the best tools to date.

REFERENCES

[1] G. Butch, J.Rambo and I. Jacobson, “The UML language. User’s guide : educational and methodical manual, translated from English,” Moscow : DMK Press, p. 248, 2006.

[2] K. Larman, C. Larman, “Application of UML and design patterns, second edition : educational and methodical manual,” Kiev: Williams, p. 620, 2004.

[3] F. Ah. Novikov. “Analysis and projectorization of UML : school-methodical work,” Saint-Petersburg: YTMO, p. 286, 2007.

[4] A. A. Boyko, V. v. Kukartsev, D. V Ereemeev, V. S. Tynchenko, V. V. Bukhtoyarov, and A. A. Stupina, “Imitation-dynamic model for calculating the efficiency of the financial leverage,” Journal of Physics: Conference Series, vol. 1353, no. 1, p. 12123, Nov. 2019, doi: 10.1088/1742-6596/1353/1/012123.

[5] N. O. Dorodnykh, “The use of UML class diagrams for the formation of production knowledge bases,” Software Engineering, vol. 4. pp. 3-9, 2015.

[6] V. S. Tynchenko, A. v. Milov, V. V. Tynchenko, V. V Bukhtoyarov, and V. V. Kukartsev, “Intellectualizing the process of waveguide tracks induction soldering for spacecrafts,” International Review of Aerospace Engineering, vol. 12, no. 6, pp. 280–289, 2019, doi: 10.15866/irease.v12i6.16910.

[7] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, S. G. Dokshanin, and V. V. Kukartsev, “Research of methods for design of regression models of oil and gas refinery technological units,” IOP Conference Series: Materials Science and Engineering, vol. 537, no. 4, p. 042078, Jun. 2019, doi: 10.1088/1757-899X/537/4/042078.

[8] O. Antamoshkin, V. Kukarcev, A. Pupkov, and R. Tsarev, “Intellectual support system of administrative decisions in the big distributed geoinformation systems,” International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, vol. 1, no. 2, pp. 227–232, 2014, doi: 10.5593/sgem2014/b21/s7.029.

[9] A. S. Mikhalev, V. S. Tynchenko, V. V. Kukartsev, L. N. Korpacheva, V. A. Kukartsev, and A. V. Rozhkova, “Storage and analysis of natural

resources information in various territories,” Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012181, Nov. 2020, doi: 10.1088/1742-6596/1661/1/012181.

[10] D. Torre, M. Genero, Y. Labiche, and Elaasar, M. “How consistency is handled in model-driven software engineering and UML: an expert opinion survey,” Software Quality Journal, pp. 1-54, 2022.

[11] V. Tynchenko et al., “Mathematical Modeling of Induction Heating of Waveguide Path Assemblies during Induction Soldering,” Metals (Basel), vol. 11, no. 5, p. 697, Apr. 2021, doi: 10.3390/MET11050697.

[12] M. Latifaj, F. Ciccozzi, M. Mohlin, and E. Posse, “Towards automated support for blended modelling of UML-RT embedded software architectures,” in 15th European Conference on Software Architecture ECSA 2021, Sep. 2021, Virtual (originally Växjö), Sweden.

[13] V. V. Bukhtoyarov, V. S. Tynchenko, E. A. Petrovsky, V. V. Kukartsev, and A. I. Kuklina, “Evolutionary method for automated design of models of vortex flowmeters transformation function,” Journal of Physics: Conference Series, vol. 1118, no. 1, p. 012041, Dec. 2018, doi: 10.1088/1742-6596/1118/1/012041.

[14] C. Moral, A. de Antonio, X. Ferre, and J. Ramirez, “A proposed UML-based common model for information visualization systems,” Multimedia Tools and Applications, vol. 80, no. 8, pp. 12541-12579.

[15] V. V. Kukartsev, A. A. Boyko, and O. A. Antamoshkin, “The simulation model of fixed assets reproduction of mechanical engineering enterprises,” in 2018 International Russian Automation Conference, Oct. 2018, pp. 1–6. doi: 10.1109/RUSAUTOCON.2018.8501776.

[16] V. S. Tynchenko et al., “Software to Predict the Process Parameters of Electron Beam Welding,” IEEE Access, vol. 9, pp. 92483–92499, 2021, doi: 10.1109/ACCESS.2021.3092221.

[17] A. A. Boyko, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, E. A. Chzhan, and A. S. Mikhalev, “Dynamic simulation of calculating the purchase of equipment on credit,” Journal of Physics: Conference Series, vol. 1333, no. 3, p. 032009, Nov. 2019, doi: 10.1088/1742-6596/1333/3/032009.

[18] C. Sáenz-Adán, B. Pérez, F. J. García-Izquierdo, and L. Moreau, “Integrating Provenance Capture and UML with UML2PROV: Principles and Experience,” IEEE Transactions on Software Engineering, pp. 53-68, Feb. 2020.

[19] A. O. Stupin, V. V. Kukartsev, V. S. Tynchenko, V. A. Kukartsev, A. I., Cherepanov, and A. V. Rozhkova, “Management modelling of the natural resources extraction station by agency modelling means,” Journal of Physics: Conference Series, vol. 1661, no. 1, p. 012196, Nov. 2020.

[20] I. E. Germanaite, K. Zaleckis, R. Butleris, and Lopata, A. “General Spatial Pattern and Meta-Pattern Model for Problems That Need Analytical Approach in Complex Spatial Systems,” Applied Sciences, vol. 12, no. 1, p. 302, 2021.

[21] V. S. Tynchenko, V. V. Kukartsev, V. V. Tynchenko, E. A. Chzhan, and L. N. Korpacheva, “Automation of monitoring and management of

conveyor shop oil-pumping station of coal industry enterprise,” IOP Conference Series: Earth and Environmental Science, vol. 194, no. 2, p. 022044, Nov. 2018.

- [22] A. V. Milov, V. S. Tynchenko, V. V. Kukartsev, V. V. Tynchenko, and V. V. Bukhtoyarov, “Use of artificial neural networks to correct non-standard errors of measuring instruments when creating integral joints,” Journal of Physics: Conference Series, vol. 1118, no. 1, p. 012037, Dec. 2018.

IoT-Based Cyber-Physical Distribution System Planning

1st Shaben Kayambo

Centre for Intelligent Systems (CIS)
School of Engineering and Technology
Central Queensland University
Rockhampton, Australia
shaben.kayambo@cqumail.com

2nd Biplob Ray

Centre for Intelligent Systems (CIS)
School of Engineering and Technology
Central Queensland University
Rockhampton, Australia
b.ray@cqu.edu.au

3rd Narottam Das

Centre for Intelligent Systems (CIS)
School of Engineering and Technology
Central Queensland University
Rockhampton, Australia
n.das@cqu.edu.au

4th Mary Tom

Centre for Intelligent Systems (CIS)
School of Engineering and Technology
Central Queensland University
Rockhampton, Australia
m.tom@cqu.edu.au

Abstract—Recently, there has been a large-scale integration of renewable sources to keep pace with the enormous increment of consumers' load demand. The integration of these devices on a large scale can violate the technical and regulatory constraints of the smart power grid. Therefore, a proper control system must be designed to increase their power output and maximize the system owners' benefit while meeting power grid standards. This paper is a collection of information to provide its reader with a holistic view of the implementation of a cyber-physical system to aid in distribution system planning. The goal of this paper is to review the current literature and identify research gaps that have yet to be fully explored. This is followed by making proposals for areas to be researched in order to improve the knowledge bank surrounding this important topic. This paper will begin by addressing the concept of a cyber-physical system by elaborating on its structure through its various layers. It will also describe the role of the Internet of Things (IoT) in improving the monitoring and control abilities of cyber-physical distribution systems. Furthermore, a detailed literature review is given to highlight the current state of distribution system research being done surrounding the implementation of cyber-physical systems to improve system planning. Based on this review, an analysis is given to better explain how a successful cyber-physical based distribution system can be constructed to achieve automation. Finally, gaps in current research are discussed to provide an understanding on which aspects of cyber-physical based distribution system planning have not been fully explored yet and where more exploration needs to be done. A list of research questions are provided to help aid potential researchers dive deeper into the area of fully automating distribution systems that consist of both renewable and non-renewable sources of energy.

I. INTRODUCTION

A. Cyber-Physical System (CPS)

In recent times, there has been a continuous effort to develop appropriate techniques so that the monitoring and control of the physical system can be improved. On the other hand, researchers are also focusing on cyber system-related issues to improve the performance of communication and computing systems [1]. Nowadays, new technology has evolved in the form of a Cyber-Physical System (CPS) that integrates both the physical and cyber systems [1]-[2]. This can be defined as a multi-dimensional system with integrated computing and communication technology (cyber system) in order to monitor the stability, reliability, and efficiency of physical systems. Using different types of sensors, the cyber system collects data from the physical system and sends the control signal to the physical system so that appropriate action can be taken [2]. The working of a cyber-physical system is presented in Figure 1.

B. Need for Cyber-Physical Power Distribution System (CP-PDS)

As of late, with the enormous load growth and rising environmental concerns, there has been a large-scale integration of renewable energy sources all over the world [3]. These devices provide significant benefits in terms of voltage profile enhancement, load reduction, pollution minimization and benefits maximization. However, the large amount of renewable energy integration with the grid is causing different

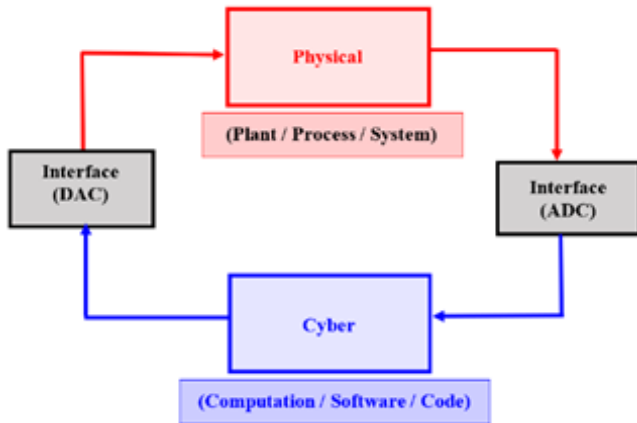


Fig. 1. A sample Cyber-Physical System (CPS).

technical constraint violations such as reverse power flow, voltage stability issue, etc. Apart from changing distribution system characteristics, they also degrade the power quality. These factors change the concept of a distribution system to an active distribution system where a variety of control methods must be adopted to improve energy efficiency and enhance system performance [4]. To do so, the future distribution system must be integrated with computing and communication networks so that the measurement and control instructions can be transmitted to and from the control center [5]. This type of active system management and control cannot be attained without integrating information and communication technology (ICT). ICT is important for timely and accurate decision-making with the utilization of an active control function. However, the random failure of the ICT system may lead to adverse situations such as blackouts, deterioration of system performance, etc. In such context, to maintain the efficient and secure operation of the power systems, the physical distribution system must be integrated with the cyber system, which can be regarded as a Cyber-Physical Power Distribution System (CPPDS) [5]-[6].

C. Structure and Function of Cyber-Physical Distribution Systems

A cyber-physical power distribution system (CPPDS) consists of two parts, namely the cyber system and the physical power system [7]. There are different types of traditional system equipment and new-age technologies such as renewable energy sources, energy storage elements, and electric vehicles in physical power systems. On the other hand, the cyber system consists of communication networks, metering instruments, controllers and computing devices. The main function of this cyber system is to automatically operate the physical power system. A typical structure of CPPDS can be divided into mainly three layers [7].

1) *Physical Layer*: In this layer, there are distribution system equipment and corresponding cyber components such as different types of renewable sources, intelligent electronic devices, battery storage systems, fixed capacitors, circuit breakers, feeder terminal units, manual and remote-control switches etc. Apart from this, this layer also includes the control and measurement units and communication systems. In this layer, the acquisition of measuring data and the control commands execution is achieved through data interfaces between primary equipment and secondary devices.

2) *Network Layer*: A two-way data communication between the control center and the physical system is necessary for stable operation, fast malfunction response, and renewable energy sources management [8]. This layer describes a distributed communication network that exhibits the transmission process of mass data using various network components. The power line carrier networks work through coupling capacitors, carrier equipment, power lines, and modems. The Ethernet networks work using switches, optimal fibers, cable, routers etc., whereas the wireless networks work through mobile communication and WIMAX technologies [9].

3) *Control Layer*: The control layer uniformly collects all the transmitted data from different communication systems and accordingly generates control demands and sends them to the physical layer in order to respond to different physical operation scenarios. This layer consists of multiple sub-servers to transfer traffic from different communication networks. After analyzing the received data, this layer acquires the current physical state of the distribution system and generates control demand for each piece of equipment of the physical layer.

4) *Operation Procedure*: Initially, local metering units and sensors measure the physical operating state of the distribution system, which includes bus voltages, renewable sources output, load values, and the on/off status of automation equipment. Then, this physical layer converts these physical states into digital states before sending them to the control layer through multiple communication networks. Based on the received information, the control layers determine the operational state of the distribution system and generate a series of control commands. These commands are sent back to the physical layer to be executed by system equipment for regulating the unsuitable physical state of the distribution system.

D. Role of IoT in the operation of Distribution System

To combat the challenges of the traditional distribution grid, the concept of the smart distribution grid has come into existence where complete coordination can be established with a two-way flow of information and electricity between the consumers and the power suppliers under a fully automated power network. However, one of the major concerns with the smart grid implementation is the increased use of information and computation technologies, which rely on the Internet as well as computing and processing power to run. Moreover, the requirements of response time and reliability are especially critical, as the entire system operation has to be carried out automatically and to do so, the data transmission and decision-making technologies shall be optimized. In order for the smooth process of all the necessary data, IoT-based modern information communication technologies play an important role in improving the monitoring and control abilities of the cyber-physical distribution systems.

II. COMPREHENSIVENESS

A detailed review of the existing literature in this field reveals that in recent times, the planning of IoT-based cyber-physical systems is gaining researchers' attention and, in most articles, researchers concentrated on the modelling of system equipment and improving reliability with adequate control techniques. In [10], researchers designed an accurate cyber-physical system with different sets of communication, control and physical system equipment under a smart grid environment. They also generated several attack scenarios to explore cyber-physical impacts on the voltage profile and rotor stability. Authors in [11] presented a comprehensive survey on the cyber-physical system considering the smart grid implementation. They discussed the development scope by providing taxonomy and insightful guidelines. Moreover, they also identified the key features and different design decisions. From the perspective of industry automation, researchers presented a detailed review of the IoT-based cyber-physical system in [12]. The design and characteristics of IoT architecture and IoT applications were discussed, and the authors have mainly considered the three key aspects of IoT systems, namely, networking, computing and control. Authors in [13] designed a cyber-physical system-based navigation system for fulfilling electric vehicle charging demands. The power distribution system was taken as the physical layer and interaction between the electric vehicle and control center is studied to assist the vehicle owner with the proper charging station. A cyber-physical system was studied in [14] for application in active distribution systems where an uncertain load control strategy was proposed based on the output values of renewable energy sources. This paper also discussed the modelling of control systems and control techniques considering a hybrid system. As of late,

there is a growing number of installations of distributed energy resources in the power grid, which could impact the grid operation through coordinated attacks. In this regard, the authors in [15] designed an IoT-based cyber-physical grid model and presented a risk assessment methodology to analyze the impact of IoT integrated cyber systems on physical grid operations. In [16], a secure control system was designed for identifying the possible cyber-attack in the communication channel of a cyber-physical system. In [17], the authors proposed an analytical method to quantify the impact of networks and information systems on the reliability analysis of cyber-physical distribution systems. Regarding the assessment of reliability, the impact of element failures, network topologies and information network traffic was taken into account. Researchers in [18] designed a cyber-physical distribution system to deploy feeder remote terminal units with an aim to increase system reliability. In [19], the scheduling problem of packet transmission over wireless communication networks was studied to widen the stability margin. In this regard, the authors proposed a new algorithm for scheduling the distributed data traffic. Authors in [20] designed an IoT-based cyber-physical system to handle industrial informatics regarding location, sensor and unstructured data for big data mining.

III. ANALYSIS

The cyber-physical system consists of a control layer and a physical layer. In between these layers, there is a network layer through which the information flow occurs through different communication channels. With the integration of different devices in the physical system, there is a significant increment of data transferring between the control and physical layers, which heightens the risk of cyber attacks. Therefore, special concentration must be given to the development of IoT-based smart computation, control and communication techniques for identifying possible cyber-attacks in the communication channel so that the reliability, safety, and secure operation of the cyber-physical system can be ensured.

A cyber-physical system consists of networking computing and communication devices. The random failure of computing software, communication network link and intelligent services can influence the control ability of the distribution system. On the other hand, to integrate a dependable i.e. failure-free control system, distribution companies have to compromise in the economic benefits since such system demands high construction and maintenance costs. In such a context, the failure rate of cyber components must be considered for cyber-physical planning of distribution systems. Moreover, the economic, reliability and safety aspects of the cyber-physical system must be taken into account for realistic planning. Apart from the failure of the cyber system,

there are quite a few pieces of equipment in the physical system and there is a high chance that they can also fail. Therefore, the uncertainty in physical system failures also needs to be considered for system planning.

Under the circumstances of a smart grid, the integration of renewable energy sources is one of the main aspects to reduce the distribution system grid dependency. Lately, there has been an acceleration in the integration of renewable sources in the system, which increases various technical and regulatory constraints and violations. This is basically due to the intermittency in load demand and renewable sources output. If the renewable power generation is more when the load demand is less, there may be a possibility of over voltage. On the other hand, if renewable energy generation is less when the load demand is more, there will be an unhealthy voltage profile. Therefore, the system operators must address the proper energy management technique, which can be accomplished only with real-time measurement and control. In this regard, the IoT-based advanced information and communication technologies must be utilized to assist the system operator to efficiently control the generation-demand mismatch by ensuring a two-way information flow.

Under any type of system disturbance, the system must act immediately so that the adverse effect of the disturbance can be minimized and accordingly, a system blackout can be prevented. In order to achieve this, the cyber-physical distribution system must be equipped with automation devices such as a phasor management unit (PMU), remote-controlled switch (RCS), circuit breaker (CB), automatic recloser (AR), and feeder terminal unit (FTU). Moreover, IoT technologies must be utilised, which enable a two-way communication flow to make the system fully automated.

IV. RESEARCH GAPS AND QUESTIONS

Following a detailed literature review, it is worth mentioning that researchers have done great work and made a significant contribution in the advancement of this field. However, there are some areas which are important from the cyber-physical distribution system planning point of view which are yet to be explored. In this regard, the research gaps of existing literature are summarised in the following section:

- Researchers mostly concentrated on the modelling of system equipment, improving reliability with control techniques. But recently, with the integration of various devices in the physical system, there is an increase in data transfers through the communication network, which increases the risk of cyber-attacks. In this regard, advanced IoT-based computing, control, and communication technologies must be used that can detect the possible cyber-attack and accordingly maximize system reliability and safety.
 - In most literature, researchers concentrated on only the modelling of cyber-physical system equipment. However, there may be a random failure of computing software, communication network link and intelligent services, which can influence the control ability of the distribution system. Therefore, the uncertainty in cyber system failure must be considered regarding realistic planning. However, researchers mostly neglected this factor.
 - In existing articles, the increment of the penetration level of renewable energy sources was accomplished with the integration of the DFACTS device. However, this is a costly solution that may not be affordable for the distribution system owner. In this regard, IoT-based control methods must be applied which can efficiently handle the uncertainty of renewable energy by ensuring two-way information and power flow between the supplier and consumers. Thus, it will help to increase the renewable penetration in the physical distribution system without affecting grid standards.
 - The physical distribution system must be converted into a fully automatic system so that the adverse effect of sudden disturbances can be prevented. However, in the existing literature, the physical distribution system was mostly converted into a semi-automatic system. The recent development of the smart distribution grid demands the system to be equipped with the automation devices such as PMU, RCS, CB, AR, and FTU to achieve the highest level of reliability and IoT technologies must be used to develop a two-way communication flow.
- The research questions are summarised below:
- Can the utilization of smart computing and control technologies identify possible cyber-attacks and take preventive measures to ensure safety and reliability?
 - Will the consideration of intermittency in cyber-failure result in a practical cyber-physical distribution system planning?
 - Will the IoT-based control methods increase the penetration level of renewable sources? If so, will the planning solution be economically feasible for the distribution system owner?
 - What will be the impact on system reliability with the installation of automation devices in the physical distribution system?

V. PROPOSED WORK

To keep pace with the enormous load growth, there is an integration of a large number of renewable energy sources. In such context, it is essential to configure these renewable energy systems in such a way so that their power generation capacity is increased and there is a maximization of the economic benefits of the power grid while meeting the

grid security requirements. To accomplish this, a proper control system must be designed. Therefore, regarding the execution of planning activities, distribution system planners must consider the following major directions:

1) To increase the penetration level of the renewable energy sources for coping with the load growth and improving the financial benefit received by the system owner.

2) To integrate various automation devices so that the overall reliability of the system can be significantly improved.

3) To convert the physical distribution system into a cyber-physical distribution system with the integration of IoT technologies.

In this regard, a planning method is proposed for the cyber-physical distribution system that will consider the incorporation of different types of renewable energy sources under a cyber-physical environment. The main objective of this planning study will be to meet the enormous load growth in an efficient way considering the financial benefit maximization. Moreover, different types of automation devices will be equipped with the distribution system so that there will be an improvement in the fault management procedures, which ultimately increase the system's reliability. On the other hand, under the smart grid environment, there is an integration of new dynamics loads in the distribution system and therefore, there will be a huge amount of data interchanging between the control layer and physical layer. It increases the chances of cyber-attacks in the communication channel of the cyber-physical system. In such context, an IoT-based advanced and secure control system will be designed for identifying and taking measured steps to defend against possible cyber-attacks. Moreover, IoT-based control technologies will also be utilized to increase the penetration level of renewable energy sources.

In the first step, solar energy systems, wind energy systems, battery storage systems, and different types of automation/protective devices will be integrated with the physical system. Then, in the next step, different types of devices will be introduced in the cyber system and an investigation will be carried out to find the effect of the cyber system on the operation of the physical system. Since the optimal location, capacity and quantity are important parameters to be appropriately chosen. In such a context, a suitable soft computation technique will be utilized to find out the capacity, types, location, and quantity of different devices. The random failure of the cyber-physical system equipment as well as the uncertainty associated with the non-

renewable and renewable power output will also be taken into consideration to derive the optimal planning solution. The popular Monte Carlo Simulation can be utilized to check the voltage stability issues due to cyber failures and large renewable penetration.

REFERENCES

- [1] E. F. Orumwense and K. Abo-Al-Ez, "A systematic review to aligning research paths: Energy cyber-physical systems," *Cogent Engineering*, vol. 6, no. 1, pp. 1-21, Dec. 2019.
- [2] L. Shi, Q. Dai, and Y. Ni, "Cyber physical interactions in power systems: A review of models, methods, and applications," *International Journal of Electrical Power System Research*, vol. 163, pp. 396-412, Oct. 2018.
- [3] H.B. Tolabi, M. Ali, and M. Rizwan, "Simultaneous Reconfiguration, optimal placement of DSTATCOM, and photovoltaic array in a Distribution system based on Fuzzy-ACO Approach", *IEEE transactions on Sustainable Energy*, vol. 6, no. 1, pp. 210-218, 2014.
- [4] C. D'Adamo, S. Jupe, and C. Abbey, "Global survey on planning and operation of active distribution networks—Update of CIGRE C6.11 working group activities," in *Proc. 20th International Conference in Exhibition of Electrical Distribution*, 2009, pp. 1-4.
- [5] S. Suryanarayanan, R. Roche, and T. M. Hansen, "Cyber-physical-social systems and constructs in electric power engineering," *Institute of Engineering Technology, London, U.K., Tech. Rep.*, 2016.
- [6] B. Falahati and Y. Fu, "A study on interdependencies of cyber-power networks in smart grid applications," in *Proceeding of IEEE PES Innovation in Smart Grid Technology*, Washington, DC, USA, 2012, pp. 1-8.
- [7] W. Liu, Q. Gong, H. Han, Z. Wang, and L. Wang, "Reliability modeling and evaluation of active cyber physical distribution system," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7096-7108, Nov. 2018.
- [8] J. Gao, Y. Xiao, J. Liu, W. Liang, and C. L. P. Chen, "A survey of communication/networking in smart grids," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 391-404, Feb. 2012.
- [9] S. W. Lai and G. G. Messier, "Using the wireless and PLC channels for diversity," *IEEE Transactions on Communications*, vol. 60, no. 12, pp. 3865-3875, Dec. 2012.
- [10] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu, "Cyberphysical security testbeds: Architecture, application, and evaluation for smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 847-855, Jun. 2013.
- [11] M.H. Cintuglu, O.A. Mohammed, K. Akkaya, and A.S. Uluogac, "A Survey on Smart Grid Cyber-Physical System Testbeds", *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. , 446-464, 2017
- [12] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A Survey on Industrial Internet of Things: A Cyber-Physical Systems Perspective", *IEEE Access*, vol. 6, pp. 78238-78259, 2018.
- [13] C. Wang, D. Wu, H. Zeng, and V. Centeno, "CPS based electric vehicle charging directing system design in smart grid," in *Proceeding IEEE Innovative Smart Grid Technology Conference*, Minneapolis, MN, USA, Sep. 2016, pp. 1-5.
- [14] Y. Wang, D. Liu, and Q.-S. Li, "A hybrid system based CPS model and control of loads in active distribution network," in *Proceeding IEEE International Conference on Power System Technology*, Wollongong, NSW, Australia, Sep./Oct. 2016, pp. 1-8.
- [15] D.J.S. Cardenas, A. Hahn, and C. Liu, "Assessing Cyber-Physical Risks of IoT-Based Energy Devices in Grid Operations", *IEEE Access*, vol. 8, pp. 61161-61173, 2020.
- [16] F. Battisti, G. Bernieri, M. Carli, M. Lopardo, and F. Pascucci, "Detecting integrity attacks in IoT-based Cyber Physical Systems: a case study on Hydra testbed", *2018 Global Internet of Things Summit (GloTS)*, Bilbao, Spain, pp.1-6, 2018.

- [17] W. Liu, Q. Gong, H. Han, Z. Wang, and L. Wang, "Reliability modeling and evaluation of active cyber physical distribution system," *IEEE Transactions on Power System*, vol. 33, no. 6, pp. 7096–7108, Nov. 2018.
- [18] S. Wang, D. Liang, L. Ge, and X. Wang, "Analytical FRTU deployment approach for reliability improvement of integrated cyber-physical distribution systems," *IET Generation Transmission and Distribution*, vol. 10, no. 11, pp. 2631–2639, Aug. 2016.
- [19] C. Qu, W. Chen, J. B. Song, and H. Li, "Distributed data traffic scheduling with awareness of dynamics state in cyber physical systems with application in smart grid," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2895–2905, Nov. 2015.
- [20] C.K.M. Lee, C.L. Yeung, and M.N. Cheng, "Research on IoT Based Cyber Physical System for Industrial Big Data Analytics", *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, pp. 1-6, 2016.

The Million Improvised Electric Rickshaws in Bangladesh: Preliminary Survey and Analysis

Shahriar Khan
Dept of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

Mushfiq Uz Zaman Chowdhury
Dept of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
1721959@iub.edu.bd

Mostakin Rabbi
Dept of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
1822142@iub.edu.bd

Asma Khatun,
Dept of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh

Abstract—Electric vehicles have had great success in the US, as seen from the rapid rise of Tesla cars and its pioneer Elon Musk. But there is an ongoing less known parallel revolution in Bangladesh, with the rise of about a million electric rickshaws. A preliminary survey of these electric rickshaws was conducted, and their stakeholders were interviewed. These rickshaws are basically a home-grown solution to the local transportation problem, and have developed without any central organized support or design. In view of their new technology and questionable stability, they don't yet have official recognition, and are permitted only in side roads and in rural areas. Several different designs were encountered in the streets of Dhaka city. The main problem is that their high speeds are not what the rickshaw chassis was designed for, which makes them prone to accidents. They are driven by four lead-acid batteries in series with an inverter supplying a brushless DC motor. This paper recommends enforcing speed limits, lowering seats, installing coiled spring suspensions, and adding brakes on rear wheels, all of which are likely to improve stability and safety. As the electric rickshaw is contributing greatly to the economy and convenience of the people, it is suggested that the vehicles be accommodated phase-by-phase within new rules and regulations. The preferred strategy would be to work with the industry, rather than be prescriptive and impose difficult and expensive standards for their construction and operation.

Keywords—electric vehicles, EV, Bangladesh, motor, BLDC motor, lead acid, battery, power electronics, inverter, standards, regulations, survey, recommendation, center of gravity, suspension.

I. INTRODUCTION

The Electric Vehicle (EV) has been a major success in the US, as evidenced by the success of manufacturer Tesla, and its pioneer Elon Musk. However, there is a parallel ongoing revolution in Bangladesh, where more than a million rickshaws have become powered as electric rickshaws, changing the living standards of the people and the economy of the country.

A. Proliferation of Electric Rickshaws

The older pedal-driven rickshaws were slow and depended on the strength and health of the driver. But they are being fast replaced by electric rickshaws, less tiring to drivers, and keeping with the worldwide trend towards

electric vehicles. The descriptive term *electric rickshaws* is preferred in this paper, than the less-descriptive terms *Easy bikes*, *Auto rickshaws*, *Battery-run Three Wheelers (BRTWs)* or *Battery rickshaws*.



Fig. 1. An electric rickshaw with a motor assembly attached to a pedal driven rickshaw

Because of the new technology and their questionable safety record, the electric rickshaws do not have official recognition, operating papers, or road permits. But they have become popular in rural areas and among the side roads of the capital city of *Dhaka*. The million electric rickshaws are a silent revolution, contributing to incomes nationwide, convenience to passengers, jobs for the drivers, and revenue for the economy. The world is moving towards electric vehicles, and sooner or later, the country is likely to move towards modification, acceptance and integration of electric rickshaws.

A preliminary survey and analysis of electric rickshaws was conducted, so as to find possible areas of improvement, to make suggestions for recognition by authorities, and for bringing them under standard rules and regulations.

II. LITERATURE REVIEW

The pedal driven rickshaw, popular in the country for at least 70 years, was mostly made by small entrepreneurs, and mostly did not follow any central or optimized design. Recommendations by academia for an improved design had been mostly ignored by the industry.

The first electric rickshaws in the country were being reported as early as 2009 [1]. More recently, they have been analyzed and documented in the literature [2,3,4,5]. Better control of the vehicles [6,7], better charging stations [8], and powering with solar energy [9,10] have been proposed. New power electronic drives for their motors have been proposed [11].

Electric rickshaws and their positive impacts on the country been reported in newspaper articles [12]. According to a 2017 report, "more than 5,00,000 battery-operated auto-rickshaws operate across the country, consuming 450 megawatts of electricity per day, and are being charged through residential connections [13]. From 2013 to 2018, the local battery market went up from 3000 crore to 8000 crores, largely from the growth of EV rickshaws [14].

In 2017, a joint study by *Rahimafrooz* and the *Asia Foundation* showed that EV rickshaws "currently transport 2.5 crore passengers in both townships and rural areas every day and has created jobs for 30 lakh people." According to Prof. Ejaz Hossain of BUET, easy bikes have expanded massively without any government intervention, but there is no vision and control. He said that the government needs to formulate a policy to allow the sustained growth of electric rickshaws [15].

A June 20, 2021 report announced a decision to ban electric rickshaws for their poor braking and poor safety record [16] At least 13,000 illegal motor-run rickshaws and vans have been destroyed so far. A few days later, a number of reports suggested the ban should be reconsidered [17,18].

The socio-economic impact of electric rickshaws have been studied [19,20]. The theory behind the motor and power electronics have been discussed in text books [21,22].

III. METHODOLOGY

A. Survey and Interviews

With the goal of investigation, interviews were conducted of stakeholders in the industry, including passengers, drivers, owners, technicians, etc. Their numbers were as follows:

- (a) Electric rickshaw drivers: 10 persons
- (b) Technicians: 5
- (c) Owners (*Mohajons*): 20
- (d) Importers, assembling shops: 2
- (e) Battery distributors: 1
- (f) Battery scrap shop: 3-4 staff
- (g) Office peon: 1

- (h) People in authority: 4-5
- (i) Resident of Rajshahi: 1

The questions mainly explored their operating locations, how they are made, types of batteries used, their safety record, the experience of passengers and drivers, etc.

The questions were mostly informal and conversational, without a set script and a set list of questions, and were mostly related to their professional connection with electric rickshaws. When told that the intentions of this study was to investigate and promote electric rickshaws, they were quite cooperative.

The survey included locations outside of *Dhaka*, and at 14 places in the *Dhaka City Corporation*. The *Dhaka* locations included *Uttara, Dakkhinkhan, Khilkhet, Ashulia, Jigatola, Rayer Bazaar*, etc.

IV. CONSTRUCTION

It was found from the surveys, that the electrical and motor assemblies are being designed and assembled by local technicians with less in-depth knowledge on the subject.

Most electric vehicles have a set of four 12 V lead-acid batteries, and an inverter for DC to three-phase conversion powering a Brushless DC motor with maximum speed of 3000 rpm (figure below). The handle is rotated for controlling the speed.

All electrical components were found to be imported, except for the batteries which were made in Bangladesh or China. At about tk 15,000 for a set of four, the batteries are the most expensive component.

The total cost for the electric components is about tk 20,000 (US\$ 220). In addition, there is the added cost of mechanical changes to the chassis, such as adapting the rear axle for fitting with the motor (figure 3).

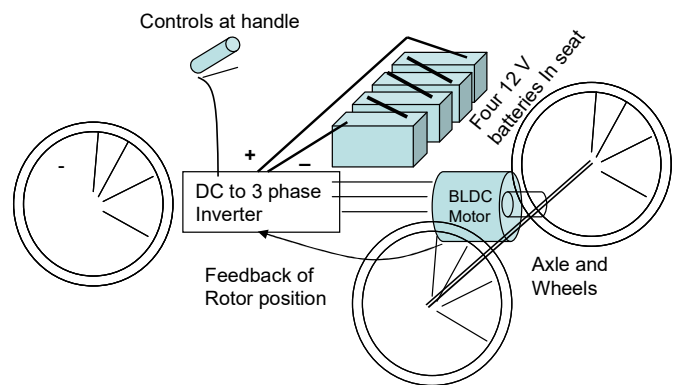


Fig. 2. Electrical components and assembly of the EV rickshaw.

TABLE I. COST OF ELECTRICAL COMPONENTS OF ELECTRIC RICKSHAWS

Component	Cost (in taka)
Batteries (set of four, locally made)	15,000 (total)
BLDC motor (imported)	2,000
Power electronics and hand control (imported)	2,000
Recharger (imported)	1,500
Total	20,500

According to the five technicians surveyed, the operation and maintenance of electric rickshaws are rather simple.

A. Comparison with Electric cars.

The Electric rickshaw costs about US\$ 500, compared to the US\$ 50,000 for the electric car in the US (factor of 100). The Lithium ion batteries are a major expense for electric cars, but they are prohibitively expensive for electric rickshaws, and are mostly not even being considered as an option.

V. MECHANICAL ISSUES

Different areas of Bangladesh were found to have different design and constructions for the rickshaw body. Less skilled technicians are doing the mechanical constructions differently, without the assistance of uniform standards or professional designs. The constructions were mostly improvised, following the designs being handed down for years.

The chassis is basically made by welding of steel pipes, the standards of which are not properly regulated. There have been reports of the electric rickshaws breaking into two when being driven at high speeds.

The motor to rear-axle connection for two different types of electric rickshaws in Dhaka city are shown below.



Fig. 3. A common electric rickshaw (seen in Dhaka city), with the motor (in left, gold) connected to a modified rear axle



Fig 4. An older chain driven model (seen in Dhaka city) with the motor (in left, gold) connected with a chain to the axle.

A. Speeds

The rickshaws had originally been designed to be pedal driven at low speeds, and not motor-driven at higher speeds. No rickshaw, electric or pedal-driven, is equipped with a speedometer. The maximum speed of an electric rickshaw is about twice that of a pedal-rickshaw. Speeds can be greater in wide open paved roads with less obstructive traffic, such as in rural areas. In the main roads of Dhaka, speeds can become high late at night, when there is less traffic. Speeds as high as 25 mph (40 km/h) have been observed in the main roads of Dhaka city, and in rural areas.

B. High seats

The major complaint about EV rickshaws is that they overturn too easily, causing injury to driver and passengers. This draws attention to their high seats.

The problem with present pedal rickshaws and especially electric rickshaws is height of the seats (about 3 ft. 2 in above ground) which increases likelihood of overturning during turns. High seats also increases wind resistance, implying greater power consumption and shorter battery life.

C. Hard Suspension

The suspension of most rickshaws is not on coiled springs, but on two pieces of curved steel bars, appearing to be mostly inflexible (figure below). A ride is often hard that any spring effect can hardly be felt by passengers. An unyielding suspension means the whole chassis will rotate, as a rear wheel hits a pothole or a bump, increasing the likelihood of overturning.



Fig. 5. The suspension is the circular iron behind the wheel and above the axle.

D. Brakes

The complaints of poor braking are quite justified, as most rickshaws, electric and pedal-driven, only have brakes on the front wheel. According to drivers interviewed, it is possible to have brakes attached to the rear wheels at some extra cost.

E. Better Designs

There are some better constructed rickshaws with lower seating and softer spring suspensions (figures below). Costing tk 55,000, just tk 5,000 more than the common electric rickshaws, these better designs are still in the minority (US\$ 1 = BD tk. 90)



Fig. 6. An electric rickshaw, with lower seats and spring suspension on the roads of Dhaka



Fig. 7. Spring suspensions in a better constructed electric rickshaw. The motor and attachment with the rear axle is also visible.

A. Characteristics of the Roads

Electric rickshaws are better adapted to bitumen-paved roads, which can sustain their relatively-high speeds. They do ply on brick-paved roads, but are subject to vibration because of the uneven surface. In rural areas, pathways may not be paved at all (plain soil), and the ride may be uncomfortable or impossible, especially in rainy weather. In all cases, the lack of an effective suspension makes the ride quite bumpy.

VI. BATTERIES AND CHARGING

According to those interviewed, the charging time for the four 12 V batteries was about 6 hours, which led to an operating time of about 4 hours with about 125 km of travel.

Charging is often done in the slots of 10 pm - 6 am and 1 pm - 4 pm, because this is when there is less demand for transportation. The cost was about tk. 20 per hour of charging. This translates to about tk. 100 per day and tk 3000 per month.

As described by technicians, the battery set lasts for about 8-12 months, after which they must be replaced.

It was observed that most of the recharging stations were not properly documented, and it was unclear how the electricity service providers were paid.

A. Recycling of batteries

Everyday, thousands of used-up batteries are replaced by new batteries. According to 3-4 people interviewed at battery scrap shops, the plates inside the battery are recycled into new batteries. There appears to be some syndication in the scrap industry, meaning it may be controlled by a small number of people, and newcomers find it difficult to enter the business.



Fig. 8: The inside of a battery scrap shop, where the batteries and the lead plates are visible.

VII. LOCATIONS WHERE PERMITTED

In the streets of capital city of *Dhaka*, EV rickshaws are generally found in narrow roads, and in the suburbs. They are not permitted in the main areas of *Dhanmondi, Bashundhara, Uttara, Banani, Gulshan, etc.* They are not permitted in the wide main roads and highways of the city before about 11:00 pm, after which they may be found in small numbers.

Electric rickshaws run mainly on small streets in the areas of *Pallibi, Darus Salam, Turag, Dakkhin Khan, Uttar Khan, Banasree, Azimpur, Kamrangirchar, Zingatola, Basabo-Madartek, Jatrabari, Kadamali, Demra, Jurain, Shyampur, Mohammadpur, Adabor, Rampura, Khilgaon, Sipahibag, Badda, Meradia, Moghbazar, Modhubagh, Mohakhali Wireless Gate* and *Uttara*.

In much of the rural areas, most of the pedal rickshaws have been replaced by electric rickshaws. According to preliminary reports, electric rickshaws are popular in all districts, The authors have seen these vehicles in *Bogra, Rangpur, Gazipur, Comilla, Jhinaidaho, and Magura*. Those interviewed were in *Dhaka, Gazipur, Rangpur, Magura, and Jhinaidaho*.

According to residents of the district of *Rajshahi*, about 85 % of rickshaws in the city areas, and 70 % of rickshaws in rural areas are electric.

VIII. ECONOMICS

Electric rickshaws are creating a silent revolution in the transportation sector, and have become popular among lower and middle-income groups, and in areas where there are few other means of transport. Low fare was a major reason for the popularity of the electric rickshaw. Fares for a typical trip range from tk 10 - 50. According to the owners (*Mohajons*) and the drivers, their purchase cost is about tk. 50,000. The daily rent was tk 400 in Dhaka city and tk. 200 in rural areas, payable by the driver to the owner (*Mohajon*).

TABLE 2. COMPARISON OF PEDAL RICKSHAW WITH ELECTRIC RICKSHAW

(US\$ 1 = BD tk. 90)	Human-powered rickshaw	Electric rickshaw
Cost	tk 10,000	tk 50,000
License	tk 10,000	- NA -
Daily rent and parking in Dhaka	tk 200	tk 400
Daily rent and parking, rural areas	tk 100	tk 200

The cost per day for charging and parking is tk 100. After paying the daily rent as 200, the earnings left over for the driver per day is in the range of tk. 800-1000.

The drivers was found to be working about 10 hours per day, making a total of 40 trips, with a total mileage of about 100 km. In comparison, human-powered rickshaws are slow and expensive for the passengers. Their earnings are less at tk. 600 - 900 per day.

A. Employment for the workforce

Although employment has improved greatly over the last few decades, there is still some shortage of job opportunities, especially in the rural areas. From our survey and the interviews, it was found that the electric rickshaw has produced many job opportunities. During the pandemic, some under-age youth became drivers, owing to the closure of the schools.

While interviewing the mayor of Magura municipality. it was found that crime rate has reduced significantly owing to improved standard of living of the people.

During our interviews, it was found that the electric rickshaws were a major source of employment in agriculture-dependent areas, with no nearby industries. For migrating workers, driving the vehicles was a secondary income source and agriculture was the primary income source. These drivers had families and agricultural activities in the home in rural areas. During the time of harvesting and crop cultivation they left the cities and go back to their villages, leading to a shortage of EV-rickshaw drivers, and consequently high fares in the city.

The availability of drivers was affected by the availability of jobs in nearby industries. Workers laid off in a garments factory would become drivers for EV rickshaws, leading to an over-supply of transportation, more competition among the drivers, and lower fare for the customers.

Similarly, when a factory re-starts its operation by employing workers, there is a shortage of drivers, and rickshaw fares go up.

B. Impact on Environment:

Although the vehicles are zero emission, they are charged by power stations running on gas and coal. They do not directly pollute the air, but move the pollution to the fossil-fueled power stations.

IX. RECOMMENDATIONS

Regardless of the problems with EV rickshaws, they are a convenient mode of transportation and have brought prosperity to the country, But the pedal-driven rickshaw chassis is designed to be powered by humans at low speeds, and is not ready for higher-speed operation with motors. Recommendations are now made as to how to best face the challenges for electric rickshaws in the coming years.

A. Strategy of Working with the EV rickshaw Industry

A prescribed product that is imposed upon manufacturers is unlikely to be accepted without opposition by manufacturers, especially if it leads to greater expense. It is not feasible to mandate nationwide a single design and construction prescribed by a central authority.

Small entrepreneurs tend not to innovate when the profit margin is too low. They tend to be risk-averse small businesses and low income people. But helping in new product development by academia can overcome these barriers to modernization.

Technical support may be provided to manufacturers in the country, so that the academicians work with the manufacturers and not impose a central, apparently optimized design.

To safely promote this technology in the country, the authorities may choose to:

(a) Enforce speed limits for the electric rickshaw, so as to limit speed, the greatest cause of accidents.

(b) Mandate vehicle construction requirements for strength of chassis and to improve their stability and comfort. A maximum seat height may be imposed to improve stability and reduce drag.

(c) Mandate brakes on the rear wheels (mostly not present at this time)

(d) Mandate an effective coiled spring suspension (replacing present curved bars) for greater comfort, and improved stability when going over a bump or pot hole. Coiled springs and low seats are already found in some better designed electric rickshaws.

(e) To register and license the vehicles and drivers, phase by phase, across the country, bringing them under government rules and regulations.

(f) To better document the charging stations of electric rickshaws, so that the utility companies are properly paid through proper metering. Lithium ion batteries are not discussed in the industry owing to their much greater expense.

(g) Require drivers to undergo some hours of training, and then a take a test, just like for a vehicle driver's license. This is because drivers may be moving at relatively high speeds side by side with cars and heavy vehicles,

X. CONCLUSION AND FUTURE WORK

Electric vehicles like Tesla have become a booming industry and are in the news, but there is a parallel ongoing silent revolution in Bangladesh with electric rickshaws. Although these rickshaws are providing convenience in transportation and contributing to the economy, they lack official recognition, have no licensing, and have not been brought under standard rules and regulations.

The surveys and interviews conducted by this paper identified some simple bottlenecks and problems, such as excessive speeds, no brakes on the rear wheels, high seats and center of gravity, hard suspension and a weak structure. The motor allows high speeds, which the rickshaw was not originally designed for. Different areas of Bangladesh have different models of electric vehicles without any standard design. Lithium ion batteries are much lighter and popular for electric cars in the US, but are not discussed for electric rickshaws because of their much greater expense.

Based on the survey, recommendations are made about how to best face the challenges for EVs in the country in the coming years. Maximum speed limits can be enforced, the seats and center of gravity can be lowered, and softer suspension may be implemented with springs. Lower seating and better spring suspensions are already being seen in some rickshaws.

This industry should be nurtured by cooperation between authorities, manufacturers, owners, drivers, and electric utilities, This will ensure that electric vehicles are yet another innovative step towards continued development of the country.

REFERENCES

- [1] R. Sarker, Electric rickshaws in Rangpur. The Daily Star, August 17, 2009.
- [2] Laboni Sarker, S. R. Jaynab, S. Khan, "The Undocumented Electric Vehicles in Bangladesh," National Conference on Electronics and Informatics-2019, at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019,
- [3] Laboni Sarker, Safia Rahman Jaynab, S. Khan, "Study on the Undocumented Electric Vehicles in Bangladesh," International Journal of Industrial Electronics and Electrical Engineering, ISSN(p): 2347-6982, ISSN(e): 2349-204X, Volume-8, Issue-3, Mar.-2020.
- [4] M. S. Z. Chowdhury, Maisha Anjum, Zannatul. Mawa, Shahriar Khan, "Future of Electrical Vehicles in Bangladesh," National Conference on Electronics and Informatics-2019, , at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019.
- [5] M. R. Ahmed and A. K. Karmaker, "Challenges for Electric Vehicle Adoption in Bangladesh," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-6,
- [6] A. Mohammad, M. A. Abedin and M. Z. R. Khan, "Microcontroller based control system for electric vehicle," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 693-696,
- [7] Rezanul Haque, S. Khan, "The Modified Proportional Integral Controller for the BLDC Motor and Electric Vehicle," IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, Toronto, Canada, 21-24 April, 2021.

- [8] A.S.M.M.Hasan, "Electric Rickshaw Charging Stations as Distributed Energy Storages for Integrating Intermittent Renewable Energy Sources: A Case of Bangladesh" *Energies* 13, no. 22: 6119, 2020.
- [9] A. Mallik, M. A. Arefin, F. Rashid, & Asfaquzzaman, 2017, "Solar Based Plugged-in Hybrid Engine Driven rickshaw (Auto-Rickshaw) & its Feasibility Analysis for Bangladesh." *International Conference on Mechanical, Industrial and Materials Engineering (ICMIME2017)*, Dec. 2017.
- [10] K. S. Reddy, S. Aravindhan, & T. K. Mallick, "Techno-Economic Investigation of Solar Powered Electric Auto-Rickshaw for a Sustainable Transport System," *Energies*,10(6), 2017.
- [11] A. A. Mamun, S. Arfin, S. Khan, "High Gain DC-DC Converter for Three-Wheeler Electric Vehicles," *IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, Toronto, Canada, 21-24 April, 2021.
- [12] "Bangladesh on the road to an electric future," *The Daily Star*, Wednesday, April 27, 2022 |
- [13] M. R. Rasel, "PDB for bringing the battery-run vehicle under traffic," *Dhaka Tribune*, September 27, 2017.
- [14] J. Chakma, "Automotive battery market revving up." *The Daily Star*, April 26, 2018.
- [15] "Easy bikes outgrowing limitations," *The Daily Star*, Sept. 19, 2017.
- [16] "Government to ban battery-run rickshaws, vans" *The Dhaka Tribune*, June 20th, 2021
- [17] Tanim Asjad, "Rethinking ban on battery-run vehicles," *The Financial Express*, June 25, 2021.
- [18] Shahin Akhter, "Battery-run rickshaws, vans continue to run defying ban amid restrictions," *The New Age*, July 06, 2021.
- [19] M. A. Rahim, M. U. H. Joardder, S, M. N. Hoque, M. M. Rahman, N. H. Sumon, "Socio-economic & environmental impacts of battery-driven auto-rickshaw at Rajshahi city in Bangladesh." *International Conference on Mechanical industrial and Energy Engineering*, 2012.
- [20] M. S. Rana, F. Hossain, S. S. Roy, S. K. Mitra, "Battery Operated Auto-rickshaw and Its Role in Urban Income and Employment-Generation," *International Journal of Advancements in Research & Technology*, 1(5), 2012.
- [21] Shahriar Khan, *Semiconductor Devices and Technology*, Third Edition, ISBN: 978-094-33-5983-4, by S. Khan, Dhaka, Bangladesh, June 3, 2018.
- [22] Shahriar Khan, *Electrical Energy Systems*, Fourth Edition, ISBN: 978-984-33-7638-1, by S. Khan, Dhaka, Bangladesh, Feb 2021.

The USB Powered Miniature Tesla coil, with Filament bulb, Fluorescent lamp and Discharge to Body

Simoom Rahman
 Dept. of EEE
 Independent University,
 Bangladesh
 Dhaka, Bangladesh
 2120115@iub.edu.bd

Shahriar Khan
 Dept. of EEE
 Independent University,
 Bangladesh
 Dhaka, Bangladesh
 skhan@iub.edu.bd

Abstract:--First invented in 1891, the Tesla coil supplied high voltage at low current, demonstrating spectacular feet-long arcing discharge. The Tesla coil has been used more for demonstration and entertainment, and less for teaching and research, partly because of its inherent dangers. Today's diodes, transistors and microprocessors allow low voltage Tesla coils with new capabilities and improved safety. Still, there are few commercial applications, and the Tesla coil remains mostly for demonstration and exhibition. The improvised miniature 5 V USB port supplied Tesla coil built by the author is worthy of investigation and documentation because of its exotic phenomena like lighting up a nearby fluorescent lamp, making a common filament bulb act like a plasma ball, and producing an imperceptible continuous discharge on the finger. The fluorescent lamp lights up not from the RF from the coil, but from the induced currents in the mercury vapor. The filament bulb produces moving plasma, proving that conditions exist for such interesting phenomena. The visible continuous arc at the finger, far from giving a shock, was imperceptible because of its RF frequencies. The constructed miniature Tesla coil illustrates boosting of voltage, the air-core transformer, resonance at tuned frequency, and other electrical principles. With some simple precautions, this easily constructed low-voltage, Tesla coil with \$20 of components, shows much promise for promoting teaching and research at schools, colleges and universities.

Keywords:-- Tesla, coil, fluorescent, lamp, arc, discharge, filament, bulb, plasma, ball, air core, transformer, RF, UV..

I. INTRODUCTION (HEADING 1)

First invented by Nikolai Tesla in 1891, the high voltage Tesla coil impressed audiences, and inspired wireless energy transfer and long distance RF communication. After the invention of solid-state devices in the 1960s, new opportunities and adaptations with the Tesla coil have become possible. The dangers of its high voltages have been mostly overcome with low voltage supply, and miniaturization with solid state diodes, transistors, capacitors, microprocessors, etc. Today, the Tesla coil is used to impress and entertain, with the decorative Plasma globe, music by modulated streamers, etc.

Mainly because of its few applications, and the dangers of its high voltages, the Tesla coil is less used in academic curriculums and research. Today's commercial applications are not as great as Tesla had hoped for, but there are great

opportunities for using the low-voltage, safe Tesla coil as a teaching aid, and for motivating and inspiring the younger generation.

A. Construction for this study

The low-voltage DC powered Tesla coil was constructed with a bug zapper using theories of resonance, transformers, and with instructions found online. It created moving plasma inside a filament bulb, caused a nearby compact fluorescent lamp to glow, and produced visible but almost imperceptible arcing discharge on the finger.

These exotic phenomena have been demonstrated in the past for larger Tesla coils, but have been less investigated in the literature for smaller Tesla coils. The construction of this study and its associated phenomena merit further description, examination and analysis. The fluorescent lamp glows not because of the effect of the RF on the phosphors, but by the RF-induced current in the vapor in the lamp. The common filament bulb acts like a plasma ball with visible moving plasma, due to some similarities in construction with the plasma ball. The visible continuous discharge on the finger does not hurt as the frequency is much greater than 10 KHz.

The constructed miniature Tesla coil may be used to promote interest, and to demonstrate numerous electrical engineering principles to school, college, and university students.

II. THE TESLA COIL - PAST AND PRESENT

The Tesla coil is a capacitor-tuned oscillator that drives an air-core resonant transformer to produce high voltages at low currents [1,2,3]. A high voltage transformer steps up the AC mains voltage up to a high enough voltage to jump a spark gap acting as a switch at the primary, generating high voltage in the secondary.

The secondary is connected to ground and to a metallic dome. An iron core transformer would have too much losses due to eddy currents and hysteresis. The only available path for the high voltage is to go through the air to the ground as a discharge streamer. Output voltages can range from 50 kV to

millions of volts, at low frequency RF between 50 kHz to 1 MHz.

A. Dangers and Safety

Nikolai Tesla used to impress his audience with spectacular discharges without getting any electric shock, and even glowing in the dark himself.

For a mains-supplied (220 V or 110 v) Tesla coil, every part of the circuit is capable of giving a shock or being fatal. The most dangerous part of the circuit is the intermediate section at thousands of volts, where arcs can jump fatally for severally cm. Owing to charged capacitors, for many minutes, the circuit can give a shock after disconnection from the mains.

A mains-supplied Tesla coil can damage electronics in the room, like smart phones, hearing aids, cardiac pacemakers, etc., and those connected to the same consumer power supply. Discharge from the coil can travel in any direction, inflicting harm to equipment or personnel. Dangerous levels of Ozone may be created in the room.

The coil must be connected to a three-pronged electric plug, with a ground pin, as insufficient grounding may shock nearby personnel, and damage the equipment.

B. Literature Review

Considering that the Tesla coil was invented 130 years ago, there are relatively few publications in the literature. The miniature low-voltage Tesla coil is a relatively new invention, only some decades old. Today's trends are towards miniaturization, with DC supplies and solid state devices [4,5,6]. The streamer discharge from the coil can be modulated with sound, a feature which has been used by the musical group *Arc Attack* [7]. New developments are making it better than ever as a teaching tool [7].

III. CONSTRUCTION

In our construction, the DC supply from the 5 V, 250 mA USB port is fed to the PCB of the Bug Zapper, which is connected to the Tesla coil through capacitors and a spark gap (figures below). The yellow wires at the right were for recharging the batteries and have no function in our circuit.

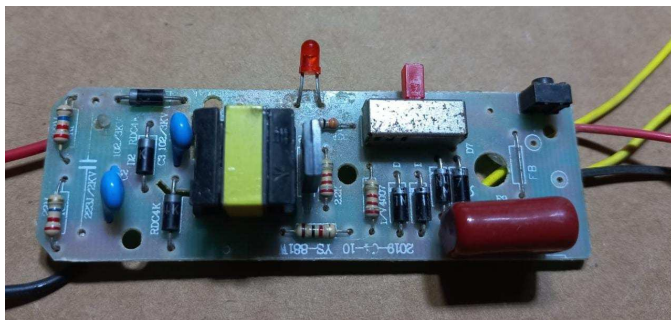


Fig 1. The printed circuit board of the bug zapper, supplying the Tesla coil.

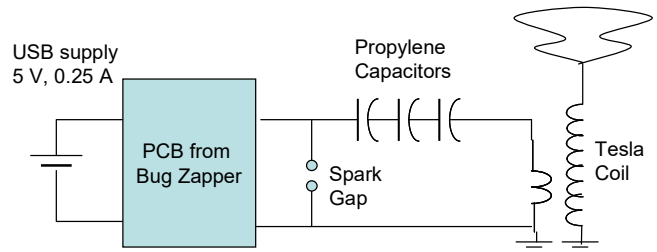


Figure 2. The PCB from the Bug zapper is connected to the Tesla coil through tuning capacitors and a spark gap

The output of the bug zapper (2–2.5 kV) is connected to three propylene capacitors and the Tesla coil.

The metal gas pipes are used as pillars to make the base. The bolts are drilled on the cardboard, which holds the electronics (figure below). The cardboard base was of minimum size so as to fit the bug zapper PCB, spark gap, and polypropylene capacitors. The spark-gap was with 2-inch nuts and bolts.

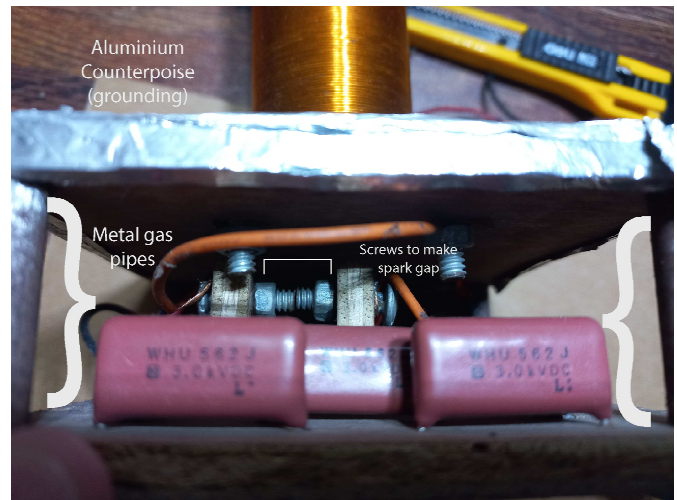


Fig. 3. The cardboard base with Spark gaps and capacitors

A. The Tesla Coil

The Tesla coil was built around a PVC pipe of external diameter 3.8 cm and length 9.5 cm, with the metal drawer knob placed on top. The primary has 1.5 turns, insulated from the primary with electrical tape.



Fig. 4. The Tesla coil had 1.5 turns in the primary and 300 turns in the secondary

The secondary coil has about 300 turns of 0.2 mm enameled copper wire, wrapped around the plastic pipe, and is connected to the metal knob on top and the aluminum foil below for grounding. Without this grounding, no discharge was observed. The 0.2 mm copper wire was sufficient for the current, minimized expense, and fit more wire around the pipe.

B. Components and Costs

The bug zapper racket was the most expensive component at \$ 6, and the total came to about \$ 17. Further details of the components and their costs are given below.

TABLE I. DETAILS OF COMPONENTS AND THEIR COSTS

Component	Remarks	Cost (\$)
PCB from <i>Bug Zapper</i>	Mosquito racquet	6.00
Copper wire for Coil	0.2 mm diameter, 100 ft	3.00
Capacitors	3KV Polypropylene	3.00
Plastic pipe for coil	4" long	1.00
Nuts, Bolts, Aluminium foil, connecting wires		2.00
Card board	6mm wide, 10" x 4"	1.00
Drawer knob	Used as dome	1.00
Total		17.00

C. Further details on Construction

A number of changes and adjustments were needed in order to have a well-functioning coil.

Bug zapper racket are available on the streets from \$ 4.00 but a more expensive \$ 6.00 racket was used for greater reliability.



Fig. 5. The bug zapper racket from which the PCB circuit was taken out..

Early attempts with 6.5 turns in the primary gave a weak arc. The turns in the primary were progressively reduced one by one, with the arc getting stronger each time. The present 1.5 turns in the primary gave the strongest arc.

The polypropylene capacitors in the primary of the coil were available two years ago in the street markets, but are no longer available now. So alternatives must be found for new constructions.

The spark gap was expected to be between 1-3 mm, and was adjusted for the best performance around 1 mm. Its nearby constructions are shown below.

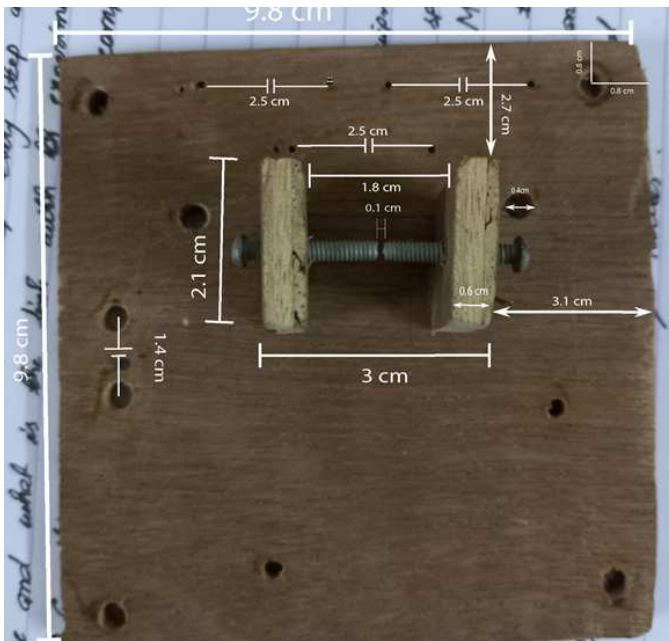


Fig. 6. Dimensions of the spark gap and nearby constructions.

D. The Lamps

The filament bulb used for showing plasma was rated for 25 W, and was 9 cm long and 4 cm wide. The fluorescent lamp was 11 cm long and 3.5 cm. Both lamps were bought from local stores.



Fig. 7. The fluorescent lamp and the filament bulb used in the construction.

IV. EXPERIMENTAL RESULTS

With the circuit turned on, the spark at the spark gap appears as a bright point of light (figure below) (that may be harmful to the eyes over many seconds).



Fig. 8. The arc at the spark gap appears as a bright point light.

A. The Fluorescent Lamp

A compact fluorescent lamp lit up, when held by the glass near our constructed coil. But the phosphorescent coating on the lamp is mainly sensitive to ultra-violet (UV) light, whereas the Tesla coil only produces Radio Frequencies (RF).

The explanation is the currents and magnetic fields in the air are small and insufficient to cause the fluorescent coating to glow. But the high voltage on the dome and the electrostatic field are strong. These induce currents in the vapors in the glass tube, giving off UV light, causing the phosphors to glow. To a young observer, this is amazing considering the low power of the USB supply.

Phosphors on fluorescent lamps are a mixture of rare-earth oxides; mostly Yttrium (Y) oxide, with Europium (Eu), Terbium (Tb), Cerium (Ce) and Lanthanum (La). Their peak sensitivity is at 253.7 nm, which falls within UV.



Fig. 9. Glowing compact fluorescent lamp held by the hand



Fig 10. Glowing compact fluorescent lamp in the dark

B. Discharge Through the body

An arc through the body, as with a *Stun Gun*, is expected to be shocking and painful. Also unpleasant are touching the spark of a piezo-electric lighter, or a stove lighter. But our constructed coil gave a continuous discharge, hardly perceptible to the fingers (figure below).

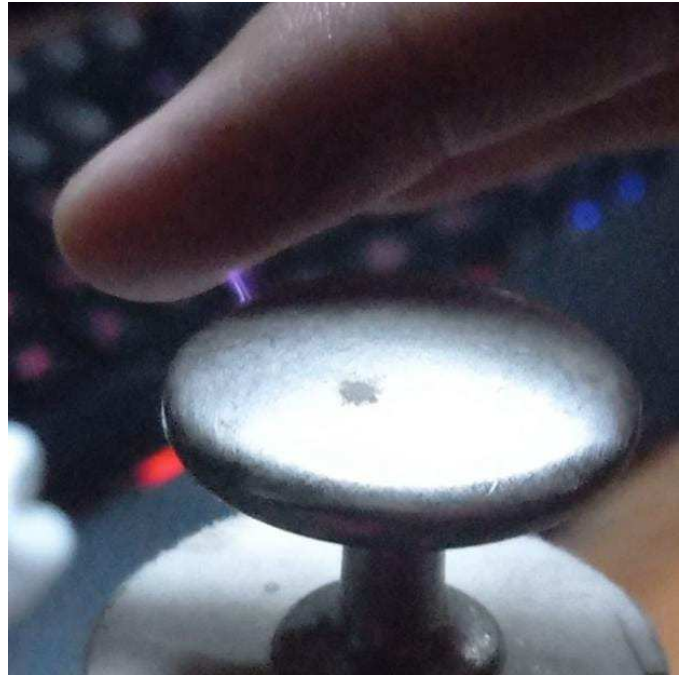


Fig. 11. Continuous steady discharge on finger

Tesla had found that discharges greater than 10 KHz do not give an electric shock, and may be hardly perceptible to the body. Considering the low power of the USB supply, this discharge is likely to be within acceptable safety standards, provided it lasts no longer than a second or two.

C. Moving Plasma in a Lightbulb

When a filament bulb's metal base was touched on our Tesla coil, moving plasma was seen inside the bulb (figure below). With a hand placed on the bulb, the plasma moved with the movement of the hand.



Fig. 12. Moving plasma discharge inside a common lightbulb,

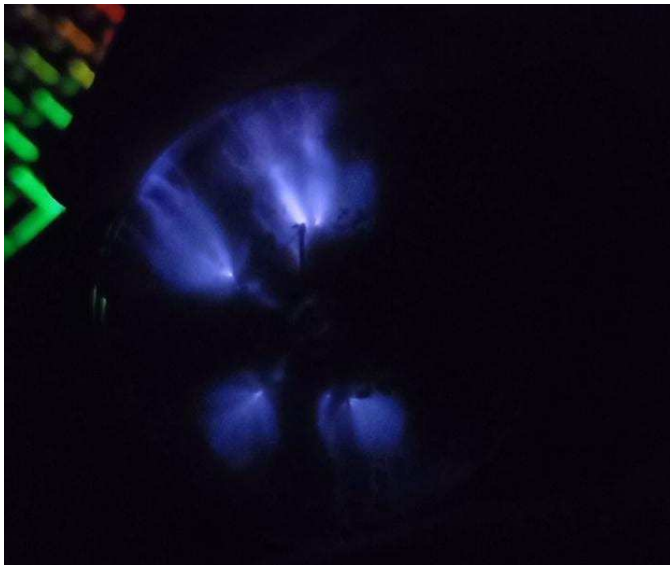


Fig. 13. The moving plasma in the dark.

This is similar to the commercially available decorative plasma ball (also known as plasma globe /sphere). The artistic discharges move with the movement of a hand placed on the glass.

The streaming plasma on our filament bulb placed on our Tesla coil is significant, because of

(a) the low voltage of the Tesla coil

(b) the filament bulb is designed to preserve a white hot tungsten filament, and not to function as a decorative plasma sphere.

(c) The electric light bulb is at slightly less than atmospheric pressure (0.7 atmosphere), with Argon, Nitrogen, Krypton, Xenon, etc. The plasma globe often has neon at close to atmospheric pressure. Argon and Xenon make the plasma more intricate and beautiful in appearance.

V. THE TESLA COIL AS A TEACHING AID

The Tesla coil is suggested here as a teaching aid for a number of reasons:

A. Low cost and Easy Construction

The circuit can be easily constructed by students from easily available inexpensive components, worth less than \$ 20.

B. Inspiration and Motivation

The observed phenomena can be of great interest to students. Advanced students can try to investigate and research the observed phenomena.

C. Safety

The 5 V, 0.25 amp USB supply ensures the output power is too low to cause any serious bodily harm by the arcing discharge or the RF radiation. For added safety, students may approach the coil for only a few seconds at a time.

D. Moving plasma inside a lightbulb

The Tesla coil shows moving plasma inside a filament bulb, which can inspire the student to understand the flow of current through inert gases.

E. Glow of a Fluorescent Lamp

The student will understand that the glow of the fluorescent lamp is not so much from the RF discharge from the coil, but from the currents induced inside the tube, giving off ultra-violet light, making the phosphor coating glow.

F. Barely perceptible visible discharge on the skin.

The student can observe that arcs and discharges on the skin need not give a shock and will be barely perceptible, as they are at Radio frequencies. In comparison, the bug zapper and piezoelectric lighter have a one-time discharge which can give a painful and uncomfortable shock.

VI. DANGERS OF THE SETUP

Any Tesla coil can pose dangers, both known and unknown. When supplied by a 220 V or 110 V, it has dangers which are well known and well established. The dangers of a USB supplied 5 V Tesla coil falls largely in the unknown.

A. Electromagnetic Radiation

The construction can induce a current in the nearby fluorescent lamp, and so it can induce currents in the body, which are not perceptible. Prolonged exposure, for more than a few seconds may be harmful, the details of which are mostly unknown.

B. Discharge through the Finger

The discharge current through the finger is not perceptible, the common explanation for which is that the current remains close to the skin surface because of the skin effect. However, according to the literature, the current still has the potential to cause a small burn on the skin. The precaution to users is to allow discharges for only a few seconds, enough for demonstration, but not enough to cause such a burn.

CONCLUSION

The original Tesla coil of 1891 has undergone great transformation today, with the advent of diodes, transistors, microprocessors etc. The miniature Tesla coil of this study was built with a "bug zapper" PCB, and showed not-so-easily explained phenomena, including moving plasma streamers in a lightbulb, glowing of a compact fluorescent lamp, and an imperceptible continuous visible discharge through the finger. Partial explanations for these phenomena are provided in this paper, but still merit further research.

This circuit can be inexpensively built with a bug-zapper PCB and a coil of wire with about \$20.00, and can be the subject of further study. Further documentation and investigation of this circuit and associated phenomena is only

appropriate, considering its great potential for its usage as a teaching and research tool in schools, colleges and universities.

REFERENCES

- [1] M. Krbal and P. Siuda, "Design and construction solution of laboratory Tesla coil," 2015 16th International Scientific Conference on Electric Power Engineering (EPE), 2015, pp. 311-314,
- [2] V. A. Kolchanova, "Computational modeling of the Tesla coil parameters," Proceedings of the 8th International Scientific and Practical Conference of Students, Post-graduates and Young Scientists Modern Technique and Technologies, 2002. MTT 2002., 2002, pp. 32-33,
- [3] S. Khan, AC Circuits, Third Edition, ISBN 978-984-33-5146-6, Dhaka, Bangladesh, Dec. 2019.
- [4] Donald G. Bruns, "A solid - state low - voltage Tesla coil demonstrator," American Journal of Physics 60, 797 (1992).
- [5] M. B. Farriz, A. Din, A. A. Rahman, M. S. Yahaya and J. M. Herman, "A Simple Design of a Mini Tesla Coil with DC Voltage Input," 2010 International Conference on Electrical and Control Engineering, 2010, pp. 4556-4559.
- [6] S. Khan, Semiconductor Devices and Technology, Third Edition, ISBN: 978-094-33-5983-4, Dhaka, Bangladesh, June 3, 2018.
- [7] C. Ghiliță, S. C. Stegaru, T. Popeea and N. Țăpuș, "Portable audio-modulated Tesla coil for demonstrative actions," 2015 14th RoEduNet International Conference - Networking in Education and Research (RoEduNet NER), 2015, pp. 238-241,

The Improvised Three-Wheeler Electric Vehicle Solutions in Bangladesh: Equipment and Experimental Waveforms

Shahriar Khan
Dept. of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
skhan@iub.edu.bd

Abdullah Hasan Shahriar
Dept. of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
1820193@iub.edu.bd

Amartya Bhakat
Dept. of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
1820890@iub.edu.bd

Murad Prodhan Munna
Dept. of EEE
Independent University,
Bangladesh
Dhaka, Bangladesh
1822015@iub.edu.bd

Abstract— In the past decade or so, there has been a dramatic rise in the number of Three wheeler Electric Vehicles (EV) in Bangladesh. As the technology is new and unknown, these EVs have been largely unrecognized and unregulated by the government. But these EVs are only a part of the worldwide shift towards EVs. Most of the local small three-wheeler EVs are being produced as solutions by low-income small entrepreneurs, who cannot afford much innovation. Past attempts to prescribe the academic "perfect" pedal-rickshaws have been mostly unsuccessful. A better approach may be to provide technical support to this growing "home-made" industry, with the intention of organizing and promoting the growth of EVs. With these goals in mind, selected electrical components of the common electric rickshaw were purchased and assembled in the lab. Preliminary waveforms were obtained, with the intention of obtaining more detailed performance curves at a later date. The technical data can be used for assisting in better designs of commercial EVs in the country. The road, transportations and economic conditions for Bangladesh are applicable for many Asian countries, meaning that the electric vehicle technologies developed in this paper may be transferred to other countries as well.

Keywords—electric vehicle, EV, rickshaw, electric rickshaw, Bangladesh, experimental, waveforms, three-wheeler, waveforms, entrepreneurs.

I. INTRODUCTION (HEADING 1)

Over the last decade or so, there has been a dramatic rise in electric vehicles in Bangladesh. Owing to the new and unfamiliar technology, these electric rickshaws have been largely unrecognized and unregulated by the government. But this proliferation in electric rickshaws is only the local manifestation of the worldwide shift towards electric vehicles. This industry should be rightfully nurtured, explored, recognized and regulated.

Local EVs range from the common three-wheeler electric rickshaw to larger three-wheeler 4, 6, and 8 seaters, named *Tom Toms* or *Easy bikes*. The two-wheeler electric rickshaws are small enough to have option to be pedal-driven, and so can be called a Hybrid Electric Vehicle.

Most of the electric rickshaws are being produced by low-income small entrepreneurs, who are not equipped for, or cannot afford much innovation. Past attempts to prescribe the academically "perfect" pedal-rickshaw have been mostly unsuccessful, and not adapted by the manufacturers. Rather, a better approach may be to work with the small-entrepreneur manufacturers, and nurture the growing industry through technical support. Towards these goals, selected electrical components of a local electric vehicle were assembled in the lab, with the intention of testing them, and overcoming bottlenecks holding back their progress.

II. LITERATURE REVIEW

A wide range of electric vehicles, ranging from the hover board to heavy trucks, are gaining popularity worldwide, and have been the subject of studies [1]. DC motors are being replaced by AC motors, with power electronics to interface with batteries, which bring advantages in weight, efficiency and increased range. The induction motor was used for electric vehicles, but the Brushless DC motor (BLDC) has the lightest weight and highest efficiency, and is the most popular for EVs in Bangladesh.

The first electric rickshaws in the country were being reported as early as 2009 [2]. More recently, they have been analyzed and documented in the literature [3,4,5,6].

Better control of EVs [7,8], better charging stations [9], and powering with solar energy [10,11] have been proposed. New power electronic drives for their motors have been proposed [12].

Electric rickshaws and their positive impacts on the country been reported in newspaper articles [12]. According to a 2017 report, "more than 5,00,000 battery-operated auto-rickshaws operate across the country, consuming 450 megawatts of electricity per day, and are being charged through residential connections [14]. From 2013 to 2018, the local battery market went up from tk 30 to tk 80 billion (US\$ 300 - 890 million), largely from the growth of EV rickshaws [15]. (assuming US\$ 1 = BT tk. 90)

A 2017 study showed that EV rickshaws "currently transport 25 million passengers in both townships and rural areas every day and have created jobs for 3 million people." According to a prominent professor, easy bikes (a type of EV) have expanded greatly without any government intervention, but there is little vision and control. The government needs to formulate a policy to allow the sustained growth of electric rickshaws [16].

The socio-economic impact of electric rickshaws have been studied [17,18]. The theory behind the motor and power electronics have been discussed in text books [19,20]. Papers aimed at better organization of the growth of electric vehicles have been submitted for publication [21,22].

III. MOTORS

The first local electric rickshaws from a decade ago, had the rear-axle driven by a chain connected to a motor. This older chain model is seen less often nowadays, and the differential speed motor has gained popularity.

A. 48 V Differential speed motor

This differential speed motor for the electric rickshaw features "a high precision motorhead with 16 tooth and anti-rust aluminum shell. Compliance with specifications allows this motor to be "mixed and matched" with other electrical components such as the power electronics drive.

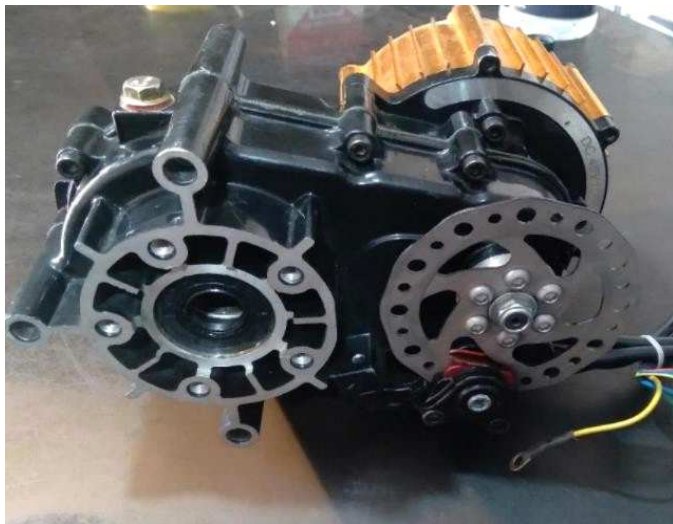


Fig. 1. The Differential speed motor is very popular for new electric rickshaws.

Further specifications are given below for a smaller and a larger motor.

TABLE 1. SPECIFICATIONS OF A SMALLER MOTOR

Power	1000 Watt
Voltage	48 V
Rated speed	3000 RPM
Max speed	30-35 KM/H

TABLE 2. SPECIFICATIONS OF A LARGER MOTOR

Power	12000 Watt
Voltage	48 V
Rated speed	450 rpm/min
Max speed	25-30 km/h
Amp	70

B. The 48 V chain-driven motor

The older chain driven motor and its specifications are given below. The addition of a gear head to a motor reduces the speed while increasing the torque.



Fig. 2. The chain-driven motor setup popular in earlier electric rickshaws

TABLE 3. SPECIFICATIONS OF THE 48 V 1 KW MOTOR

Power	1000 Watt
Voltage	48 V
Rotted speed	480 rpm/min
Max speed	25-30 KM/H
Amp	10.8

C. Larger motor for 4-8 passenger EV

Larger BLDC motors are used for heavier 4-8 seater EVs, as seen below. Higher voltages are used with 5 - 6 batteries kept in series. The specifications are also shown.



Fig 3. The larger BLDC motor used for 4-6 seater local EVs



Fig. 4. The larger BLDC motor attached to the chassis of an 8-seater EV.

TABLE 4. SPECIFICATIONS OF THE 60 /72 V, 3 KW MOTOR

Power	3000 Watt
Voltage	60 V/ 72 V
Rotted speed	600 rpm/min
Max speed	35-40 KM/H
Amp	10.8 A

D. Another 60 V BLDC motor

This different type of motor was found to be quite reliable too as they usually last for more than 3-4 years. The motor usually cost 3500-4000 tk in the local market.



Fig. 5 Another type of BLDC motor

TABLE 5. SPECIFICATIONS OF THE 60 V, 1 KW MOTOR

Power	1000 Watt
Voltage	60 V
Speed	2500 rpm

IV. SUPPORTING ELECTRICAL EQUIPMENT

A. Batteries

Numerous types of EV Batteries are available in the local market. According to the surveyed technical personnel, batteries with warranty battery last longer: one year before replacement.. Batteries without warranty are shorter lasting: 4-

6 months. According to our survey, most rickshaws use batteries without warranty.

According to those surveyed, only a few manufacturers make good batteries. Most batteries have performance levels well below the published values. Batteries made in China are reported to be cheaper and more popular those made in Bangladesh.



Fig 6. Batteries used in electric vehicles in the country. Most batteries are reported to perform below their published values.



Fig 7. Most electric vehicles in the country use four batteries connected in series.

B. Chargers

Various chargers were seen in the market, as seen below.



Fig. 8 . A commonly used charger



Fig. 9. Chargers, as seen in the field.



Fig. 10. An electric rickshaw being recharged

According to preliminary estimates, the efficiency of charging is about 83 % . This means it takes about 7.2 kWhr to recharge 5 batteries of 6 kWhr.

C. Controllers

Controllers change the DC supply from the batteries (usually 48 V) to the AC required for the motor. Some controllers encountered in practice are shown below.

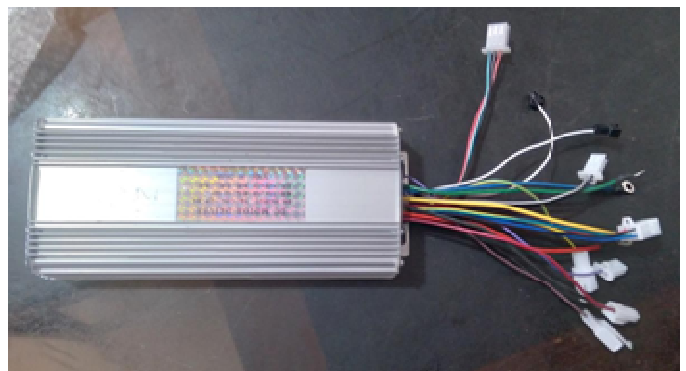


Fig. 1. A controller about to be assembled into an EV

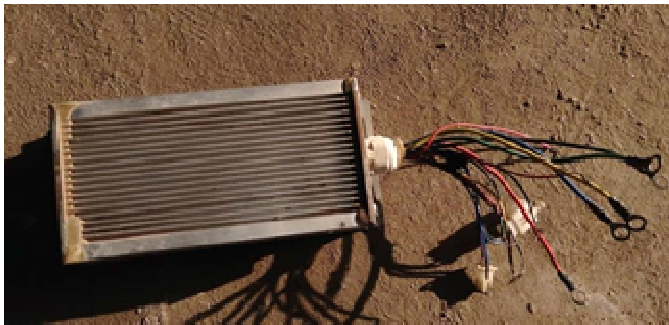


Fig. 11. A controller removed from an EV

V. EXPERIMENTAL SETUP

A. Components

Following the findings of equipment in the field, a typical BLDC motor, a controller, a speed control handle and a lighting, display and key panel were put together for the experimentation. The BLDC motor (figure below) was rated at 48 V, 800 W (name not legible as it was in a foreign language).



Fig. 12. The BLDC motor used in the experiment, rated at 48 V, 800 W.

The controller used (figure below) was of Model *Yinpin*, and rated for 48V, 800 W, 35A \pm 1.



Fig. 13. The controller *Yinpin*, rated for 48 V, 35A \pm 1, 800 W.

The display panel had the headlight and key socket attached (figure below).



Fig. 14. the headlights with the attached key socket.



Fig. 15. The display panel, headlights and key socket are attached together.

B. Cost of equipment

The total costs of the equipment was about BDT 17,500, which came to \$ 194 at an exchange rate of US\$ 1 = BDT 90 (table below).

TABLE 6. COSTS OF EQUIPMENT

Components	BD Taka	US\$
Batteries	12,000	133.00
BLDC motor	3,500	39.00
Controller + Display + Light + Key slot + Speed controller	2,000	22.00
Total	17,500	194

C. Connections

The experimental equipment were connected as shown below.



Fig 16. The batteries in series with the power electronics and the BLDC motor



Fig. 17. Authors Abdullah Shahriar and Amartya Bhakat are seen with the experimental setup

VI. EXPERIMENTAL RESULTS

As expected turning the hand controller varied the output voltage to the motor and the speed from 0 to maximum.

TABLE 7. TWISTI ANGLE OF HAND CONTROLLER VARIED THE OUTPUT VOLTAGE AND THE SPEED.

Twist angle of Hand controller	Output voltage (across motor)
0	33.4 mV
10 %	245.2 mV
20 %	5.6 V
30 %	14.3 V
40%	16.3 V
50%	20.21 V
60 %	26.01 V
65%	29.4 V
70%	32.5 V
75%	33.7 V
80%	35.3 V
85%	37.5V
90%	38.7 V
95%	39.1 V
100%	41.5 V

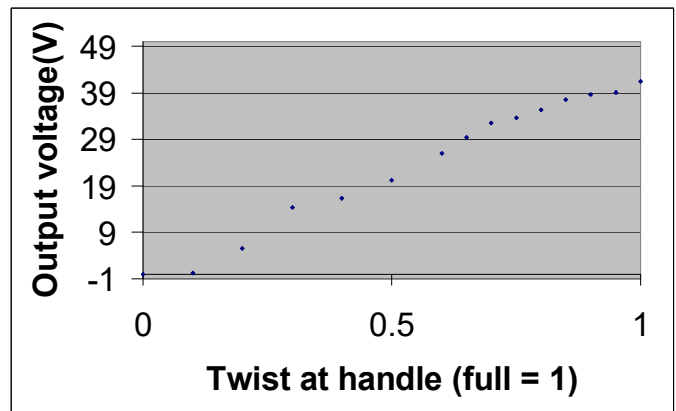


Fig. Increase in voltage at motor, with the angular twist at the handle

The output waveforms from the controller to the motor, for various positions of the hand controller, with corresponding speeds is shown below.

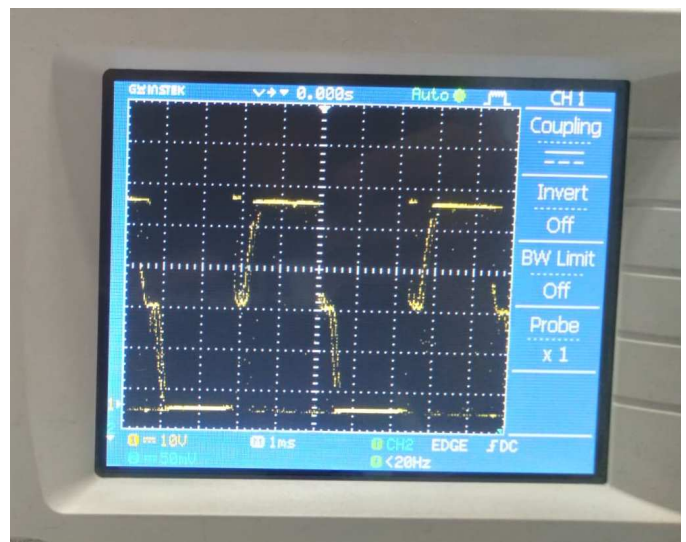


Fig. 18. The output voltage of the controller fed to the motor.

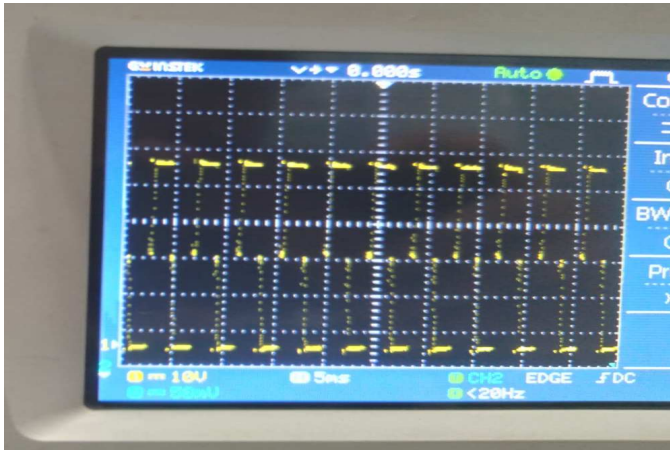


Fig. 19. The output voltage of the controller fed to the motor for a different speed. .

VII. CONCLUSION

With the intention of organizing and promoting the growth of electric rickshaws in the country, a brief survey of the electrical assembly was conducted. Selected equipment were purchased and put together, as for a commercial electric rickshaw. Preliminary waveforms were obtained, with the intention of obtaining more detailed performance curves at a later date. These can be used for assisting in better designs of commercial electric rickshaws.

A limitation of the existing setup was that no significant mechanical load was attached to the motor. A mechanical load will be attached in future, attempting to replicate the EV mechanical load as much as possible. The motor behavior will be simulated at a later date with available parameters, and compared to the observed experimental waveforms.

REFERENCES

[1] M. Weiss, K. C. Cloos, E. Helmers, "Energy efficiency trade-offs in small to large electric vehicles." *Environ Sci Eur* 32, 46 (2020).

[2] R. Sarker, Electric rickshaws in Rangpur. *The Daily Star*, August 17, 2009.

[3] Laboni Sarker, S. R. Jaynab, S. Khan, "The Undocumented Electric Vehicles in Bangladesh," National Conference on Electronics and Informatics-2019, at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019,

[4] Laboni Sarker, Safia Rahman Jaynab, S. Khan, "Study on the Undocumented Electric Vehicles in Bangladesh," *International Journal of Industrial Electronics and Electrical Engineering*, ISSN(p): 2347-6982, ISSN(e): 2349-204X, Volume-8, Issue-3, Mar.-2020.

[5] M. S. Z. Chowdhury, Maisha Anjum, Zannatul. Mawa, Shahriar Khan, "Future of Electrical Vehicles in Bangladesh," National Conference on

Electronics and Informatics-2019, , at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019.

[6] M. R. Ahmed and A. K. Karmaker, "Challenges for Electric Vehicle Adoption in Bangladesh," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-6,

[7] A. Mohammad, M. A. Abedin and M. Z. R. Khan, "Microcontroller based control system for electric vehicle," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 693-696,

[8] Rezanul Haque, S. Khan, "The Modified Proportional Integral Controller for the BLDC Motor and Electric Vehicle," IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, Toronto, Canada, 21-24 April, 2021.

[9] A.S.M.M.Hasan, "Electric Rickshaw Charging Stations as Distributed Energy Storages for Integrating Intermittent Renewable Energy Sources: A Case of Bangladesh" *Energies* 13, no. 22: 6119, 2020.

[10] A. Mallik, M. A. Arefin, F. Rashid, & Asfaquzzaman, 2017, "Solar Based Plugged-in Hybrid Engine Driven rickshaw (Auto-Rickshaw) & its Feasibility Analysis for Bangladesh." International Conference on Mechanical, Industrial and Materials Engineering (ICMIME2017), Dec. 2017.

[11] K. S. Reddy, S. Aravindhan, & T. K. Mallick, "Techno-Economic Investigation of Solar Powered Electric Auto-Rickshaw for a Sustainable Transport System," *Energies*,10(6), 2017.

[12] A. A. Mamun, S. Arfin, S. Khan, "High Gain DC-DC Converter for Three-Wheeler Electric Vehicles," IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, Toronto, Canada, 21-24 April, 2021.

[13] "Bangladesh on the road to an electric future," *The Daily Star*, Wednesday, April 27, 2022 |

[14] M. R. Rasel, "PDB for bringing the battery-run vehicle under traffic," *Dhaka Tribune*, September 27, 2017.

[15] J. Chakma, "Automotive battery market revving up." *The Daily Star*, April 26, 2018.

[16] "Easy bikes outgrowing limitations," *The Daily Star*, Sept. 19, 2017.

[17] M. A. Rahim, M. U. H. Joardder, S. M. N. Hoque, M. M. Rahman, N. H. Sumon, "Socio-economic & environmental impacts of battery-driven auto-rickshaw at Rajshahi city in Bangladesh." International Conference on Mechanical industrial and Energy Engineering, 2012.

[18] M. S. Rana, F. Hossain, S. S. Roy, S. K. Mitra, "Battery Operated Auto-rickshaw and Its Role in Urban Income and Employment-Generation," *International Journal of Advancements in Research & Technology*, 1(5), 2012.

[19] S. Khan, *Semiconductor Devices and Technology*, Third Edition, ISBN: 978-094-33-5983-4, by S. Khan, Dhaka, Bangladesh, June 3, 2018.

[20] S. Khan, *Electrical Energy Systems*, Fourth Edition, ISBN: 978-984-33-7638-1, by S. Khan, Dhaka, Bangladesh, Feb 2021.

[21] S. Khan, M. Zaman, M. Rabbi, A. Khatun "The Million Improvised Electric Rickshaws in Bangladesh; Preliminary Survey and Analysis," Accepted for publication in IEMTRONICS 2022, Toronto, Canada, to be held on June 1-4, 2022.

[22] S. Khan, A. H. Shahriar, A. Bhakat, M. P. Munna " The Small Three-Wheeler Electric Vehicle Solutions in Bangladesh; Survey, Progress and Documentation" Accepted for publication in IEMTRONICS 2022, Toronto, Canada to be held on June 1-4, 2022.

The Small 3-Wheeler Electric Vehicle Solutions in Bangladesh; Survey, Progress and Documentation

Shahriar Khan
Dept. of EEE
Independent University,
Bangladesh
 Dhaka, Bangladesh
 skhan@iub.edu.bd

Abdullah Hasan Shahriar
Dept of EEE
Independent University,
Bangladesh
 Dhaka, Bangladesh
 1820193@iub.edu.bd

Amartya Bhakat
Dept. of EEE
Independent University,
Bangladesh
 Dhaka, Bangladesh
 1820890@iub.edu.bd

Murad Prodhan Munna
Dept. of EEE
Independent University,
Bangladesh
 Dhaka, Bangladesh
 1822015@iub.edu.bd

Abstract— The ongoing dramatic growth in electric vehicles (EV) worldwide has been accompanied by a parallel growth in small 3-wheeler EVs in Bangladesh. The estimated one million local small electric vehicles means there is about one electric vehicle for every 170 residents of the country. A survey was conducted of the local EV industry, trying to find the status and patterns and trends in the growth. A wide range of vehicles of 2, 4, 6, and 8 seats were seen. An EV in the US can be 20-40 kW with speeds of 120 km/h, whereas a local 3 kW three-wheeler can carry up to 8 passengers at up to 35 km/h. Unlike the US, where Hybrid electric vehicle (HEV) means additional IC engine, in the local context, HEV means the option to be human-powered or solar powered. The two-seater electric rickshaws had option to be pedal driven. Amazingly, solar power supplemented EVs are already being used for 8-seaters, showing the great advances made by the local industry. These EVs have grown without central organizational support or official recognition. Rather they are solutions by low-income entrepreneur manufacturers in response to the special needs of the local market. A problem was that vehicles are all known by different names all over the country. The terms Tom Tom and Easy bike were used interchangeably for various vehicles. We found it best to classify local EVs according to their number of seats. Low power (may be pedal driven) and high power (cannot be pedal driven). The trend towards EVs have been surveyed and documented here for better recognition, regulation and generally for better progress.

Keywords—rickshaw, electric rickshaw, three-wheeler, Bangladesh, experimental, solar power, hybrid, electric vehicle, Tom Tom, Easy bike.

I. INTRODUCTION (HEADING 1)

The worldwide growth of EVs in the last decade, has been accompanied by a parallel dramatic growth in small three-wheeler EVs in Bangladesh. Battery powered three wheeler rickshaws are replacing human-pulled rickshaws at a rapid rate. The estimated 1 million electric vehicles nationally means there is one EV for about every 170 residents of the country.

A survey was conducted of local EVs, trying to find the status and patterns in the growth. The end goal was better documentation, recognition and promotion of the EV industry.

A problem encountered was that different types of electric vehicles are known by different names. An electric rickshaw costs US\$ 600-800, compared to US\$ 200 for a pedal driven

rickshaw. Amazingly solar (plus battery) powered small EVs are already in the market, showing the great advances made by the local EV industry. The growth of EVs may be better regulated, recognized, and benefitted by better organization and support, as has been attempted in this paper.

II. LITERATURE REVIEW

A wide range of electric vehicles ranging from the hover board to heavy trucks are gaining popularity across the world, and have been the subject of studies [1]

The first electric rickshaws in the country were being reported as early as 2009 [2]. More recently, they have been analyzed and documented in the literature [3,4,5,6]. While petrol engines have a maximum efficiency of about 35 %, EVs, which are zero emission, can have efficiencies of 90 %. However, EVs move the emissions from the car to the power station, where the efficiencies can be 60 % for combined cycle power plants.

Better control of EVs [7,8], better charging stations [9], and powering with solar energy [10,11] have been proposed. New power electronic drives for their motors have been proposed [12]

Electric rickshaws and their positive impacts on the country have been reported in newspaper articles [12]. According to a 2017 report, "more than 5,00,000 battery-operated auto-rickshaws operate across the country, consuming 450 megawatts of electricity per day, and are being charged through residential connections [14]. From 2013 to 2018, the local battery market went up from 3000 crore to 8000 crores, largely from the growth of EV rickshaws [15]

A 2017 study showed that EV rickshaws "currently transport 2.5 crore passengers in both townships and rural areas every day and has created jobs for 30 lakh people." A prominent professor said easy bikes have expanded massively without any government intervention, but there is no vision and control. The government needs to formulate a policy to allow the sustained growth of electric rickshaws [16].

A June 20, 2021 report announced a decision to ban electric rickshaws for their poor braking and poor safety record [17] At least 13,000 illegal motor-run rickshaws and vans have

been destroyed so far. A few days later, a number of reports suggested the ban should be reconsidered [18,19].

The socio-economic impact of electric rickshaws have been studied [20,21]. The theory behind the motor and power electronics have been discussed in text books [22,23]. Papers aimed at better organization of the growth of electric vehicles have been submitted for publication [24,25]

III. NEEDS OF THE LOCAL MARKET

An analysis of the EV industry in Bangladesh should first analyze its great differences with the US and Western markets (table below).

EVs in the US are four-wheeled, are large at 20 - 40 kW, whereas Bangladeshi EVs are small and three-wheeled at 1 - 3 kW only.

Hybrid EVs in the US means there is the option to power with the internal combustion engine. In Bangladesh, hybrid EV means the option to be pedal-powered, for the smaller 1 - 1.2 kW electric rickshaws, or solar-powered for the slightly larger vehicles.

The roads and highways in the US are wide and long, stretching for thousands of miles. In contrast, the roads in Bangladesh are narrow and stretch for hundreds of miles. US roads are smooth, paved with bitumen and concrete. In Bangladesh, roads are paved with bitumen, bricks or are unpaved. The maximum speed limit in the US can be 128 km/h, compared to the 30-35 km/h capabilities of most local EVs. The users and riders in the US are mostly high-income whereas they are not so in Bangladesh. Ride-sharing is rare in the US, whereas it is common with the local 4-8 seater EVs.

TABLE . NEEDS OF THE US VS. NEEDS OF THE LOCAL MARKET

	Electric Vehicles in US	Electric Vehicles in Bangladesh
Width of Roads	Wide roads, considering large country	Narrow roads, as country is smaller country size
Length of Roads	Freeway, Highways of Thousands of miles	Highway networks of Hundreds of miles
Quality of Roads	Smooth bitumen and concrete paved roads,	Bitumen, rough brick paved and soil and sandy roads
Speed limits	75 mph limits on freeways and 35 mph on city roads	Much lower speed limits
Users and Riders	Mostly high-income, typical for US	Mostly low-income, typical for a middle income country
Ride-sharing	Ride sharing (mini-busses, micro-buses less common)	Ride sharing common with EVs with capacities of 4 - 8
Who drives	Self - driven	Professional drivers.
Power of vehicle	20 - 40 KW	1 kW to 3.5 kW
Meaning of Hybrid EV	Has supplementary IC engine	Has supplementary pedal-power or solar power

IV. CLASSIFICATION

According to our survey, local EVs range from the popular smaller 2-seater electric rickshaws, to the larger 8-seater Tom Toms used for shared rides from point to point.

A problem is that different vehicles are known by different names by locals, all over the country. The term Easy bike is

used interchangeably for small and large EVs. The term Tom Tom is used interchangeably for larger EVs. We find it best to classify local EVs according to their seating capacity. Lower power 2-seater electric rickshaws may have added pedal drive, and higher power 4-8 seater EVs are too heavy to be pedal driven.

The low power electric rickshaws mainly have local made components, except for the electric motor, controller, and sometimes the batteries. The larger EVs may parts of the chassis imported from abroad (China).

The low-power 1 - 1.2 kW EV may be hybrid, with option to pedal-drive, meaning it be driven even with discharged batteries.

The motor and chassis may be of two types – The older chain driven and newer cog-wheel differential driven now gaining popularity.

The high power EVs have up to 3 kW motor, and are too heavy to be pedal-driven. In case of discharged batteries, they must be towed to the next charging station. A solar panel may be connected for supplemental power. More comparison is provided in a later section.

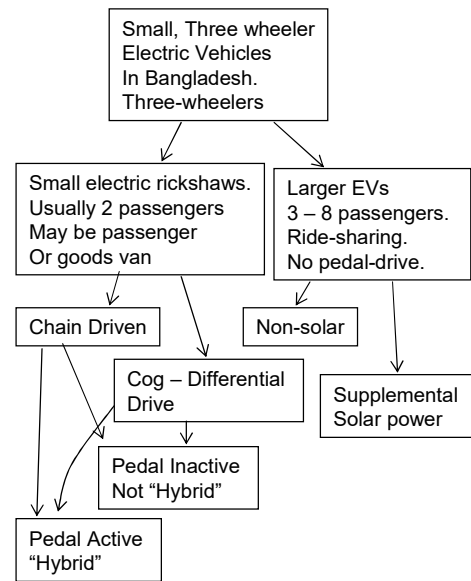


Fig. 1. Classification of local EVs

According to our conducted survey, the following were some types of three-wheeler small electric vehicles available in the country.

- 2-seater Electric Rickshaw.
- Electric Van/Auto Van.
- Electric covered van.
- 6-seater Mishuk Auto /Tom Tom/ Borac .
- 8-seater Mishuk Auto Rickshaw
- 8-seater Solar Electric Auto

These have been described below

V. THE ELECTRIC RICKSHAW

As the name suggests, the electric rickshaw has is basically the 2-passenger pedal-driven rickshaw fitted with the electric motor and accessories. Its cost is in the range of tk 50,000 – 80,000 (compared to tk. 15,000 for a pedal-driven rickshaw). It is the largest in number of all types of EVs in the country. It is seen frequently seen in the districts and rural areas of the country. In Dhaka city, it is seen in the side roads (e.g. *Rayerbazar*) and in the suburbs (e.g. *Shatarkul*).



Fig 2. The electric rickshaw found in the side roads and suburbs of Dhaka, and outside of Dhaka city

According to mechanics and the nameplate, the motor is rated for 1.2 kW, 48 V (four 12 v batteries). The maximum rickshaw speed is 30 -35 km/h. The charging time is 6-8 hours, which allows operation for 8-10 hours or 120 km -140 km. The rated capacity of the battery is 90-120 Amp.hr. Its controller is rated for 48-volt, 1200 W.

A. Construction and Assembly

Electric rickshaws may or may not have the option to be pedal driven (Hybrid EV). For a hybrid electric rickshaw, the pedal driven chain will be attached to the center of the rear axle. The four batteries may be placed in the seats (figure below), or outside, saving space in the seats for other purposes. The motor can be connected with the rear axle either with a chain in earlier models, or a direct differential drive in newer models.

The motor controller can be attached at a convenient place on the body of the rickshaw.



Fig. 3 The differential drive motor, with pedal-driven chain attached to the center of the rear axle.



Fig. 4. The differential drive attached at the center of the axle. There is no pedal-driven chain.



Fig. 5. The older type chain driven rickshaw.



Fig. 6. Another view of the differential drive motor, with the option to pedal-drive. The silver colored controller is seen above, attached to the wooden body.

B. Improved Versions of the Electric Rickshaw

Newer and somewhat improved versions of the electric rickshaw are now being produced in smaller numbers (figure below):



Fig. 7. An alternative version of the electric rickshaw, with lower chassis



Fig. . Another alternative version of the electric rickshaw with lower chassis.

These advantages of the new versions include

(a) they have lower seats and centers of gravity, meaning their likelihood of toppling over and wind resistance are less.

(b) they have softer suspensions, relevant for their higher speeds than the pedal-driven rickshaw. Their softer suspensions allow for a smoother ride and greater stability.

VI. 6-SEATER ELECTRIC VAN/AUTO VAN:

The electric van or Auto van is the electric rickshaw modified to carry up to six passengers. Seen in rural areas, this EV has a questionable safety record.



Fig. 8.. The electric passenger van - rear view



Fig 9. The electric passenger van - front view. There is no chain for driving by pedal.

A. Construction

The lower chassis of the electric van is shown below. In this case, a chain driven axle is seen.



Fig. 10. Chain-driven motor for the electric van. There is no chain for riding by pedal.

VII. ELECTRIC COVERED VAN/AUTO VAN:

The electric covered van is the same as the electric rickshaw but is locally modified to just to carry loads.



Figure 11. The electric covered van, being examined by authors Amartya and Murad .

The observed electric van had a chain-driven motor, and did not have a pedal driven chain (non-hybrid EV).



Fig. 12. Chain-driven motor of electric van, with no option to drive by pedal.

VIII. THE 4-SEATER MISHUK AUTO RICKSHAW

The 4-seater EV, sometimes called the Mishuk Auto Rickshaw is seen below. It is too heavy to be conveniently pedal-driven.



Fig. 13 The Mishuk Auto rickshaw, sometimes known as the Auto rickshaw.

The body appears well designed and constructed, and apparently some components of the chassis are imported (from China).

A. Specifications

The four-passenger Mishuk Auto rickshaw was found to use a motor rated for 1 kw, 48 v (4 x 12 v batteries), 3000 – 3400 rpm. The batteries were rated for 120-200 Amp.hr, and charged for 6 - 8 hours and ran for 8-10 hrs for 120 – 140 km. The max speed is 30 - 35 km/h. The controller was 48-v, 1200 W

B. Visible construction.

The attached motor and controller are shown below.



Fig. 14. Electric motor of the 6-seater Mishuk, as seen from above



Fig. 15 . Electric motor – view from below



Fig. 16. The controller attached to the chassis in the Mishuk Auto Rickshaw

IX. 8-SEATER MISHUK AUTO /TOM TOM/ BORAC AUTO:

The 8-seater Mishuk Auto seen below is also known as Tom Tom or Borac or Easy Bike in different parts of the country. This EV is often seen in rural areas and side roads of Dhaka city.



Fig. 17. The 8-passenger capacity of the Mishuk Auto

A. Specifications

The 8 passenger capacity requires a larger 60v motor. Here five batteries giving 60 v (12 x 5 pieces) are used for the 3 kW motor.

B. Construction

The universal drive connected to the axle is seen below



Fig. 18. The 3 kW motor and drive for the 8-seater Mishuk or Tom Tom.



Fig. 19. The motor connected to the wheels for a Mishuk Auto/ Tom Tom. A well-functioning suspension (yellow) is also visible.

X. THE 8-SEATER SOLAR AUTO

The 8-seater auto or solar Tom Tom is becoming popular in Rajshahi division and other areas the country. The picture below was taken in the *Mugda* area of Dhaka . This is similar to the 8-seater Tom Tom, but with an added solar panel and drive. The solar panel, connected to a charge controller (figure below) in this case increases the running time.



Fig. 20. The 8-seater Tom Tom, with Solar panel attached.



Fig 21. The Charge controller inside the solar Tom Tom

A. Specifications

The motor is rated for 1000 kW, 48 V (12 v 4pieces) with rated speed of 2500-3000 rpm and max speeds of 35-40 km/h. The battery charging time is 6-8 hours, battery capacity is 200 Amp.hr, and the running time is 10-12 hours, allowing a range of 120 – 140 km. The supplemental solar power is 350 w, allowing an extra range of 40-50 km.

XI. COMPARISON OF LOCAL EVS

A comparison of the different types of EVs are seen below. They can have 2, 4, 6 or 8 seats. Most motors are rated 1 – 1.2 kW, with some rated higher at 3 kW.

TABLE. COMPARISON OF VARIOUS TYPES OF EVs IN THE COUNTRY

	2-Seat Electric Rickshaw	6-Seat Electric Van, Auto van	4-Seat Mishuk Auto Rickshaw	8-seat Mishuk Auto, Tom Tom or Borak	8-seat, Solar Tom Tom
Passengers	2 seats	6 seats	4 seats	8 seats	8 seats
Motor	"DC motor"	"DC motor"	"DC motor"	"DC motor"	"DC motor"
Brand	Bengal Tiger (Happy motor)	Luke	Kun Ray, PK plus	Lunyee	Luke
Model		SLZ-48V		LY-60300	MT005
Power	1200w	1000 w	1000w	3000 w	1000 w
Human Powered	Pedal power possible	Pedal power possible	No pedal power	No pedal power	No pedal power
Controller	48-volt, 1200 W	48-volt, 1000 W	48-volt, 1200 W	60-volt, 1200 W	48-volt, 1000 W
Max speed	30 km/h - 35 km/h	30 km/h - 35 km/h	30 km/h - 35 km/h	35 km/h - 40 km/h	35 km/h - 40 km/h
Rated rotation			3000 – 3400 rpm	3000 – 3400 rpm	2500-3000 rpm
Battery	48 v (12 v 4 pieces)	48 v (12 v 4 pieces)	48 v (12 v 4 pieces)	60 v (12 v 4 pieces)	48 v (12 v 4pieces)
Battery charging time	6-8 hrs.	6-8 hrs.	8-10 hrs	8-10 hrs	6-8 hrs
Running time	8-10 hours.	8-10 hours.	8-10 hours.	8-10 hours.	10-12 hours
Battery Capacity	90-120 AH	90-120 AH	120-200 AH	220 AH	. 200AH
One charging mileage	120 km - 140km	120 km - 140km	120 km - 140km	120 km - 140km	120 km - 140km
Solar power	(none)	(none)	(none)	(none)	350 w
Extra miles with solar	- NA -	- NA -	- NA -	-NA-	40-50 km

CONCLUSION

A wide range of electric vehicles are now being made in the country, as solutions to the local needs of transportation. Most of the electric rickshaws are solutions by low-income small entrepreneurs who have improvised and innovated in the absence of official recognition and regulation.

The estimated one million local small electric vehicles means there is about one electric vehicle for every 170 residents of the country. The range of vehicles today may have 2, 4, 6, or 8 seats. Known by various names, varying according to region, the imported parts include the motor, batteries, controller, and parts of the drivetrain and chassis.

There is an ongoing worldwide move towards electric vehicles, as seen in the rising stock market valuation of EV manufacturer Tesla. The proliferation of small EVs in Bangladesh is another manifestation of this worldwide trend. The country is somewhat ahead as small electric rickshaws may or not have pedal connections. These Hybrid EVs have supplementary power with pedals or with solar panels, which are unknown in the West.

It is expected that the survey and analysis of this paper will help to organize, recognize, regulate and then promote these small electric vehicles in the country.

REFERENCES

[1] M. Weiss, K. C. Cloos, E. Helmers, "Energy efficiency trade-offs in small to large electric vehicles." *Environ Sci Eur* 32, 46 (2020).

[2] R. Sarker, Electric rickshaws in Rangpur. *The Daily Star*, August 17, 2009.

[3] Laboni Sarker, S. R. Jaynab, S. Khan, "The Undocumented Electric Vehicles in Bangladesh," National Conference on Electronics and Informatics-2019, at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019,

[4] Laboni Sarker, Safia Rahman Jaynab, S. Khan, "Study on the Undocumented Electric Vehicles in Bangladesh," *International Journal of Industrial Electronics and Electrical Engineering*, ISSN(p): 2347-6982, ISSN(e): 2349-204X, Volume-8, Issue-3, Mar.-2020.

[5] M. S. Z. Chowdhury, Maisha Anjum, Zannatul. Mawa, Shahriar Khan, "Future of Electrical Vehicles in Bangladesh," National Conference on Electronics and Informatics-2019, , at Atomic Energy Centre, Dhaka, Bangladesh, 4-5 December, 2019.

[6] M. R. Ahmed and A. K. Karmaker, "Challenges for Electric Vehicle Adoption in Bangladesh," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-6,

[7] A. Mohammad, M. A. Abedin and M. Z. R. Khan, "Microcontroller based control system for electric vehicle," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 693-696,

[8] Rezanul Haque, S. Khan, "The Modified Proportional Integral Controller for the BLDC Motor and Electric Vehicle," IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, Toronto, Canada, 21-24 April, 2021.

[9] A.S.M.M.Hasan, "Electric Rickshaw Charging Stations as Distributed Energy Storages for Integrating Intermittent Renewable Energy Sources: A Case of Bangladesh" *Energies* 13, no. 22: 6119, 2020.

[10] A. Mallik, M. A. Arefin, F. Rashid, & Asfaquzzaman, 2017, "Solar Based Plugged-in Hybrid Engine Driven rickshaw (Auto-Rickshaw) & its Feasibility Analysis for Bangladesh." International Conference on Mechanical, Industrial and Materials Engineering (ICMIME2017), Dec. 2017.

[11] K. S. Reddy, S. Aravindhan, & T. K. Mallick, "Techno-Economic Investigation of Solar Powered Electric Auto-Rickshaw for a Sustainable Transport System," *Energies*,10(6), 2017.

[12] A. A. Mamun, S. Arfin, S. Khan, "High Gain DC-DC Converter for Three-Wheeler Electric Vehicles," IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021, Toronto, Canada, 21-24 April, 2021.

[13] "Bangladesh on the road to an electric future," *The Daily Star*, Wednesday, April 27, 2022 |

[14] M. R. Rasel, "PDB for bringing the battery-run vehicle under traffic," *Dhaka Tribune*, September 27, 2017.

[15] J. Chakma, "Automotive battery market revving up." *The Daily Star*, April 26, 2018.

[16] "Easy bikes outgrowing limitations," *The Daily Star*, Sept. 19, 2017.

[17] "Government to ban battery-run rickshaws, vans" *The Dhaka Tribune*, June 20th, 2021

[18] Tanim Asjad, "Rethinking ban on battery-run vehicles," *The Financial Express*, June 25, 2021.

[19] Shahin Akhter, "Battery-run rickshaws, vans continue to run defying ban amid restrictions," *The New Age*, July 06, 2021.

[20] M. A. Rahim, M. U. H. Joardder, S. M. N. Hoque, M. M. Rahman, N. H. Sumon, "Socio-economic & environmental impacts of battery-driven auto-rickshaw at Rajshahi city in Bangladesh." International Conference on Mechanical industrial and Energy Engineering, 2012.

[21] M. S. Rana, F. Hossain, S. S. Roy, S. K. Mitra, "Battery Operated Auto-rickshaw and Its Role in Urban Income and Employment-Generation," *International Journal of Advancements in Research & Technology*, 1(5), 2012.

[22] S. Khan, *Semiconductor Devices and Technology*, Third Edition, ISBN: 978-094-33-5983-4, by S. Khan, Dhaka, Bangladesh, June 3, 2018.

[23] S. Khan, *Electrical Energy Systems*, Fourth Edition, ISBN: 978-984-33-7638-1, by S. Khan, Dhaka, Bangladesh, Feb 2021.

[24] S. Khan, Mushfiqz Zaman, Mostakin Rabbi "The Million Improvised Electric Rickshaws in Bangladesh; Preliminary Survey and Analysis," Submitted for publication.

[25] S. Khan, A. H. Shahriar, M. P. Munna, "The Developing Electric Rickshaw Solutions in Bangladesh: Equipment and Experimental Waveforms," Submitted for Publication.

Analog Front-End CMOS Temperature Sensor Interface for Optogenetic Devices

Shahrzad Ghasemi, Soliman A. Mahmoud
Electrical Engineering Department
University of Sharjah
Sharjah, United Arab Emirates
{u18105750, solimanm}@sharjah.ac.ae

Abstract— This work presents a CMOS temperature sensor interface based on a second-generation current conveyor (CCII) and a transimpedance amplifier (TIA) with a high linear and temperature-independent pseudo-resistor. Since the LED in optogenetic implantable devices generates heat, a temperature sensor is required to monitor the temperature. Thus, the reverse current of the LED can be employed as its own sensing element. This paper will utilize the LED reverse current as a temperature-sensitive parameter (TSP) to sense the junction temperature. The sensor interface operates under a supply voltage of ± 0.6 V. The reverse current driving capability is ± 1.7 μ A. The total power consumption of the CMOS sensor interface is 192 μ W. The circuit provides an operational frequency of up to 2.08 MHz. The transimpedance gain of the proposed temperature sensor interface is equal to 300×10^3 V/A. The proposed circuit is designed using 90 nm CMOS technology and simulated using LTspice.

Keywords— CMOS temperature sensor interface, optogenetics, second-generation current conveyor (CCII), transimpedance amplifier (TIA).

I. INTRODUCTION

Over the past few decades, biomedical researchers have been focused on utilizing microtechnologies and nanotechnologies in healthcare applications [1]. Microtechnologies and nanotechnologies' main advantage is developing devices with micrometer and nanometer scales. These technologies provide the interface between nervous system disorder treatments and electronic components. The invasive micrometer and nanometer scales technologies are known as neural probes [2].

Optogenetic technology is a combination of optics and genetics. The technology is used to provide stimulation and suppression in brain tissue photosensitive cells. The stimulation and suppression are achieved by light-emitting. Therefore, the stimulus can be provided through light pulses. The light sources of optogenetic are optical fibers and light-emitting diodes (LEDs). The LED neural probe overcomes light losses and maximizes light power delivery [3].

Manufacturing and implanting optogenetic neural probes have some challenges. One of these main concerns is preventing tissue damage. During the stimulation, the overheating process could damage the human body tissues. Hence, providing a temperature sensor in the optogenetic device is particularly important. The challenge of implementing a temperature sensor for an optogenetic probe occurs when the sensor is required to be designed by taking the power consumption, temperature dependency, and dimensions into consideration [3], [4].

The human body temperature remains almost constant (37°C) over a specific range of environmental temperatures. Nevertheless, the human body is sensitive to temperature fluctuations [5]. Approximately a 2 °C temperature rise

represents a threshold temperature to prevent the brain from overheating damage. However, the threshold temperature varies based on the brain activity state and other factors discussed in [6], [7].

In order to design and implement a temperature sensor for optogenetics, the sensor interface must include an analog front-end (AFE) element, analog-to-digital converter (ADC), and digital signal processor (DSP). One of the techniques of implementing a temperature sensor is the current-mode temperature sensor. The LED reverse current (I_R) can be utilized to sense the temperature variation. Thus, I_R is considered a temperature-sensitive parameter (TSP). The LED that is used as a light source for optical stimulation can be utilized as a sensing element. Based on that, the LED used in optogenetic can be used in two phases, which are stimulation and sensing phases [8].

Gallium nitride (GaN) and indium gallium nitride (InGaN) LEDs are widely utilized in optogenetic devices to target a range of opsins, which are photoreceptive proteins, owing to the tunability feature of the emission wavelength across the visible spectrum [9]. As presented by [10] and [11], the I_R of the InGaN and GaN LEDs can be used for junction temperature (T_J) determination, respectively. Therefore, the I_R can be used to measure the T_J and surface temperature (T_S), which is related to T_J [10]. As introduced in [8] and [4], the I_R is used for temperature sensing purposes. The research proposed a method to determine the T_S of implanted devices by reusing LED, which is utilized for light emission, as a temperature-sensitive element. Thus, the I_R can be employed to measure the T_J , and I_R is named TSP for the optogenetic temperature sensor interface. The I_R is exponentially proportional to the T_J inverse [8]. The technique has been proposed by [4], [8] with the power consumption equal to 260 μ W. This current mode temperature sensor optogenetic has been implemented in 0.35 μ m CMOS technology and it operates under supply voltage +5 V. Moreover, a current mode temperature sensor interface technique has been presented by [12]. The designed temperature sensor AFE has been designed in 0.25 μ m CMOS technology. The circuit has been designed with a supply voltage of ± 0.75 V. In order to achieve compatibility with different types of LEDs, the gain of the designed temperature sensor interface is digitally controlled. The current summing network (CSN) was utilized to provide the digital programmability. The total power consumption of the presented temperature sensor varies between 240 μ W and 700 μ W.

In this paper, an AFE CMOS temperature sensor interface is proposed. The realized temperature sensor interface consists of voltage biasing and current sensing using a second-generation current conveyor (CCII). Moreover, it consists of a transimpedance amplifier (TIA). The TIA is used to sense the current of the CCII and convert

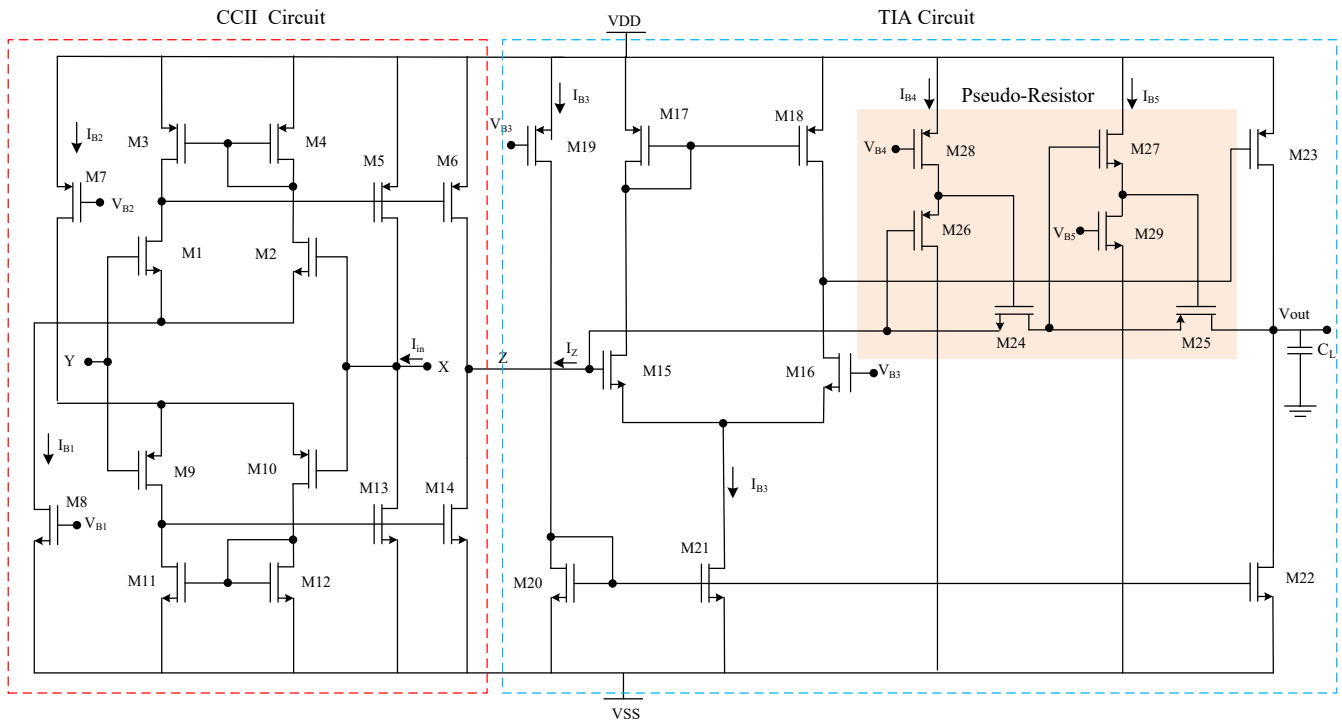


Fig. 1. The proposed CMOS realization of the temperature sensor AFE circuit.

it to a voltage signal. The feedback resistor of the TIA is realized using a pseudo-resistor. Therefore, the proposed AFE is designed based on CCII and TIA with an embedded pseudo-resistor in the feedback of the TIA. The temperature-independent pseudo-resistor has been used to prevent the AFE from temperature dependency. The CCII has been employed as a voltage buffer to ensure stable biasing voltage using Y and X terminals. Additionally, it is used for current sensing to convey the LED reverse current from X terminal to Z terminal. The TIA has been employed to convert and amplify the output current of the CCII to an amplified voltage signal. The amplified voltage signal is the input signal of the ADC block. The temperature sensor AFE operates under a supply voltage of ± 0.6 V.

This paper is organized as follows: The proposed temperature sensor interface design is discussed in Section II. The implemented circuit is evaluated and tested in Section III. The circuit is simulated and evaluated using LTspice with 90 nm CMOS technology. Finally, Section IV concludes the work.

II. PROPOSED CMOS TEMPERATURE SENSOR INTERFACE

The proposed CMOS AFE temperature sensor structure is designed using a CCII cascaded with a TIA. The designed AFE circuit is shown in Fig. 1.

The LED I_R is the X terminal input current of the CCII, and it conveys to the Z terminal current of the CCII. Moreover, the X terminal voltage follows the Y terminal voltage, which represents the LED biasing voltage. Therefore, the CCII provides voltage and current buffers. The CCII circuit has been realized by relying on the CCII presented by [13]. The input stage of the CCII circuit is designed by utilizing two complementary differential pairs. These two pairs are NMOS matched differential pair and

PMOS matched differential pair. The NMOS and PMOS pairs are connected in parallel. Thus, the voltage follower between X and Y terminals has been provided. Moreover, to achieve a rail-to-rail operation, the differential pairs are biased with a constant tail current. The CCII is designed to provide a high input impedance. Hence, the current that passes through the Y terminal is equal to zero in the realized circuit. On the other hand, the output stage is class AB; thus, it guarantees high current driving capability. The X terminal current is mirrored to the Z terminal using matched transistors pairs.

The voltage follower between X terminal and Y terminal has been provided using two matched differential pairs. These differential pairs are (M1, M2) and (M9, M10), representing NMOS and PMOS differential pairs, respectively. These differential pairs are biased using transistors (M7, M8) with currents I_{B2} and I_{B1} , respectively. Two matched current mirror pairs (M3, M4) and (M11, M12) are used for current mirroring for the complementary differential pairs in the input stage. The current of the X terminal is conveyed to the Z terminal using two matched transistors pairs (M5, M13) and (M6, M14).

As mentioned previously, The CCII output current is the input signal of the next block used as a converter and amplifier. The proposed TIA is a modified version of the two-stage operational transconductance amplifier (OTA) given by [14]. The modification has been achieved by replacing the feedback resistor with a pseudo-resistor. The TIA circuit is utilized for sensing, converting to voltage, and amplifying the output current of the CCII. The output voltage is the input signal of the ADC block.

Differential pair (M15, M16) has been used to implement the TIA circuit. In order to get a single output voltage, transistor M17 mirrors the current to transistor

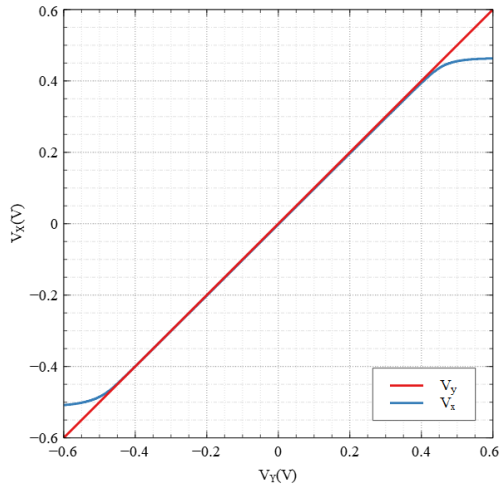


Fig. 2. The voltage follower between Y and X terminals of the proposed temperature sensor AFE.

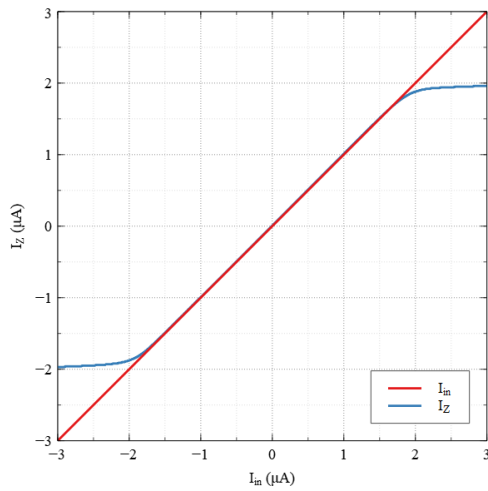


Fig. 3. The X and Z terminals current of the temperature sensor AFE.

M18. The differential pairs are biased through transistors M19, M20, and M21 with biasing current I_{B3} . The output stage of the TIA consists of transistors (M23, M22) and it forms class A output stage.

The pseudo-resistor has been designed using two MOS transistors. One NMOS transistor M24 and one PMOS transistor M25. The combination of these two MOS transistors mitigates the nonlinearity of the pseudo-resistor. Each transistor is connected to a source-followers.

These two source-followers afford dynamically adjusting the two transistors' gate to source voltages. This adjustment will provide constant V_{GS} on the pseudo-resistor for large output voltage swings. The resistor value of the pseudo-resistor can be varied by adjusting the biasing currents I_{B4} and I_{B5} . The proposed pseudo-resistor is used to design a temperature-independent feedback resistor within a specific range of temperature variation.

The output signal, which is voltage, of the temperature sensor AFE is given by

$$V_{Out} = R_f I_{in}. \quad (1)$$

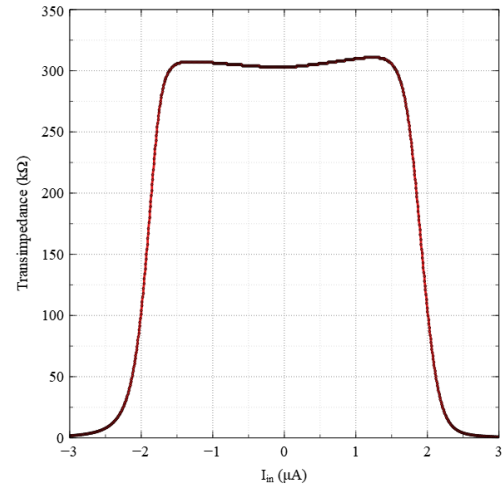


Fig. 4. Transresistance gain provided by the feedback pseudo-resistor.

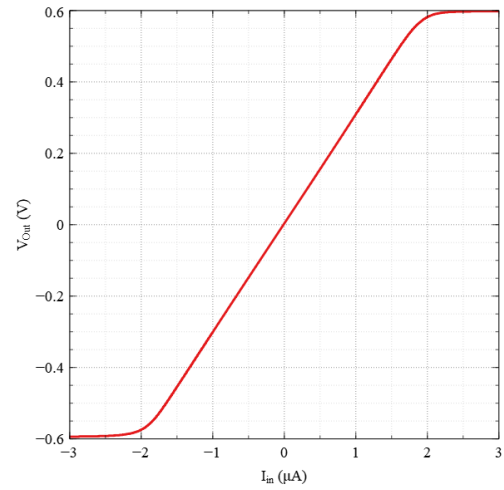


Fig. 5. The DC characteristics of the temperature sensor AFE.

The total power dissipation of the temperature sensor AFE with the pseudo-resistor circuit can be driven as

$$P = 2V_{DD}(I_{B2} + 2I_{M3} + 2I_{M5}2I_{B3} + I_{M23} + I_{B4} + I_{B5}). \quad (2)$$

III. SIMULATION RESULTS

The temperature sensor interface circuit has been simulated by LTspice using 90 nm CMOS process technology based upon the BSIM4 (level 54) MOSFETs model under the supply voltage ± 0.6 V. The circuit has been simulated at 37 °C as a base human body temperature.

The NMOS and PMOS differential pairs are biased with 1.3 μ A and 3.9 μ A tail currents, respectively. The circuit is used to provide constant bias voltage by buffering the voltage from the Y terminal to the X terminal of the CCII. The voltage follower action between Y and X terminals has been simulated by varying the Y terminal DC voltage. Fig. 2 represents the voltage follower of the CCII used for the temperature sensor interface to provide stable biasing voltage. The voltage of the Y terminal is transferred to the X terminal in the dynamic range of ± 0.4 V.

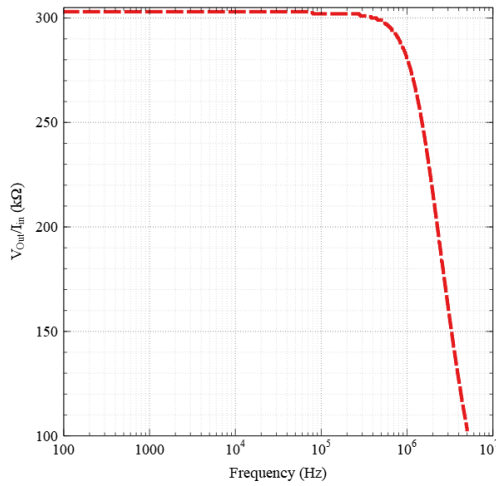


Fig. 6. The magnitude response of the temperature sensor AFE.

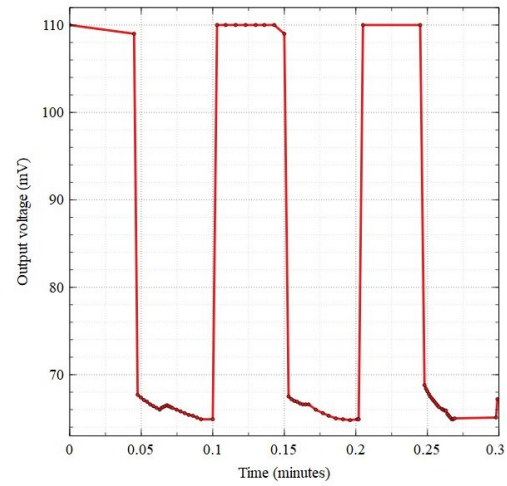


Fig. 8. The output voltage time response of the AFE.

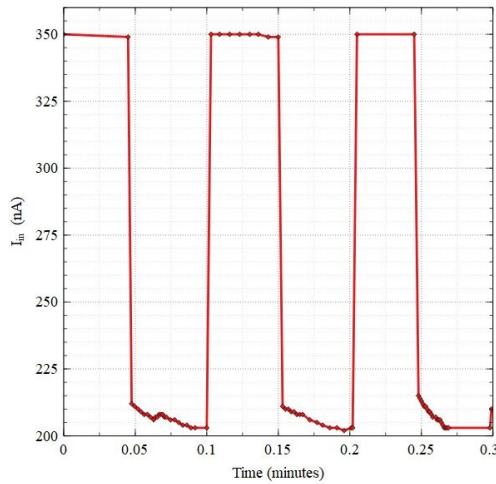


Fig. 7. The input and output currents time response of the sensor AFE.

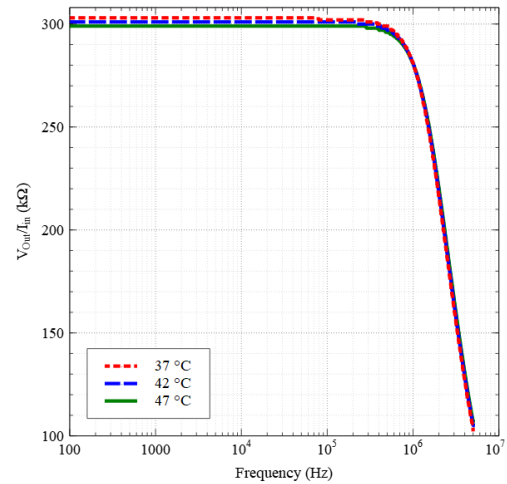


Fig. 9. The magnitude response of the temperature sensor AFE under temperature variation.

The voltage follower has been provided with a bandwidth equal to 171 MHz and the DC voltage gain equal to 1.0038.

As mentioned earlier, the CCII conveys the reverse current of the LED from X terminal to Z terminal. Fig. 3 shows the X and Z terminals currents when the X terminal current is varied from -3 to 3 μA . As shown in the figure, the X terminal conveys the current to the Z terminal with a dynamic range $\pm 1.7 \mu\text{A}$. The total power consumption of the CCII circuit is less than 119 μW .

As discussed in the previous section, the output current of the CCII block amplification and conversion occurs by the TIA block. The feedback resistor of the TIA has been designed using a pseudo-resistor. Therefore, the total transimpedance gain of the AFE temperature sensor interface is equal to the pseudo-resistor resistance value. The constant transimpedance gain has been achieved by pseudo-resistor for the input current range of $\pm 1.8 \mu\text{A}$, as shown in Fig. 4. The transresistance gain of the temperature sensor interface which is provided by the pseudo-resistor is equal to $300 \times 10^3 \text{ V/A}$. Therefore, the feedback resistor value equals 300 k Ω with V_{B3} equals zero, representing the common-mode voltage.

The DC characteristic of the temperature sensor interface has been simulated by varying the input current

and measuring the output voltage. The TIA converts the input current to voltage linearly with constant transimpedance gain equal to $300 \times 10^3 \text{ V/A}$, as shown in Fig. 5. The total power consumption of the TIA with pseudo-resistor is 72 μW .

The magnitude response of the temperature sensor interface is demonstrated in Fig. 6. The bandwidth of the implemented circuit is equal to 2.08 MHz.

Since the optical stimulation has been achieved by applying current pulses, the designed circuit has been simulated and evaluated by applying input pulses. The reverse current of the current pulses is temperature-dependent and changes with temperature variations. The transient response of the input current pulses is demonstrated in Fig. 7. The output voltage of the temperature sensor interface circuit represents the converted and amplified signal of the input signal. The output voltage is represented in Fig. 8 for the input current given in Fig. 7. The total power consumption of the temperature sensor interface is 192 μW .

The circuit is implemented to sense the reverse current of the LED in order to sense the temperature variation. The temperature variations can affect the transistors' parameters used in the analog circuit [15]. Therefore, the

designed circuit is required to be temperature independent within a specific temperature range. Considering implantable optogenetic devices, human body temperature, which is equal 37 °C, is considered a base temperature. The magnitude response of the temperature sensor interface has been simulated for temperatures 37 °C, 42 °C, and 47 °C. The temperature variation result is shown in Fig. 9. The summary of the performance parameters of the proposed temperature sensor interface is given in TABLE 1.

TABLE 1. TEMPERATURE SENSOR AFE RESULTS

Parameters	Proposed Sensor Interface
CMOS technology (nm)	90
Voltage supply (V)	±0.6
Transimpedance gain (V/A)	300×10 ³
3-dB bandwidth (MHz)	2.08
Power consumption (µW)	192
Voltage dynamic range (V)	±0.4
Current driving capability (µA)	±1.7
I _Z -I _{in} Offset current (nA)	10
IRN (nA/√Hz) at 100 Hz	1.3
Temperature Range (°C)	28-60

IV. CONCLUSION

This paper presented a CMOS temperature sensor interface based on a CCII and a TIA with a temperature-independent pseudo-resistor. The circuit was designed and simulated using 90 nm CMOS technology under ±0.6 V supply voltage. The method used the LED for the optical stimulation in optogenetic implantable devices as a TSP to permanently sense the T_s, which depends on the T_j. Therefore, the temperature sensor AFE was designed to provide a biasing voltage for the LED and sense the reverse current. Moreover, the TIA has been utilized for signal conversion and amplification. The total power consumption of the designed circuit is equal to 192 µW. The transimpedance gain of the circuit is equal to 300×10³ V/A. In future work, the sensor interface may include the reduction of the total power consumption.

REFERENCES

[1] M. Ferrari, *BioMEMS and biomedical nanotechnology: volume II: micro/nano technologies for genomics and proteomics*, vol. 2. Springer Science & Business Media, 2007.

[2] M. M. Maharbiz, R. Muller, E. Alon, J. M. Rabaey, and J. M. Carmena, "Reliable next-generation cortical interfaces for chronic brain-machine interfaces and neuroscience," *Proc. IEEE*, vol. 105, no. 1, pp. 73–82, 2016.

[3] S. Goncalves et al., "LED Optrode with Integrated Temperature Sensing for Optogenetics," *Micromachines*, vol. 9, no. 9, p. 473, 2018.

[4] F. Dehkhoda, A. Soltan, N. Ponon, A. O'Neill, A. Jackson, and P. Degenaar, "A current-mode system to self-measure temperature on implantable optoelectronics," *Biomed. Eng. Online*, vol. 18, no. 1, pp. 1–15, 2019.

[5] K. Jackson et al., "Brain temperature and its fundamental properties: a review for clinical neuroscientists," *Front. Neurosci.*, vol. 8, pp. 307–324, 2014.

[6] C. Childs, "Human brain temperature: regulation, measurement and relationship with cerebral trauma: part 1," *Br. J. Neurosurg.*, vol. 22, no. 4, pp. 486–496, 2008.

[7] E. A. Kiyatkin, "Brain hyperthermia during physiological and pathological conditions: causes, mechanisms, and functional implications," *Curr. Neurovasc. Res.*, vol. 1, no. 1, pp. 77–90, 2004.

[8] F. Dehkhoda, A. Soltan, N. Ponon, A. Jackson, A. O'Neill, and P. Degenaar, "Self-sensing of temperature rises on light emitting diode based optrodes," *J. Neural Eng.*, vol. 15, no. 2, p. 26012, 2018.

[9] F. Wu, E. Stark, P.-C. Ku, K. D. Wise, G. Buzsáki, and E. Yoon, "Monolithically integrated µLEDs on silicon neural probes for high-resolution optogenetic studies in behaving animals," *Neuron*, vol. 88, no. 6, pp. 1136–1148, 2015.

[10] B. Wu et al., "Junction-temperature determination in InGaN light-emitting diodes using reverse current method," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 241–245, 2012.

[11] E. Jung, J. K. Lee, M. S. Kim, and H. Kim, "Leakage current analysis of GaN-based light-emitting diodes using a parasitic diode model," *IEEE Trans. Electron Devices*, vol. 62, no. 10, pp. 3322–3325, 2015.

[12] S. Ghasemi and S. A. Mahmoud, "Current-mode self-sensing temperature sensor using DC-CCII for optoelectronic devices," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-5, 2021.

[13] T. M. Hassan and S. A. Mahmoud, "Fully programmable universal filter with independent gain-ω0-Q control based on new digitally programmable CMOS CCII," *J. Circuits, Syst. Comput.*, vol. 18, no. 05, pp. 875–897, 2009.

[14] G. Palmisano, G. Palumbo, and S. Pennisi, "Design procedure for two-stage CMOS transconductance operational amplifiers: A tutorial," *Analog Integr. Circuits Signal Process.*, vol. 27, no. 3, pp. 179–189, 2001.

[15] H. Veldandi and R. A. Shaik, "Low-voltage PVT-insensitive bulk-driven OTA with enhanced DC gain in 65-nm CMOS process," *AEU-International J. Electron. Commun.*, vol. 90, pp. 88–96, 2018.

A High-order Temperature-compensated Bandgap Voltage Reference with Low Temperature Coefficient

Shalin Huang

School of Microelectronics and Communication Engineering School of Microelectronics and Communication Engineering
 Chongqing University
 Chongqing, China
 shalin_huang@163.com

Mingdong Li

School of Microelectronics and Communication Engineering
 Chongqing University
 Chongqing, China
 594242555@qq.com

Peng Yin

School of Microelectronics and Communication Engineering School of Microelectronics and Communication Engineering
 Chongqing University
 Chongqing, China
 yinpeng9527@cqu.edu.cn

Fang Tang

School of Microelectronics and Communication Engineering
 Chongqing University
 Chongqing, China
 eefrank@cqu.edu.cn

Abstract—A high-order temperature-compensated bandgap voltage reference (BGR) with a low-temperature coefficient (TC) for high-precision applications is proposed, manufactured in a 0.18- μm CMOS process. Strong-inversion MOSFETs and forward-biased Bipolar Junction Transistors (BJTs) are employed in the proposed high-order temperature-compensated circuit, which eliminates the curvature in base-emitter voltage (V_{BE}), so to achieve a low TC. Measurement results prove that a minimum TC of 0.7 ppm/ $^{\circ}\text{C}$ over the temperature range of -25°C to 125°C is realized with a resistance trimming network. The line sensitivity is 0.0146%/V when supply voltage changes from 3.2 V to 3.7 V.

Index Terms—BGR, High order, Temperature coefficient, Line sensitivity

I. INTRODUCTION

Precise voltage references undoubtedly are important in data converters, operational amplifiers linear regulators, etc [1], [2]. As it is all known, the base-emitter voltage (V_{BE}) of a diode-configure bipolar transistor (BJT) is well characterized in temperature. Therefore, bandgap reference (BGR) technology, whose output is highly independent of temperature [3], is popular in on-chip reference voltages. To get an accurate reference voltage, attention must be paid to process variations, voltage, temperature.

For first-order temperature-compensated BJT-based BGRs, the standard design method is to use the component ΔV_{BE} , which is proportional to the absolute temperature, to compensate for the component V_{BE} that is complementary to the absolute temperature. Most modules need a reference

voltage with a low TC for power supply, so the purpose of the BGR design is to find a way to generate an insensitive-temperature output. While the temperature sensitivity is usually limited to 20 ~ 100 ppm/ $^{\circ}\text{C}$ [4]–[6] for first-order BJT-based BGRs, which is not enough for high-performance applications. Therefore, a high-order compensation circuit must be used to eliminate the interference of curvature content. In the reported designs [7]–[11], a TC of sub-1 ppm/ $^{\circ}\text{C}$ is still hard to achieve. This paper proposes a high-order temperature-compensated BGR with a low TC, by simply combining two strongly inverted MOSFETs and two forward-biased BJTs. The generated compensation current is expressed in the form of $T \exp(f(T))$ after theoretically analyzing and can be designed to approach the $T \ln(T)$ term in V_{BE} , realizing the non-linear compensation.

II. PROPOSED STRUCTURE

In the vast majority of BJT-based bandgap [1]–[3], [12], the bandgap-energy-related component is provided by V_{BE} of the diode-configure BJT. According to previous works, the nonlinear relationship between V_{BE} and temperature can be expressed as [13]

$$V_{BE}(T) = V_{GO}(T_r) - \left[\frac{V_{GO}(T_r) - V_{BE0}(T_r)}{T_r} - (\eta - \theta) \frac{k}{q} \ln T_r \right] T - (\eta - \theta) \frac{k}{q} T \ln T \quad (1)$$

where $V_{GO}(T_r)$ is the silicon bandgap voltage at a reference temperature T_r and it is 1.2 V at 0 K, $V_{BE0}(T_r)$ is the base-emitter voltage at T_r , η is process related, with a nominal value of 3 ~ 4 [14], θ is collector current temperature-dependent order, expressed as $I_C(T) = I_{C0}(T/T_r)^{\theta}$, where

This work was partly sponsored by Natural Science Foundation of Chongqing, China, No. cstc2019jcyj-zdxmX0014, and by the Research Foundation of Chongqing Science and Technology Bureau No. cstc2018jzcx-cyztzxX0049.

The second and the third term in (5) are used to compensate for the linear and the nonlinear term in V_{BE} , respectively. The resistor R_1 is fulfilled by an 8-bit resistor array to trim the linear compensation term, as shown in Fig. 3(b). MOSFETs are working in the linear region to be acted as switches, and they have the same unit size, in Fig. 3(b). The size of switches should be large enough to minimize the effects owing to the drain-source voltage V_{DS} . Therefore, the size of $19.8\mu/600n$ is chosen. The fixed resistor on the right side is to make sure the trimming code has a wide range. The equivalent resistance value of the trimming networks varies from 33.42% to 100% of R_t , broadening the trimming range.

III. SIMULATION AND MEASUREMENT RESULTS

Fig. 4 compares the simulation currents of $\frac{1}{p}I_{PT}, I_{CT}$ and I_{b5} (base current of Q_5). The result shows that ignoring I_{b5} is reasonable in Section II. Fig. 5 compares the output

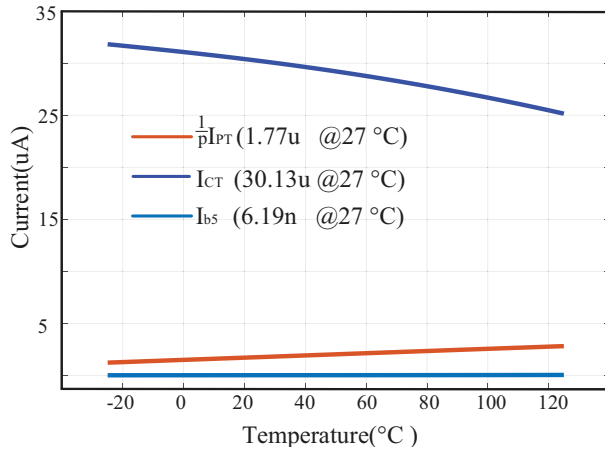


Fig. 4. Simulated currents of $\frac{1}{p}I_{PT_C}$, I_{CT_C} and I_{b5}

temperature dependency of this proposed circuit with that of an optimized first-order BGR. To compare fairly, the output voltage of optimized first-order BGR is equal to that of the proposed compensated BGR. The TC performance of the first-order BGR is about $3.166 \text{ ppm}/^\circ\text{C}$, and that of the proposed BGR is $0.623 \text{ ppm}/^\circ\text{C}$. Therefore, it can be seen that the existence of this high-order compensation circuit significantly improves the output's performance in terms of TC (about 5 times better). In order to verify the sensitivity of the designed BGR to mismatch and process variations, Monte Carlo simulations of all components (MOS transistors, BJT, resistors and capacitors) were carried out 1000 times. Fig. 6(a) shows the temperature curve of V_{ref} , and Fig. 6(b) shows the corresponding statistic TC distribution. Consider the TC distribution is non-Gaussian, therefore sigma is not meaningful [3], [15]. The result shows that the average TC of 1000 runs is $6.794 \text{ ppm}/^\circ\text{C}$ from -25°C to 125°C and the proportion of TC greater than $12 \text{ ppm}/^\circ\text{C}$ is 13.3%.

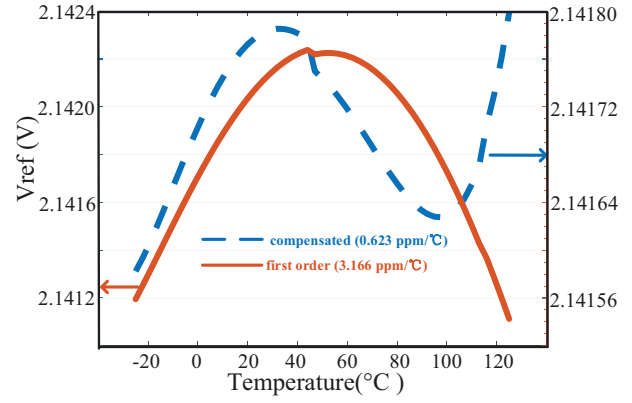


Fig. 5. TC performance comparison between the optimized first-order BGR and the proposed BGR with high-order compensation

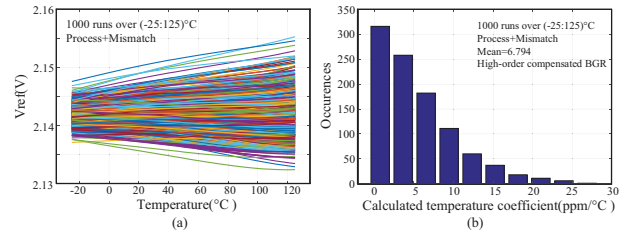


Fig. 6. (a) Monte Carlo simulation results and (b) the histogram of the calculated TC from (a)

The chip micrograph is shown in Fig. 7. Fig. 8 presents the line sensitivity of this proposed BGR. The change of reference voltage is 0.156 mV and the line sensitivity is calculated as $0.0146\%/V$ at room temperature, when the supply voltage is from 3.2 V to 3.7 V , embodying high stability. Six samples were tested for TC, in which

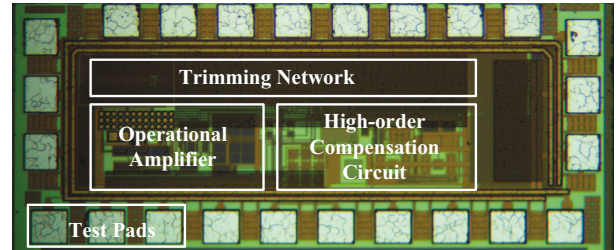


Fig. 7. Chip microphotograph with an active area of $908 \mu\text{m} \times 282 \mu\text{m}$

the OYO1000 programmable temperature chamber with a temperature range of -70°C to 150°C was used for temperature scanning. The voltage was measured using a Keysight 34401A digital multimeter with a resolution of $1 \mu\text{V}$. The temperature dependencies of 6 samples are shown in Fig. 9, where the peak to peak measured difference is 0.82 mV . A minimum measured TC of $0.706 \text{ ppm}/^\circ\text{C}$, a maximum one of $1.535 \text{ ppm}/^\circ\text{C}$, and an average one of $1.076 \text{ ppm}/^\circ\text{C}$ are achieved respectively, making the proposed BGR competitive in terms of temperature sensitivity.

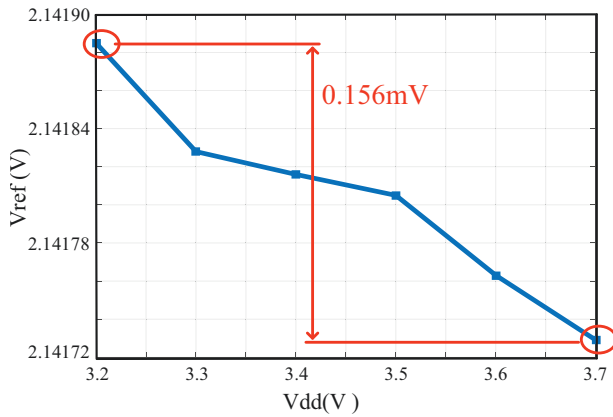


Fig. 8. Measured output voltage at room temperature as the function of supply voltage

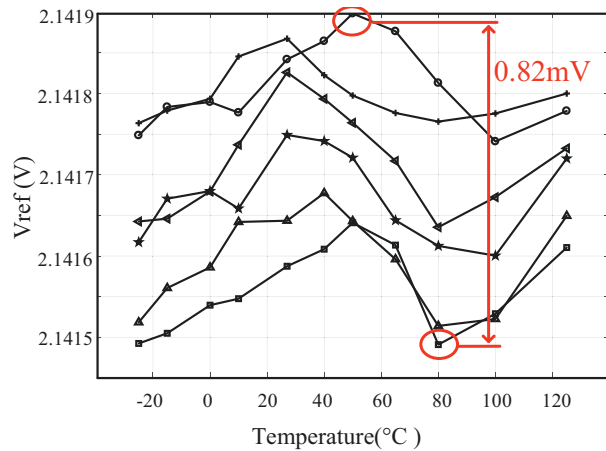


Fig. 9. Measured output voltage versus temperature of 6 samples

TABLE I
PERFORMANCE SUMMARY AND COMPARISON

	[7] TCASI'21	[8] JSSC'21	[9] JSSC'21	This work
Tech.(nm)	65 CMOS	180 CMOS	130 CMOS	180 CMOS
Supply Voltage (V)	2.5~3.6	1.8	2~3.3	3.2~3.7
Line Sensitivity (%/V)	NA	NA	0.03	0.0146
Temp. Range (°C)	0~80	-40~125	-40~150	-25~125
TC (ppm/°C)	0.8(min) 0.87(max)	3.2(min) 5.5(max) 4.3(avg)	5.78(min) 13.5(max) 8.75(avg)	0.706(min) 1.535(max) 1.076(avg)
# of samples	3	18	7	6

NA: No Information Available

Table I compares the performances of this design with other reported BGR designs. The proposed BGR obtains a comparable TC with [7]; while it obtains a wider temperature range than [7]. Compared with [8], [9], the proposed BGR obtains a better TC; while it obtains a better line sensitivity than [9].

IV. CONCLUSION

This paper proposes a high-order compensation circuit to correct the curvature of bandgap reference, by simply combining MOSFETs and BJTs (biased in their strong-inversion region). The topology diagram of Fig. 1 constructs a complex function of temperature. By adjusting the parameters in this function, the compensation term can be used to compensate for the nonlinear term in V_{BE} well. Hence a low TC can be achieved. The recommended compensation current has been manufactured in the standard 0.18- μ m CMOS process. Measurement results show that the proposed BGR is competitive in terms of TC and line sensitivity.

REFERENCES

- [1] A. Bendali and Y. Audet, "A 1-V CMOS Current Reference With Temperature and Process Compensation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 7, pp. 1424-1429, July 2007.
- [2] C. Avoinne et al., "Second-order compensated bandgap reference with convex correction," *Electronics Letters*, vol. 41, no. 5, pp. 276-277, Mar. 2005.
- [3] Y. Huang et al., "BiCMOS-Based Compensation: Toward Fully Curvature-Corrected Bandgap Reference Circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 4, pp. 1210-1223, Apr. 2018.
- [4] Y. Lam and W. Ki, "CMOS Bandgap References With Self-Biased Symmetrically Matched Current Voltage Mirror and Extension of Sub-1-V Design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 6, pp. 857-865, June 2010.
- [5] Inyeol Lee and Gyudong Kim and Wonchan Kim, "Exponential curvature-compensated BiCMOS bandgap references," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 11, pp. 1396-1403, Nov. 1994.
- [6] R. T. Perry et al., "A 1.4 V Supply CMOS Fractional Bandgap Reference," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2180-2186, Oct. 2007.
- [7] N. Liu, R. L. Geiger and D. Chen, "Sub-ppm/°C Bandgap References With Natural Basis Expansion for Curvature Cancellation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 9, pp. 3551-3561, Sept. 2021.
- [8] J.-H. Boo, et al., "A Single-Trim Switched Capacitor CMOS Bandgap Reference With a 3σ Inaccuracy of +0.02%, -0.12% for Battery-Monitoring Applications," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 4, pp. 1197-1206, Apr. 2021.
- [9] K. Chen, L. Petruzzi, et al., "A 1.16 V 5.8-to-13.5 ppm/°C Curvature-Compensated CMOS Bandgap Reference Circuit With a Shared Offset-Cancellation Method for Internal Amplifiers" *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 267-276, Jan. 2021.
- [10] L. Liu, X. Liao and J. Mu, "A 3.6 Vrms Noise, 3 ppm/°C TC Bandgap Reference With Offset/Noise Suppression and Five-Piece Linear Compensation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 10, pp. 3786-3796, Oct. 2019.
- [11] H. Chen et al., "A Sub-1 ppm/°C Precision Bandgap Reference With Adjusted-Temperature-Curvature Compensation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 6, pp. 1308-1317, June 2017.
- [12] R. J. Widlar, "New developments in IC voltage regulators," *IEEE Journal of Solid-State Circuits*, vol. 6, no. 1, pp. 2-7, Feb. 1971.

- [13] G. Zhu, Y. Yang and Q. Zhang, "A 4.6-ppm/ $^{\circ}C$ High-Order Curvature Compensated Bandgap Reference for BMIC," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 9, pp. 1492-1496, Sept. 2019.
- [14] G. Zhu, Y. Yang and Q. Zhang, "A 4.6-ppm/ $^{\circ}C$ High-Order Curvature Compensated Bandgap Reference for BMIC," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 9, pp. 1492-1496, Sept. 2019.
- [15] K. K. Lee et al., "A Sub- μ W Bandgap Reference Circuit With an Inherent Curvature-Compensation Property," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 1, pp. 1-9, Jan. 2015.

FPGA IMPLEMENTATION OF ARTIFICIAL NEURAL NETWORK (ANN) FOR ECG SIGNAL CLASSIFICATION

Shatharajupally Vinaykumar
VLSI System, ECE
NIT, Trichy
Tamil Nadu, India
208220030@nitt.edu

Thilagavathy R
ECE
NIT, Trichy
Tamil Nadu, India
thilagavathy@nitt.edu

Abstract—The heart is one of the crucial parts of the human being. The graphical recording of the cardiac cycle produced by an Electrocardiograph is called an Electrocardiogram (ECG) signal. To predict the occurrence of an arrhythmia, an electrocardiogram (ECG) is generally used by doctors to identify the condition of the patient. Hence, to accurately detect the abnormalities of the heart in advance and classify those diseases without human involvement many machine learning algorithms are used. The MIT-BIH Arrhythmia database is being used to classify the beat classification performance. This paper presents the hardware implementation of a classifier using an Artificial Neural Network (ANN) to classify four abnormalities (Normal beat, Supraventricular ectopic beat, Ventricular ectopic beat, Fusion beat) of heartbeat with high accuracy. To an appropriate input vector for the classifier, several preprocessing stages have been applied. Discrete Wavelet Transform (DWT) is used to extract the features from the ECG signal. To implement this work, Xilinx Artix-7 NESYS 4 DDR FPGA board is used. This model got 86% testing accuracy in simulation and 85.6% in hardware.

Keywords—ECG (Electrocardiography), ANN (Artificial Neural Network), DWT (Discrete Wavelet Transform), Xilinx, Field Programmable Gate Array

I. INTRODUCTION

The electrocardiogram (ECG) shows the plot of the bio-potential generated by the activity of the heart and is used by physicians to predict and treat various cardiovascular diseases. Classification of ECG plays an important role in the early and accurate detection of arrhythmia types and is important in choosing the appropriate treatment for a patient. The MIT-BIH data Arrhythmia database is available online that has around about 48 ECG recordings [11] obtained from

patients and each recording contains approximately 30 min duration, which was digitized at 360 Hz per channel, and they have a resolution of 11 bits over a 10mV range. The flow diagram of the work is given in Fig. 1.

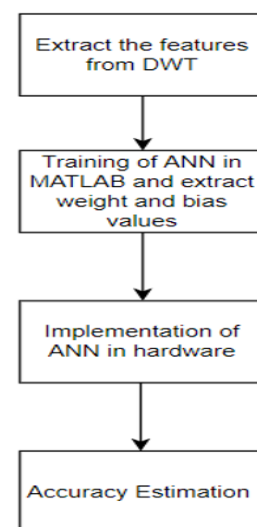


Fig1. Flow diagram of Hardware Implementation

Different types of classifiers are available for classifying the ECG signal. Among all, ANN is most widely used because of its ease of implementation in hardware.

ANN consists of several layers named input, hidden, and output layers. In each layer, it consists of several neurons. These neurons will imitate our biological neurons in the

brain. Each neuron will perform a certain task according to the application. The output of one neuron will give the input to the other layer neuron. One layer of neurons will be connected to the other layer of neurons using links. Each link is associated with a weight value, these values will get by the feed-forward backpropagation algorithm [12]. For each neuron, there is an activation function that performs certain functions according to the application. The simple ANN Architecture is given in Fig. 2.

A neuron consists of pre-activation and activation functions and Fig. 3 depicts the structure of a single neuron.

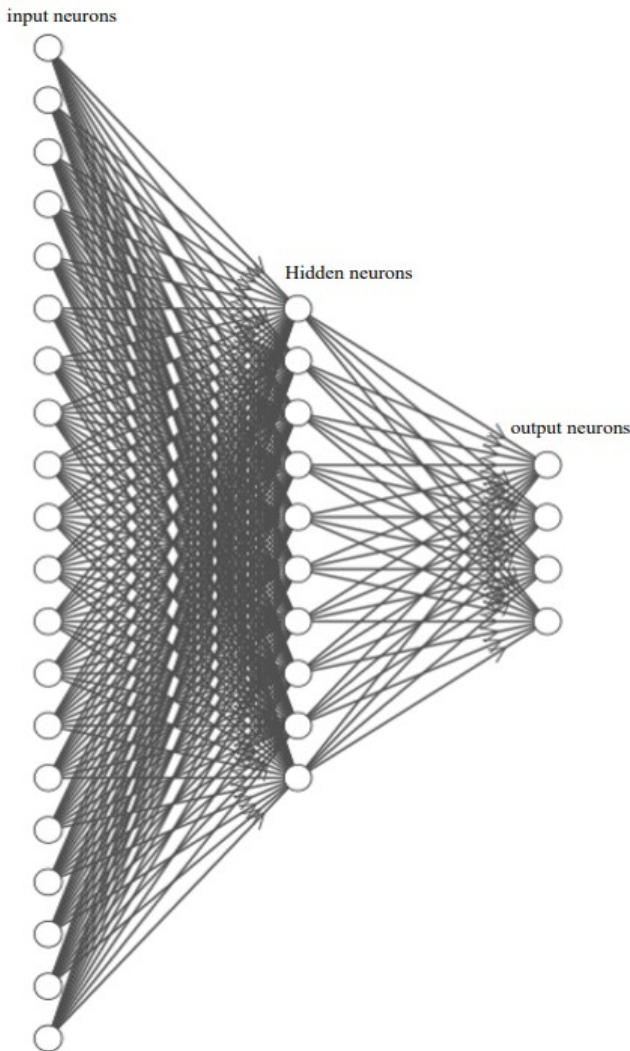


Fig 2. ANN Network of this paper

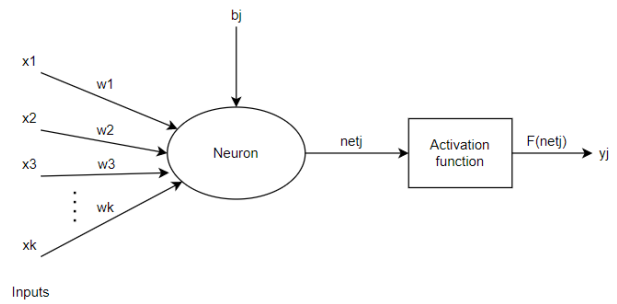


Fig 3. Structure of a single neuron

The pre-activation can be calculated by the below equation, [10]

$$a_i(x) = b_i + W_i h_{i-1}(x) \tag{1}$$

Where, $a_i(x)$ = pre-activation output of i^{th} node

b_i = bias value of i^{th} node

W = Weight matrix

h = input vector

The activation at layer i is given by,

$$h_i(x) = g(a_i(x)) \tag{2}$$

$g(\cdot)$ indicates the activation function of the neuron. The most common activation functions used in ANN are linear, sigmoid, SoftMax, ReLU, Leaky ReLU, etc. Depending on the application appropriate activation function can be used.

Field Programmable Gate Arrays [9] can be programmed or configured by the user during implementation. These boards can be rapidly used to prototype ASICs, or as a substitute for a place where ASIC will be eventually used. The programming of FPGA is done by writing Verilog code or circuit diagrams using Hardware Description Language (HDL). The basic components present in the FPGA board are CLBs, memory units that are part of look-up tables, I/O Blocks, etc. Free versions are available, High-end processors/ Processor platforms, IP cores (like MATLAB Functions), Clock Multipliers, Dedicated Memories, MAC units, ARM Bus Architecture and Bus accessories, Separate Clock tree network to minimize clock skew and delay in clock distribution to all corners of the ICs Available in different capacities and facilities.

In 2015, S. H. Jambukia, [8] gave a detailed survey on preprocessing techniques, ECG databases, feature extraction techniques, and ANN-based classifiers. In 2016, M. G. Egila, [1] Implemented FPGA- based ECG signal analysis using a least-square linear-phase finite impulse response filter. The main aim of this filter in this design is to remove low-frequency noise in the ECG signal. But in the classification part, only one node is used at the output, it will give the information only on whether the ECG signal is normal or abnormal.

In 2019, Zena N. Abdelkader, [3] proposed a design of ECG Classification using Xilinx System Generator. In this paper, an ANN classifier with two layers and four neurons with an activation function of type "tan-sigmoid" is implemented and provides 99% training accuracy, 80% validation accuracy, and 81.9% of Testing accuracy. In 2014, K. Muthuvel, [4] proposed a design to extract and classify the ECG signal using Harr Wavelet Transform and Neural Network. The only simulation result is carried out in this paper with 62% of accuracy, 73% of sensitivity, and 63% specificity.

In this work, hardware implementation of Classifier is carried out with 20 input, 10 hidden, and 4 output neurons. The sigmoid function is used for the hidden layer with piece-wise linear approximation and the Softmax function is used for output layer classification. Accuracy has been improved in MATLAB simulation and the same functionality is implemented in Verilog for hardware.

The rest of the paper is organized as follows. Section 2 explains the hardware implementation of ANN. Section 3 illustrates simulation results in MATLAB and Verilog and Error estimation. Finally, Section 4 concludes the paper.

II. HARDWARE IMPLEMENTATION OF ANN

The main aim of this work is to implement the ANN in Verilog after the ANN has been trained in MATLAB (offline) [10]. After testing the ANN using several test cases offline and measuring the value of accuracy, the same feed-forward ANN algorithm will be implemented in hardware using any one of the HDL languages and this result will compare with the MATLAB result. The weight and biases values are obtained, after training the ANN offline. These values are loaded in a text file and will be called to the main module through the test bench. Since all the values are in real numbers, they are converted to integers by multiplying with the scaling factor 1024 and rounding off that number, because Verilog will take only integers as input and will get integers as output. After getting the final output every value will be right-shifted by 10. Since at the input side all values are multiplied by 1024.

The ANN feed-forward algorithm consists of multipliers and adders. After the multiplication of inputs with corresponding weights, the accumulated sum will give as input to the activation function. The ANN model for this project is shown in Table I. This model consists of a total of 14 neurons, 10 for the hidden layer and 4 for the output layer.

Table I. ANN Model of this paper

Number of classifications	4
Number of training samples	1000
Number of input neurons	20
Number of hidden neurons	10
Number of output neurons	4

From Table I, it requires 20 multipliers 20 adders for each neuron in the middle layer, and 10 multipliers and 10 adders for the last output layer. In total it requires 240 multiplications and 140 additions. And this cost increases if the number of input neurons or the number of hidden neurons increases significantly. In this implementation, all the neurons in the hidden layer are triggered at the same time as well as output neurons, so this ANN implementation is very faster.

The above approach requires more resources, as there are more multipliers and adders. There is another approach where a much less number of multipliers and adders are used compared to the above method. In this ANN architecture, each layer will execute one after the other. Each neuron in the corresponding layer computes the product of the input at the positive edge of the clock. In this way, one can reduce the number of multiplications as equal to the number of input neurons, but the number of clock cycles will be increased. For this work, 20 input neurons are available so, it will take a total of 20 clock cycles and 1 clock cycle for bias input, and to calculate the accumulation of product and activation function output it will take two more clock cycles, in total it requires 23 clock cycles for hidden layer alone and 13 clock cycles for the output layer. But the drawback of this method is it will take more time from input to output.

The feed-forward algorithm for ANN implementation consists of parallel processing capabilities of the hardware, where each node in the layer will trigger at the same time (positive edge of the clock). The hierarchy of ANN is shown in Fig. 4. The corresponding weight and bias values are loaded into the main module through the test bench.

This algorithm will start from the module main.v. All the weight and bias values are stored in the text file and will be read from the test bench. From the test bench, these values will pass to the main module. After processing the ANN algorithm, the result will send back to the testbench from the main module. Since in each layer there are different number of neurons, we created two different modules for these layers named hidden.v and output.v .

The modules hidden.v and output.v are instantiated in the module main.v . Total of 10 neurons in the hidden layer and 4 neurons in the output layer are instantiated. During the processing of the hidden layer, all the neurons will receive inputs and corresponding weights one at a time and perform multiplication operations on them, after storing the result in registers they all accumulated. The number of clock cycles it requires to perform to compute for the hidden layer is equal to the number of input values. The main.v module sends inputs and weights to the hidden layer neurons at each positive edge of the clock. The same procedure is followed in the output layer also.

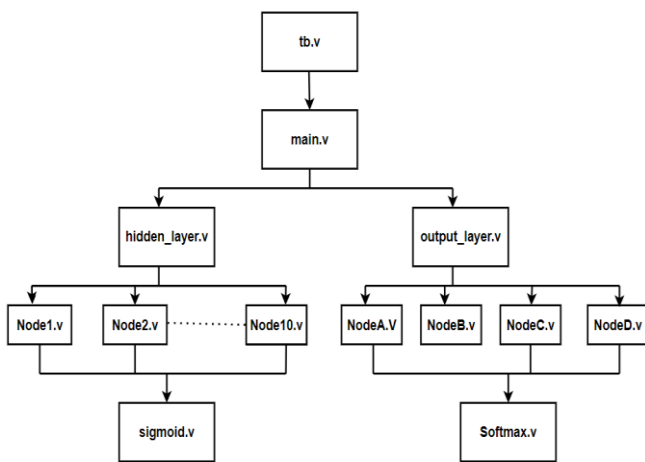


Fig 4. ANN Hierarchy

After generating the cumulative sum of the hidden layer the result will go to the sigmoid.v module , where the sigmoid activation function is performed and the result will store in the corresponding registers. And these values go to inputs to the next output layer. Like the hidden layer, the same process will perform in output_layer.v . This cumulative sum will go to the softmax.v , which is the activation function of the final layer, The output of this module is the final classification result of the ANN model.

The activation function used for the hidden layer is given by $f(x) = \frac{1}{1+e^{-x}}$ (3)

It is very difficult to implement exponential and division operations in Verilog and it will require more resources and more area [5]. The above function will be implemented by the

piece-wise linear approximation method [2]. In this method, the approximation is performed using the following expression (4). But this PLAN approximation will give the average error of 0.00587.

$$f(x) = \begin{cases} 1, & |x| \geq 5 \\ 0.03125 * |x| + 0.84375, & 2.375 \leq |x| < 5.0 \\ 0.125 * |x| + 0.625, & 1.0 \leq |x| < 2.375 \\ 0.25 * |x| + 0.5, & 0 \leq |x| < 1.0 \end{cases} \quad (4)$$

But the expression (4) contains real values, so one can convert the above expression by multiplying its terms without variable |x| with scalar factor 2¹⁰. The equation (4) is performed only with a positive value of x. Since sigmoid is an even function, the equation (5) is used for the negative value of x.

$$f(-x) = 1 - f(x) \quad (5)$$

For the final layer of the ANN model, the Softmax function is used as an activation function. The equation of the softmax is given by,

$$f(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (6)$$

Where,

z_i = input feature vector

K = number of output classes

This function represents the probability of that class. If it is getting more input value than the output corresponding to that class. But, as the equation contains exponential and division operators, it's very difficult to implement that function in hardware [6] and [7] and will utilize more resources. So, consider whatever value coming in the pre-activation function is high, given high probability to that corresponding neuron, for all remaining neurons considered as zero probability.

III. RESULTS AND DISCUSSION

In this study for Dataset classification, Intra-patient Data set [11] scheme is used. According to this, random subsampling is conducted to train and evaluate classifiers. Data is randomly selected from the whole database. To implement this work in hardware Xilinx Artix-7 NESYS 4 DDR FPGA board is used. In this work, training of the ANN model is done offline in MATLAB [10] and obtained the values of weights and bias values of corresponding layers.

Here, using nntool in MATLAB, the ANN model is trained with one hidden layer containing 10 nodes with a sigmoid activation function and 4 output nodes with a SoftMax activation function.

As shown in the flow diagram in Fig. 1, from DWT one can extract the features of ECG beat and these features are input to the ANN model. A total of 20 features are getting from each beat. From the database total of 1000, 20*1 vectors will be getting, in which 70% is used for training, 15% for validation, and the remaining 15% is used for testing for training the algorithm. And this result is tested with 1000 more data samples.

In the hidden layer total of 10 neurons are instantiating with each having a pre-activation function and sigmoid activation function. Fig. 5 shows the simulation of the hidden layer.

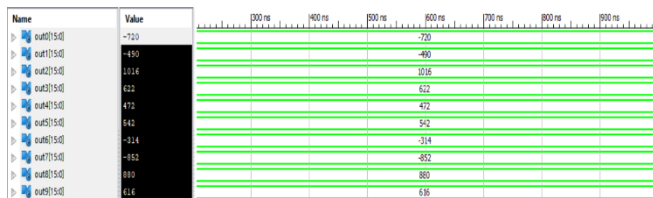


Fig 5. Verilog Simulation of Hidden layer

In the output layer total of 4 neurons are instantiating with each having a pre-activation function and softmax activation function Fig. 6 shows the simulation of the final layer.

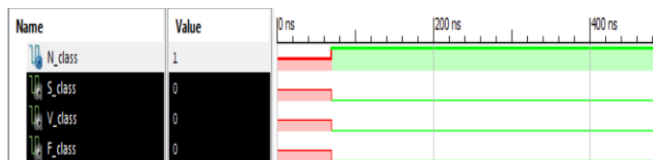


Fig 6. Verilog Simulation of Output Layer

The hardware utilization design summary is given in Table II.

Table II. Hardware utilization design summary

Logic Utilization	Used	Available	Utilization
Slice LUTs	84	63400	~1%
Fully used LUT pairs	0	84	0%
BUFGs	1	32	~3%
DSP48E1s	10	240	~4%

Since, in Verilog, it takes only integers and neglects fractional parts, there is some error produced compared to the MATLAB result. Table III, and IV shows the Error estimation of the hidden layer and output layer respectively. One random beat of sample is considered for this error estimation.

Using the confusion matrix from the MATLAB simulation, one can get the accuracy of the ANN model. From Fig. 7, it is noted that, 90.1% training accuracy, 84.7% validation accuracy, 85.3% testing accuracy, and 88.6% overall average accuracy when we take the total 1000 samples and are divided into 70% for training, 15% for validation, and 15% for testing. Once the model is trained, testing can be done using another 1000 random samples. The 86% of testing accuracy is obtained and shown in Fig. 8.

Table III. Error estimation of Hidden layer

	MATLAB	Verilog	Error
Node1	0.4704	0.4716	0.0012
Node2	0.8301	0.8232	0.0069
Node3	0.8309	0.8232	0.0077
Node4	0.2422	0.243	0.0008
Node5	0.2569	0.258	0.0011
Node6	0.9323	0.9248	0.0075
Node7	0.4512	0.4521	0.0009
Node8	0.9252	0.9218	0.0034
Node9	0.4735	0.4746	0.0011
Node10	0.2729	0.2705	0.0024
Overall average error of the random one sample			0.0033

Table IV. Error estimation of the Output layer

	MATLAB	Verilog	Error
Node1	3.5239	3.5509	0.0270
Node2	0.9593	0.9590	0.0003
Node3	-0.9900	-0.9872	0.0028
Node4	-0.5966	-0.5964	0.0002
Overall average error of the random one sample			0.0075

From the Table III and Table IV, it may be noted that the overall average error 0.0033 and 0.0075 are obtained from hidden layer and output layer respectively.

After verifying this, the hardware implementation using 1000 testing samples are carried out and the results are shown in Table V. From the Table V, it may be noted that 85.6% of testing accuracy is obtained. Out of 1000 samples, 856 samples are correctly classified and 144 samples are not correctly classified.

IV.CONCLUSION

Artificial Neural Network (ANN) is one of the main research topics nowadays everyone working with. For this biomedical system, ANN is used as a classifier with one hidden layer consisting of 10 nodes and an output layer consisting of 4 nodes corresponding to 4 classes. Piecewise Linear Approximation (PLAN) is used to implement the sigmoid activation function with an average error of 0.00587. And the accuracy observed for classification is 86% in simulation and 85.6% in hardware.

V.REFERENCES

- [1] M. G. Egila, Magdy A. El-Moursy, "FPGA-based electrocardiography (ECG) signal analysis system using least-square linear-phase finite impulse response (FIR) filter", Electronics Research Institute (ERI), Elsevier, 2016.
- [2] Ivan Tsmots, "Hardware Implementation of Sigmoid Activation Functions using FPGA", IEEE 2019.
- [3] Zena N. Abdulkader, "Implementation of ECG Classification Using Xilinx System Generator", IEEE 2019.
- [4] K. Muthuvel, "ECG Signal Feature Extraction and Classification using Harr Wavelet Transform and Neural Network", ICCPCT,2014
- [5] P. Kumar Meher, "An optimized lookup-table for the evaluation of sigmoid function for artificial neural networks," 2010 18th IEEE/IFIP International Conference on VLSI and System-On-Chip, 2010, pp. 91-95, doi: 10.1109/VLSISOC.2010.5642617.
- [6] Ioannis Kouretas, "Hardware Implementation of a SoftMax-Like Function for Deep Learning", e 8th International Conference on Modern Circuits and Systems Technologies (MOCASST),2018
- [7] X. Dong, X. Zhu and D. Ma, "Hardware Implementation of Softmax Function Based on Piecewise LUT," 2019 IEEE International Workshop on Future Computing (IWOFC, 2019, pp. 1-3, doi: 10.1109/IWOFC48002.2019.9078446.
- [8] S. H. Jambukia, V. Dabhi and H. B Prajapati, "Classification of ECG signals using machine learning techniques: A survey", International Conference on Advances in Computer Engineering and Applications (ICACEA), March 2015.
- [9] FPGA design flow, Copyright © 2008, Xilinx® Inc
- [10] Howard Demuth, Mark Beale "Neural Network Toolbox User's Guide"
- [11] Danni Ai, Jian Yang, et al., "Fast multi-scale feature fusion for ECG heartbeat classification", EURASIP Journal on Advances in Signal Processing (2015) 2015:46 DOI 10.1186/s13634-015-0231-0
- [12] Xinghuo Yu, M. O. Efe and O. Kaynak, "A general backpropagation algorithm for feedforward neural networks learning," in IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 251-254, Jan. 2002, doi: 10.1109/72.977323.

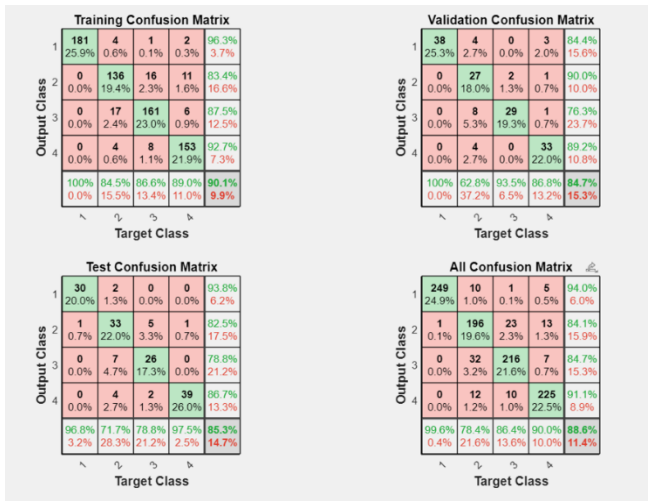


Fig 7. Confusion matrix of training, validation, and testing

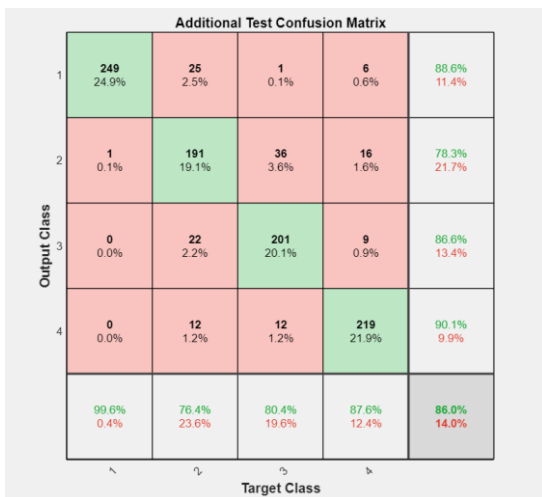


Fig 8. Confusion matrix for only testing samples

Table V. Testing accuracy in Verilog

	Mis classified samples	Correctly classified samples	Accuracy
N class	10	240	96%
S class	51	199	79.6%
V class	47	203	81.2%
F class	36	214	85.6%
Total Testing Accuracy	144	856	85.6%

Dynamic Modelling and Analysis of Solar Powered Reverse Osmosis Desalination System for Pakistan using the Bond Graph Model

Sheikh Usman Uddin
Electrical and Computer Engineering
Memorial University of Newfoundland and Labrador
 St. John's, Canada
 suddin@mun.ca

Dr. Geoff Rideout
Engineering and Applied Science
Memorial University of Newfoundland and Labrador
 St. John's, Canada
 g.rideout@mun.ca

Abstract— With the advancement of technology and increase in world's population, the demand for the consumption of fresh drinking water has increased worldwide. Utilization of clean energy sources for purification of water is the need of today's world. This paper focuses on modelling the solar powered reverse osmosis system using the bond graph modelling technique which is very effective for multidisciplinary systems. The model contains electrical, mechanical and hydraulic domains merged together using the bond graph language. It further utilizes the 20-Sim software to implement the dynamic model and simulate the system. The paper eventually provides a detailed analysis on how the system will respond to the changing system parameters.

Keywords— *Solar, Reverse Osmosis, Bond Graph, Dynamic Modelling, Simulation, 20sim Software*

I. INTRODUCTION

Pakistan economy mostly relies on agriculture and around 22.2% of that contributes to its overall gross domestic product (GDP) [1]. Approximately 42.3% of the labor in Pakistan is employed from the total labor strength [2]. In 1950s Pakistan has the water capacity of around 5000 m^3 per capita and it has decreased now to 1000 m^3 per capita [2]. According to the United Nations Educational, Scientific and Cultural Organization report, Pakistan available water supply is just above 1000 m^3 per capita which puts the country in the list of water stress countries [3]. Overall worldwide, the countries that are facing difficulty in availability of freshwater are now relying on water desalination processes using ground water as this method allows to generate fresh water supply but at the cost of electricity consumption. According to an estimate it take 10,000 ton/yr. of crude oil to produce 1000 m^3 /day of desalinated water from the process of thermal desalination [4]. This energy requirement is a big challenge as the world is already heading towards a big global climate change. The renewable energy is the solution to that problem and solar powered reverse osmosis plants can be very beneficial when it comes to elimination of utilization of fossil fuels. Hence renewable technology based reverse osmosis (RO) systems can be a good source of sustainable fresh water [5]. Thankfully, Pakistan lies on the region of good solar irradiance ranging from 5–7 kWh per m^2 per day and sunshine hours of 1,500–3,000 [6] which can be used as an

advantage for solar powered RO systems. The desalination process now is used worldwide and its utilization trend has shown an increase in the recent years [7]. The Solar powered reverse osmosis is a unique idea for converting the saline or brackish water to fresh water as the water passes through membranes and remove approximately 98% - 99.5% salt from it [8]. With all of these factors being said and issues that Pakistan as a country has, powering a reverse osmosis system for extracting drinking water for a community in Pakistan using renewable energy sources makes a very favorable solution. Modelling of such multi-disciplinary systems which involves electrical, mechanical and hydraulic domains require complex modelling techniques. This paper utilizes the bond graph modelling approach which facilitates different domains using the explicitly graphical power flow paths among interconnected system elements, and leverage analogies among different energy domains such as thermal, fluid, mechanical, and electrical. This modelling technique is then utilized to understand the behavior of desired outputs in relationship to the changing parameters of the system.

II. LITERATURE REVIEW

In this section several papers were reviewed related to the photovoltaic (PV) based reverse osmosis system and their conclusions are discussed. S. Sobana and Rames C. Panda [9] have reviewed more than 65 literatures for identifying different process parameters, dynamic modelling and control of desalination system. Their work mainly gives an account of two types of phenomenological models of desalination processes, namely, mechanistic model or membrane transport model and lumped parameter model. Many researchers have presented identified transfer functions from input – output data of reverse osmosis process. Alatiqi et al. (1989) [10] used system identification techniques to estimate a MIMO structure of RO plant at Doha. Assef et al. (1995), Riverol and Pilipovic (2005), and Robertson et al. (1996) [11] [12] [13] also developed multivariable transfer function models from the plant's input – output data. Zilouchian (2001) [14] worked and find the reverse osmosis desalination system transfer function and also in steady state matrix form using recursive least square method. Ramaswamy et al. (1995) [15] implemented connected multilayer feed forward neural network using the back propagation algorithm to identify the nonlinear

multivariable multistage flash desalination plant. Fkirin et al. (1997) [16] presented an algorithms which focus on optimal identification for the timevarying dynamic process based on linear combination of the recursive least square method. Saengrung et al. (2002) [17] modeled two reverse osmosis plants using system identification. Gambier et al. (2007) [18] derived a lumped parameter dynamic MIMO model from first principle laws and used it in carrying out diagnostics of system and even in finding faults in system. Ahmad et al. (2007) [19] developed and simulated a membrane transport model suitable for the multiple solutes system in reverse osmosis for unsteady-state condition. Chaaben et al. (2008) [20] developed a MIMO model relating input and output variable for a small photovoltaic reverse osmosis desalination unit. All these research approach provides a great insight on how to model the reverse osmosis system.

III. BOND GRAPH LANGUAGE

Due to the increasing number of sophisticated and interdisciplinary systems, multidisciplinary techniques are becoming increasingly crucial in the modelling process. Bond graph approach establishes a common vocabulary for understanding relationships and similarities among various modelling methodologies. It provides a uniform domain-independent graphical representation of multidisciplinary models [21]. Bond graph provides a clear graphical representation from which other representations can be derived, e.g. linearized state equations, system transfer functions, or differential equations. The bond graph is a modelling technique that is based on energy transfer between system components. Bonds connect the component ports, demonstrating how power is transferred between them. A half-arrow on one end of the bond indicates the direction of power transfer. In a bond graph, there are two variables: effort (e) and flow (f). The system, which is composed of many energy domains such as mechanical, electrical, and hydraulic, is turned into a single model in order to simplify the system's analysis. The figure 1 shows analogies of different bond graph variables utilized in different domains [22].

Energy Domain	Effort (e)	Flow (f)	Generalized Momentum	Generalized Displacement
Translational Mechanics	Force	Velocity	Momentum	Displacement
Rotational Mechanics	Torque	Angular Velocity	Angular Momentum	Angle
Electro-magnetic	Voltage	Current	Flux Linkage	Charge
	Magneto-motive Force	Magnetic Flux Rate	-----	Magnetic Flux
Hydraulic	Total Pressure	Volume Flow	Pressure Momentum	Volume
Thermodynamic	Temperature	Entropy Flow	-----	Entropy
Chemical	Chemical Potential	Molar Flow	-----	Molar Mass

Figure 1 – Analogies of different bond graph variables utilized in different domains [22]

IV. SYSTEM DESCRIPTION

The overall system will comprise of an electrical, mechanical and hydraulic system. The figure 2 below shows the complete block diagram of the system.

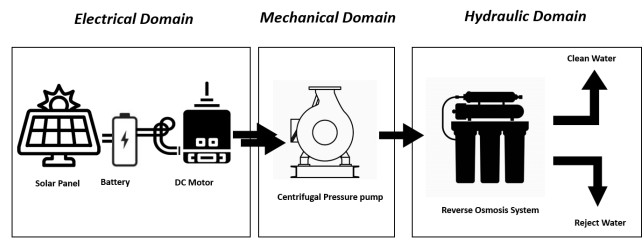


Figure 2 –Block Diagram of Complete System

A. Electrical System

The electrical system will use the energy of the sun and by using the photovoltaic cell convert that energy to electrical direct current form. This energy will be fed to the batteries through charge controllers so that it can be used to provide power to the reverse osmosis system. The major load of reverse osmosis system is the dc motors that is coupled with the centrifugal pressure pump to run the hydraulic system. The block diagram for a solar powered dc motor is shown in figure 3 [23].

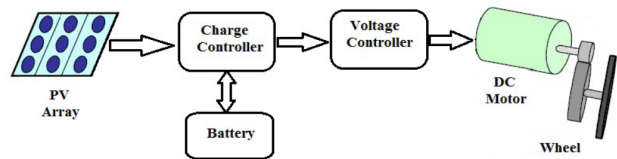


Figure 3 –Block Diagram of Solar Powered DC motor [23]

B. Reverse Osmosis Desalination System

Reverse osmosis system converts unfiltered feed water to clean water when feed water is pressurized and passed through a membrane. The water flows through the membrane and more contaminants are accumulated to the more concentrated side of the membrane whereas clean drinking water is accumulated to less concentrated side of the membrane. The cleaned water is then stored in a storage tank and usually the system is designed in a way that when the storage tank is full the RO system shut downs automatically. The brine water is drained as a water waste [24]. The brine water can be used for dishwashing and other cleaning purpose so that system efficiency is increased. The basic components of a RO process are: pre-treatment system used for cleaning brackish water, High-Pressure Centrifugal (HP) pump used to provide required pressure to membrane, and post-treatment system for smell and odor removal [25]. The overall block diagram of reverse osmosis system is shown in figure 4.

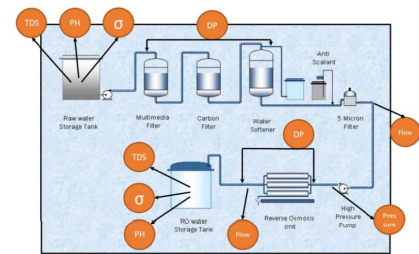


Figure 4 –Block Diagram of Reverse Osmosis System

V. SYSTEM MODELLING

A. Generalized Bond Graph Model of System

The generalized interrelation of multidisciplinary system can be seen in figure 5. The distinct systems are connected with the power variable defined as effort and flow. Electrical domain output of motor torque and motor speed is used as the input to the mechanical domain. The mechanical domain outputs the Pump Pressure and pump flow rate that is used as the input to the hydraulic domain and eventually producing clean and reject water.

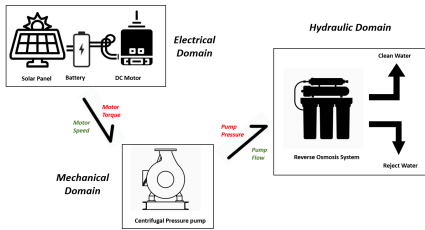


Figure 5 –Generalized Bond Graph Model of System

B. Detailed System Based Bond Graph Modelling

1) Electrical System Bond Graph Modelling

In order to reduce the complexity of the system, the solar panel and battery are modelled as a voltage source. The rated voltage of battery is 24V DC. The DC motor model is created in which motor resistance and inductance are modelled as R-element and I-element respectively. The conversion from electrical domain to mechanical domain is executed with the help of gyrator that relates effort to flow with the motor constant (K_m). The selected parameters of the model are shown in table 1. The bond graph model of the system is shown in figure 6.

Table 1 - Electrical Model Parameters

Element	Description	Value	Unit
Se	Voltage Source	24.0	V
R	Winding Resistance	5.0	Ohm
I	Winding Inductance	0.03	H
GY	Motor constant	0.02848	Nm/A

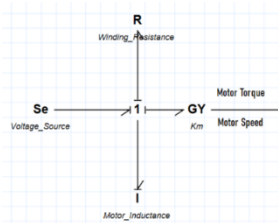


Figure 6 –Bond Graph Model of Electrical System

2) Pressure Pump Bond Graph Modelling

The pump is defined as a machine used to generate a pressure differential in order to propel liquid through a piping system from one location to another. An electric motor rotates an impeller in a centrifugal pump that adds energy to the water after it is directed into the core of the

rotating impeller. The pump is coupled directly to the motor and the load of the pump is modelled using the equations (1) and (2) [26] [27].

$$T_l = (K_1 \cdot \omega - K_2 \cdot Q) \cdot Q \tag{1}$$

$$P = (K_1 \cdot \omega - K_2 \cdot Q) \cdot \omega \tag{2}$$

Where:

T_l is the load torque of the pump

P is the pressure of the fluid

ω is the rotational speed of the machine

Q is the flow rate of the pump

K_1 and K_2 are the pump parameters

The modelling of a single stage pump is done in the bond graph by the help of the gyrator element whereas R-element represent losses due to impeller/diffuser wear and fluid disk friction, I-(inertia) element models the rotating fluid of the pump and another R-element represents the pipe loss when water is transmitted from pump discharge to the reverse osmosis system inlet. The selected parameters for the model are shown in table 2. The bond graph model of the system is shown in figure 7.

Table 2 - Pump Model Parameters

Element	Description	Value	Unit
GY	Centrifugal Pump Equations Modelled	Pressure & Flow rate	Bar & GPM
R	Impeller Loss	0.05	Bar
R	Pipe Loss	0.0005	Bar
I	Rotating Fluid Inertia	0.125	kgm ²

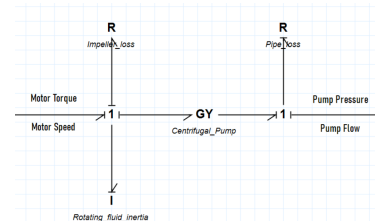


Figure 7 –Bond Graph Model of Centrifugal Pump

3) Membrane Modelling

The reverse osmosis system works opposite to the phenomena of osmosis system. By applying the external pressure which is from the centrifugal pump clean water and dirty water are separated. The feed water pressure from the pump must be large enough to overcome the osmotic pressure and the membrane resistance, as well [28] [29]. The flow inside the membrane is cross in nature resulting in clean water being diffused to the other side of membrane and all brine (dirty) water being collected on the other side. Figure 8 shows the water flow inside the membrane [30].

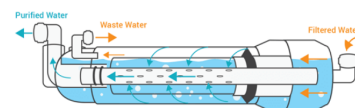


Figure 8 – Water flow inside reverse osmosis membrane [30]

Therefore, a quantity of the feed water permeates through the membrane reducing strongly the water salt concentration to get the fresh water (purified water), and the remaining feed water becomes very concentrated brine (waste/dirty) water. It can be seen that the feed water flow is used by both the clean water and rejected water. To reduce this complex modelling the system is converted to complete hydraulic system. The system is modelled using a C-element that represent the water storage under pressure inside the membrane. The membrane is considered as a tank which has radius 0.025 meters with water density of 1000 kg/m³. The area was calculated using equation (3) and used to calculate the value of capacitance using equation (4).

$$A = \pi \cdot r^2 \tag{3}$$

$$C = \frac{A}{\rho \cdot g} \tag{4}$$

Where A is the area in m^2 , π is the constant with value of 3.142, r is the radius of the membrane, C is the capacitance value of the tank, ρ is the density of water and g is the acceleration due to gravity. Two R-elements are used to model the hydraulic membrane resistance on the clean water side and on dirty water side which depends highly on the membrane temperature and conductivity. To control the pressure of reject water, a non-linear valve is modelled using the R-element. The valve is modelled using the valve law equation (5) [31].

$$Q = \frac{P^2 \cdot \rho}{2 \cdot (C_d \cdot A_0)^2} \tag{5}$$

Where:

- Q is the flow rate of the valve
- P is the pressure across valve
- ρ is the density of the water
- C_d is the discharge coefficient of the valve
- A_0 is the nominal area of the valve

The clean water pressure, hydraulic load losses and osmotic pressure are all modelled using the Se-element with its summation. In order to keep the dirty water at required pressure Se-element is also used to model that. The selected parameters for the model are shown in table 3. The generalized bond graph model of the system can be seen in figure 9.

Table 3 – Reverse Osmosis System Parameters

Element	Description	Value	Unit
C	Membrane Water Storage	2.0e-07	m ³ /bar
R	Clean water resistance	1.0	bar
R	Reject water resistance	0.0005	bar
R	Reject water valve	Eq. (7)	bar
Se	Clean water	5.0	bar
Se	Reject water	2.0	bar

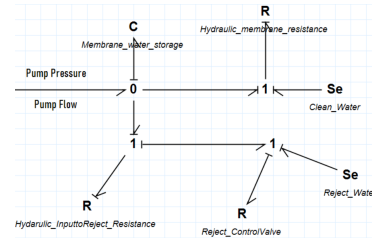


Figure 9 –Bond Graph Model of Reverse Osmosis System

C. Complete System Model

After combining all the individual systems, the complete bond graph is established as shown in figure. The multi-disciplinary systems (electrical, mechanical and hydraulic) are all combined with their effort to flow relationship as shown in figure 10.

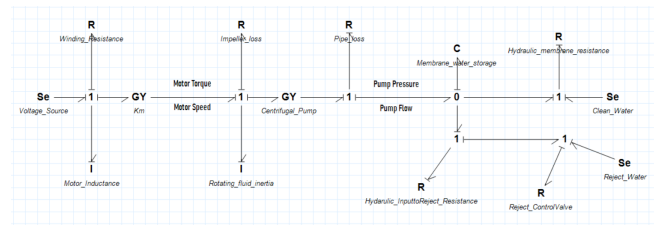


Figure 10 –Complete System Bond Graph Model

VI. ANALYSIS AND RESULTS

The system was modelled in 20sim software and using the simulation feature, dynamic response was analyzed after which system reaches the steady state response. The normal response of the system was shown in the first part. The comparative results are shown later. In comparative parts the dark color line represents the system response based on changing the system parameter whereas the light color line represents the normal system response.

A. Normal System Dynamic Response:

Figure 11 shows that the source voltage is stabilized at 24V DC and the current reaches the maximum value of 5 Ampere. Figure 12 shows that the motor torque is stabilized at 1.13Nm and the motor angular speed reaches a value of 1.25 rad/s. Figure 13 shows that at the start the pump pressure becomes negative and the flow rate is also disturbed but once the pump reaches its capacity a pressure of 0.15 bar is achieved and flow rate of 5 GPM is stabilized. Figure 14 and 15 shows that the pressure of clean and reject water is maintained at 5 bar and 2 bar where as the flow rate reaches a value of 5.2 GPM and 0.0004 GPM respectively which is optimum condition as there is minimum dirty water and the system is operating at the most efficient point.

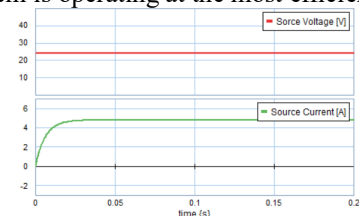


Figure 11 – Source Voltage and Current Response

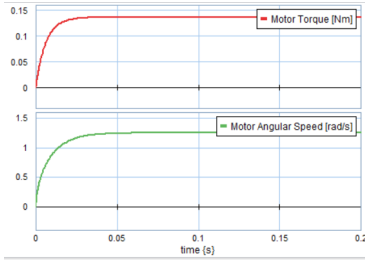


Figure 12 – DC Motor Torque and Angular Speed Response

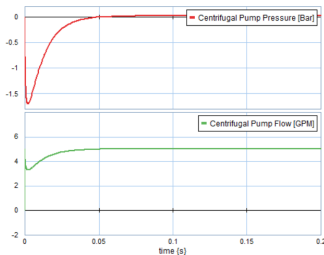


Figure 13 – Centrifugal Pump Pressure and Flow rate Response

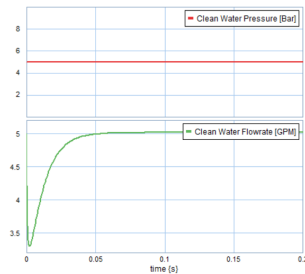


Figure 14 – Clean Water Pressure and Flow rate Response

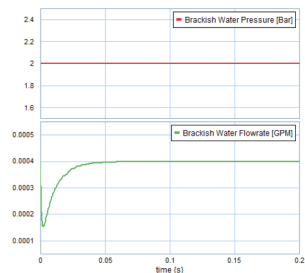


Figure 15 – Dirty Water Pressure and Flow rate Response

B. Voltage Reduced due to Intermittent Solar Source (Comparison with normal response):

The voltage source is now reduced and dynamic response is analyzed compared to the actual dynamic response of the system. This happens because the state of charge of battery is not fixed and dependent on the charging and discharging of battery which in turn depends on the availability of solar energy and conversion efficiency of the photovoltaic system. Figure 16 showed that now the voltage is reduced to 20V DC and the current is also reduced. Figure 17 verifies that torque of motor is dependent on the voltage and it can be seen that torque is reduced whereas the angular speed of the motor is still the same.

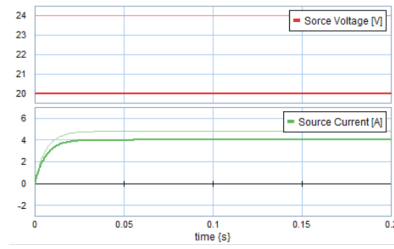


Figure 16 – Source Voltage and Current Response

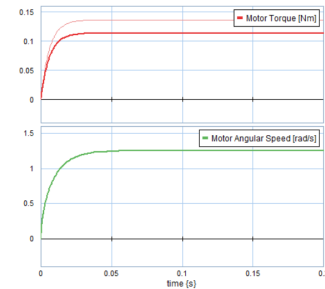


Figure 17 – DC Motor Torque and Angular Speed Response

C. Increased Clean water side hydraulic resistance (Comparison with normal response):

The clean water side hydraulic resistance is increased and the dynamic responses are analyzed as shown in fig 18, 19, 20 and 21. It can be seen that the motor speed has drastically reduced because of the load on the impeller. Also the pump now has low negative pressure and reaches the stability quickly but due to high hydraulic resistance the flow rate of pump is now reduced significantly resulting in very low flow rate of clean water and same flow rate of dirty water.

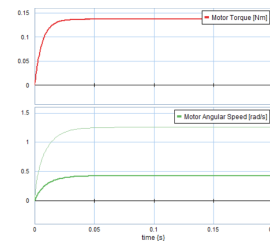


Figure 18 – DC Motor Torque and Angular Speed Response

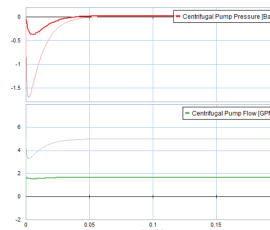


Figure 19 – Centrifugal Pump Pressure and Flow rate Response

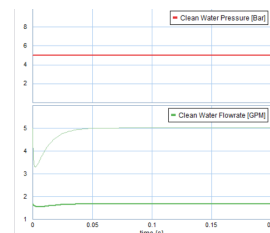


Figure 20 – Clean Water Pressure and Flow rate Response

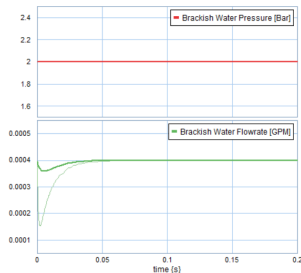


Figure 21 – Dirty Water Pressure and Flow rate Response

VII. CONCLUSION

Based on the results, it can be concluded that the system has transient and steady state response. The transient response is finished after a time period of 50 – 80 milli seconds. The paper uses mathematical models of multidisciplinary system and merge them as single model using bond graph language. Simulation is performed using the 20Sim software and the results concluded that if voltage source is decreased due to intermittent nature of renewable energy system the motor torque is also affected. Also when the hydraulic resistance of clean water side is increased the system becomes highly inefficient as all the water is thrown to the dirty side rather than being cleaned in reverse osmosis membrane. The practical understanding is that when the membrane is clogged with impurities the resistance to flow is increased resulting in poor performance of system which is visible from the modelling results. A differential pressure sensor can be used with alarm values to indicate an operator managing the system so that he can clean the membrane and efficiently run the system.

REFERENCES

[1] A. Khan and N. Awan, "Inter-Provincial Water Conflicts in Pakistan: A Critical Analysis," *J. South Asian Middle East. Stud.*, vol. 43, no. 2, pp. 42–53, 2020, doi: 10.33428/jsoutasiamideas.43.2.0042

[2] B. Ahmad, "Water Management: A Solution to Water Scarcity in Pakistan," *J. Indep. Stud. Res.- Manag. Soc. Sci. Econ.*, vol. 9, pp. 111–125, Dec. 2011, doi: 10.31384/jismss/2011.09.2.9.

[3] UNESCO, *The 2nd UN World Water Development Report: Water, A Shared Responsibility*, New York, 2006, p. 134

[4] S.A. Kalogirou, *Seawater desalination using renewable energy sources*, *Prog. Energy Combust. Sci.*, 31 (2005) 242–281.

[5] N. Ghaffour, J. Bundschuh, H. Mahmoudi, M.F.A. Goosen, *Renewable energy-driven desalination technologies: a comprehensive review on applications of integrated systems*, *Desalination*, 356 (2015) 94–114.

[6] M.S. Khalil, N.A. Khan, I.A. Mirza, *Renewable Energy in Pakistan: Status and Trends*, University of Engineering and Technology, Taxila-Pakistan Alternate Energy Development Board, Islamabad, Pakistan, 2003.

[7] C. Fritzmann, J. Löwenberg, T. Wintgens, T. Melin, *State-of-the art of reverse osmosis desalination*, *Desalination*, 216 (2007) 1–76.

[8] M. Wilf, *Fundamentals of RO-NF Technology*, Proceedings of International Conference on Desalination, Limassol, Cyprus, 2004.

[9] S. Sobana and Rames C. Panda, *Identification, Modelling, and Control of Continuous Reverse Osmosis Desalination System: Science and Technology*, 46: 551–560, 2011 DOI: 10.1080/01496395.2010.534526

[10] Alatiqi, I.M.; Ghabris, (1989) *System identification and control of reverse osmosis desalination*. *Desalination*, 75 (1–3): 119–140.

[11] Assef, J.Z.; Watters, J.C.; Desphande, P.B. (1995) *Advanced control of a reverse osmosis desalination unit*. Proc. International

Desalination Association (IDA) World Congress, Abu Dhabi, Vol. V, pp. 174–188.

[12] Riverol, C.; Pilipovik, V. (2005) *A reverse osmosis desalination industrial-scale unit*. 2: 50–54.

[13] Robertson, M.W.; Watters, J.C.; Desphande, P.B. (1996) *Model based control for reverse osmosis desalination processes*. 104 (1–2): 59–68.

[14] Zilouchian, A.; Jafar, M. (2001) *Automation and process control for a reverse osmosis plant using soft computing methodology*. *Desalination*, 135: 51–59.

[15] Ramaswamy, S.; Deshpande, P.B.; Tambe, S.S.; Kulkarni, B.D. (1995) *Neural networks for the identification of MSF desalination plant*. *Desalination*, 101 (2): 185–193.

[16] Fkirin, M.A.; Al Madhair, A.F. (1997) *Prediction of time varying dynamic processes*. *International Journal of Quality and Reliability Management*, 14 (5): 505–511.

[17] Saengrung, A. (2002) *Modeling of reverse osmosis plants using system identification and neural networks*. *Journal of Master abstract International*, 41–03: 0856.

[18] Gambier, A.; Krasnik, A.; Badreddin, E. (2007) *Dynamic modelling of a RO desalination plant for advanced control purposes*. *American Control Conferences*, 7 (9): 4854–4859.

[19] Ahmad, A.L.; Chang, M.F.; Bhatia, S. (2007) *Mathematical modelling of multiple solutes systems for reverse osmosis process in palm mill effluent*. *J. Chemical Engineering*, 132 (1–3): 183–193.

[20] Chaaben, A.B.; Andouls, R. (2008) *MIMO modelling approach for a small photovoltaic reverse osmosis desalination system*. *Research Unit RME, INSNT North Urban Centre, BP 676, 1080, Tunis, Tunisia*

[21] Mehdi TURKI, Jamel BELHADJ, Xavier ROBOAM; *Bond Graph modelling and analysis of an autonomous Reverse Osmosis desalination process fed by a hybrid system*, the 9th International Conference on Modeling and Simulation of Electric Machines, Converters and Systems, Québec, Canada, 2008.

[22] Borutzky, W., 2010, "Bond Graph Methodology: Development and Analysis of Multi Disciplinary Dynamic System Models," Springer-Verlag London Limited.

[23] Das S, Sadhu P. K, Chakraborty. *Design and Implementation of A PV Powered Tri-Cycle*. *Curr World Environ* (2016)

[24] *What is Reverse Osmosis System and how does it work?*, *Fresh Water Systems*, <https://www.freshwatersystems.com/blogs/blog/what-is-reverse-osmosis> (Accessed 10th Nov 2021)

[25] A. Maurel, "Technologie et Application", *Techniques de l'ingenieur*, pp.4- 13.

[26] A.Ben Rhouma, J.Belhadj and X.Roboam, " Design and control of a pumping system fed by hybrid Photovoltaic-Wind source without battery storage". *International conference on Electrical Engineering Design and Technologies (ICEEDT) November 5-6, 2007, Hammamet, Tunisia*.

[27] Ryan Ratliff: « Modelling of vertical centrifugal Pumps », *University of Texas*, p.269.

[28] A. Maurel, "Dessalement de l'eau de mer et des eaux saumâtres: Et autres procédés non conventionnels d'approvisionnement en d'eau douce", 2nd edition, Lavoisier, 2006.

[29] A. Maurel, "Dessalement de l'eau de mer et des eaux saumâtres: Et autres procédés non conventionnels d'approvisionnement en d'eau douce", 2nd edition, Lavoisier, 2006.

[30] *5Pack 50GPD reverse osmosis system membrane water filter; Maxwater;* https://www.maxwaterflow.com/5-Pack-50-GPD-Membrane-Reverse-Osmosis-Max-Water-Filter-Universal-RO-System-NSF_p_1404.html (Accessed 19th April 2022)

[31] *Valve Sizing Calculation; Emerson Technical;* <https://www.emerson.com/documents/automation/manual-valve-sizing-standardized-method-fisher-en-140724.pdf> (Accessed 19th April 2022)

Design and Analysis of a Solar Powered Water Filtration System for a Community in Black Tickle-Domino

Sheikh Usman Uddin
Electrical and Computer Engineering
Memorial University of Newfoundland and Labrador
St. John's, Canada
suddin@mun.ca

Abdul Azeez
Engineering and Applied Science
Memorial University of Newfoundland and Labrador
St. John's, Canada
aazeez@mun.ca

Onyinyechukwu Chidolue
Engineering and Applied Science
Memorial University of Newfoundland and Labrador
St. John's, Canada
oachidolue@mun.ca

Dr. Tariq Iqbal
Electrical and Computer Engineering
Memorial University of Newfoundland and Labrador
St. John's, Canada
tariq@mun.ca

Abstract— The demand for fresh drinking water has increased globally due to technological advancements and an increase in the world's population. The world is also witnessing climate change due to excessive emissions from conventional power generation procedures. Today's world necessitates the use of renewable sources for water purification. Using the HOMER Pro software, this study elaborates on a design for a solar-powered drinking water reverse osmosis system for a community in Black Tickle-Domino. This study also uses the HOMER Pro software's optimization feature to conduct an economic analysis to develop the most cost-effective system design. The Steady-state analysis was conducted in HOMER Pro Software whereas the dynamic modelling and analysis of the proposed design was also validated in MATLAB Simulink. Also a comprehensive instrumentation design with important protection controls schemes is designed to ensure system stability and reliability for operation.

Keywords— Solar, Reverse Osmosis, Water Desalination, HOMER Pro Software, MATLAB, Instrumentation and Protection.

I. INTRODUCTION

The human body needs water more than it needs anything else. It is surprising to hear that we will die of thirst quicker than we would die of hunger. It is equally surprising to hear that, in any given month, well over 100 communities in Newfoundland and Labrador do not have continuous access to clean, safe drinking water [1]. The Indigenous community of Black Tickle-Domino, located in Labrador, does not have piped-in water and forcing the 126 residents of the community to travel around 2 kilometer's to retrieve water. The community's regular water supply is several unmonitored local streams, brooks, and ponds. A Potable Water Dispensing Unit (PWDU) was installed in Domino in 2004.

Still, access to water was strictly limited and expensive, with residents paying up to two dollars per liter of drinking water. The PWDU is unsustainable without consistent government funding [2]. Aside from water, the high cost of transportation, such as snowmobile gas, is another significant deterrent to continuous PDWU operation. Water retrieval from a stream about 25 kilometers away is hindered by adverse weather, substantial snowfalls, and storms [3]. Local shallow water holes known as 'wells' arise as water sources, while some wells become inaccessible due to snow covering in the winter and spring. Many residents continue to use untreated brooks and still ponds for their water needs, as they did before installing the PWDU [3]. It is important to note that this is an issue concerning drinking water and severe mental and physical health risks associated with using untreated or inadequately treated water. This project mainly aims at designing and simulating a PV System to power a community-size water filtration system to solve the freshwater scarcity issues. The fresh water-scarce areas in the world now rely on water desalination processes. Desalination has emerged as a potential method to produce freshwater in recent times due to technological advancement. Keeping this in mind, a small-scale solar-based filtration system seems to be the most feasible, environment-friendly, and accessible solution for the continuous availability of water.

II. LITERATURE REVIEW

In this section, several papers related to the photovoltaic-based reverse osmosis system were reviewed, and their conclusions were discussed. The article published on the MIT website [4] in 2013 elaborates how a small Mexican village produces clean water with a solar-powered system. The system runs autonomously, producing 1,000 liters per day. The solar-powered setup can produce at least 1,000 liters of drinking water on a cloudy day - enough for 450 residents [4].

A simulation-based RO design system was used to find the performance of the solar-powered RO system. Three Pakistan cities were assessed, including Lahore, Hasil Pur and Faisalabad [5]. According to WHO quality requirements, the TDS concentrations for these three cities were reduced from 1495 to 295.44, 2190 to 237.69 and 7683 to 241.98, respectively [5]. The required energy turns out to be 60, 95, and 311 kWh/month for Lahore, Hasil Pur, and Faisalabad, respectively. The paper used PVsyst software for solar calculation and deduced that 19, 15, and 40 PV panels would be used by Lahore, Hasil Pur, and Faisalabad, respectively [5]. Another study was conducted for a photovoltaic-based reverse osmosis system in which a 500 L/hr. system was designed using a 2KW PV system with a 5KVA hybrid inverter. Analysis and comparison of cooling and no-cooling of the PV system were also carried out. It was concluded that around 18% more daily PV energy was utilized using the tracking system, and 10% more PV energy was used when the panels were cooled [6]. Abdel Kareem's [7] analysis was based on capturing desalination water using solar, wind and geothermal energy. His conclusion defined that good design is very important when off-grid PV systems are designed. He further concluded that solar tracking and cooling of PV cells eventually make the process efficient. Ali et al. [8] did a further analysis of combining reverse osmosis and adsorption desalination systems. He further used MATLAB for simulations of different scenarios. His work concludes that if both are run together, the permeate salinity is slightly decreased, which will make the process more efficient. Charcosset [9] also carried out his analysis on the same topic. His research involves merging renewable energy with different designs and mathematical models to achieve the best economical solution. His work concluded that PV-based RO plants are most efficient in energy consumption when utilized on a small scale. Ahmad and Schmid [10] study water desalination in Egypt desert using a PV system for energy. Gocht et al. [11] carried out their study for reverse osmosis using photovoltaic energy sources in Jordan. Richards and Schafer [12] worked on PV based desalination plant for remote communities in Australia. Tzen et al. [13], kalogirou [14] also carried out analyses for the same systems. Goosen et al. [15] work were slightly different from others as he studied and compared the challenges in a renewable-based desalination system.

III. REVERSE OSMOSIS DESALINATION SYSTEM

A reverse osmosis setup will be used for water filtration in this design. Reverse osmosis converts unfiltered water to clean water when water is pressurized and passed through a membrane, as shown in Figure 1. When water flows through the membrane, the unfiltered water flows to the more concentrated side of the membrane while clean water moves to the region of low concentration. The unfiltered water can also be referred to as feed water, the clean water as permeate

and the concentrated contaminated water as the waste or brine. A membrane is a tool with filtering pores that allow only clean water molecules to pass through it. In regular osmosis, equilibrium is obtained on both sides of the membrane. Still, in reverse osmosis, constant pressure is applied to the feed water while ensuring that the brine and permeate do not mix up but are collected separately [16].

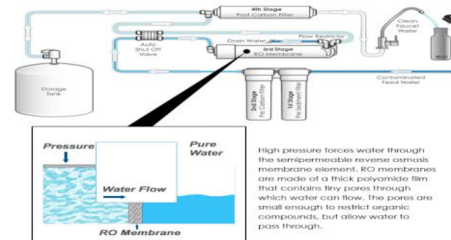


Figure 1 - Reverse Osmosis Plant Flow Diagram [16]

IV. SITE SELECTION AND SYSTEM DESIGN

A. Site Characteristics and Water Supply System:

Black Tickle - domino is in the southern Inuit community on the pond's island; this community is one of the minorities in Canada that still suffers insufficient water supply. This vicinity has no piped water supply causing 126 people of its residents to have little access to water supply or rather must travel 1 km away or more to access this supply. The community has only one portable drinking water supply located 2 km away, and it is costly for residents to access. The average water consumption per household size consisting of 3 people in this community is 394 Liters, which is less than the standard consumption rate of 274 Liters of water per day per person as the average usage in Canada [17]. Herring Cove pond is 2km from the furthest house in Black Tickle and about 1km from the nearest place [18]. The water quality in the pond was tested to check for its drinking safety; further testing [19] found out the water E-coil was absent, but a certain amount of coliform was present and recorded to be satisfactory for drinking. Based on the reports, this site has been selected.

B. Selected Water Source and Water Requirement:

For the Black Tickle community, Herring Cove Pond is the best fit for the source of water to solve the current state of water scarcity and accessibility. The pump selection depends on the bore depth and flow rate. Ponds elevation is 143 meters with the pump placement at 120 meters; 23 meters were left to avoid the accumulation of underground impurities and pump damage from stones or sands. The pump selection depends on the energy demand, total dynamic head, performance efficiency, energy discharge and maintenance. Choosing the correct sizing for the solar system is very important to meet the energy demand. A poorly sized system can affect energy production and economic cost. The average

daily water consumption per person in Canada is 272.55 Liters [17], including drinking and household activities. However, due to the cost and proximity inefficiency, residents of black tickle consume as low as 88.2 Liters, i.e. about 32% of the standard consumption rate per day. This design will provide the standard water requirement per day per person to solve the issue. Hence the total water requirement per day by whole community is 34,341.3 Liters/day based on the population of 126 people. Figure 2 below shows a schematic diagram for the water extraction setup that will be used for resolving the water scarcity at Black Tickle.

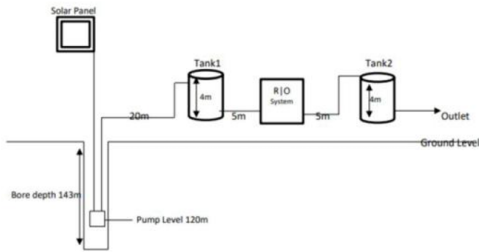


Figure 2 - Block Diagram for Ground Water Extraction

C. Water Energy Calculation:

In order to calculate the total energy required to bring the water from ground to usable level, the total dynamic head needs to be calculated. The total dynamic head is the entire pipe length for the configuration, plus the frictional elbows and horizontal pipes. Assuming that the pump is 120 meter above the water level and has to raise a water to height of two tanks which are 4 meters high each and have 5, 90° elbows will require a total dynamic head of 134 meters. The total head and water consumption requirement per day is required to calculate the pump size. The shaft power required is calculated to be 5.2 KW using the hydraulic power required to raise the water given in equation 1 and then using the shaft power required by the motor as given in equation 2.

$$\text{Hydraulic Power} = \rho * g * h * Q \tag{1}$$

$$\text{Shaft Power} = (\text{Hydraulic Power}) / (n) \tag{2}$$

Where

ρ is the water density of 1000 kg/m³

g is the gravitational constant 9.81 m/s²

h is the total dynamic head i.e. 134 meters

Q is the water flow rate i.e. 0.002361 m³/s

n is the system efficiency i.e. 60%

D. Electrical Load Requirement:

The significant components of load in the system are submersible pump motor, complete RO system load and miscellaneous load. The entire system will run for 04 hours a day to produce sufficient drinking water for the whole community. Based on the total dynamic head of 134 meters, the liquid flow rate of 37.4 gallons per minute, and pump

efficiency of 60%, a 7.0-kilowatt single phase induction motor is selected to be coupled with the submersible pump [21]. The complete RO system for producing 37.4 gallons per minute has a power requirement of 12.0 kilowatt [22]. Some miscellaneous energy required, including area lighting, control equipment, etc., would need a maximum of 1.0 kilowatt. Hence the total energy required will be approximately 20.0 kilowatts for running the system smoothly at a production rate of more than 37.4 gallons per minute. The table 1 summarizes the total electrical energy requirement of the system.

Table 1 - Electrical Load Summary

Component	Load (KW)
Submersible pump motor	7.0
Reverse Osmosis System	12.0
Miscellaneous	1.0
Total Load	20.0

V. SYSTEM STEADY STATE MODELLING AND ANALYSIS

A. System Design in HOMER PRO:

The HOMER Pro software is used for the techno-economic analysis to optimize the system and introduce the most feasible solution. The system design includes a Photovoltaic (PV) array, lead-acid battery and inverter feeding the load. The system has a daily energy requirement of 80-kilowatt hours with a peak load of 35.01 kilowatts. The PV has 32.9 kilowatts of maximum power and uses a 38.0-kilowatt inverter. The result of Homer Pro with optimization suggests 72 12-volt batteries with four batteries in each string to keep the DC bus voltage at 48 volts. The system configuration in HOMER software is shown in figure 3.

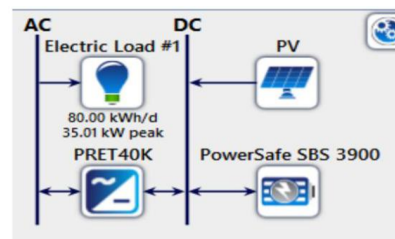


Figure 3 – System Design in HOMER Pro Software

The figure 4 below shows the available solar irradiance and clearness index at the proposed site location.

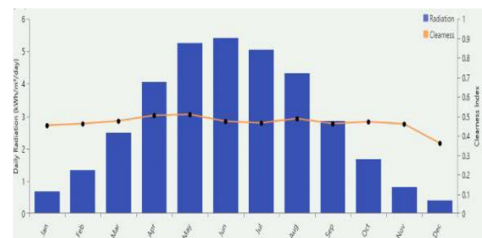


Figure 4 – Solar Irradiance at Herring Cove Pond

B. System Steady State Modelling Results:

The figure 5 shows comparison of total electrical load served, PV Panel Power output and battery input power for three consecutive days in the month of June. It can be seen that when the photovoltaic is sufficient for the load battery is being charged otherwise battery is providing power to the load.

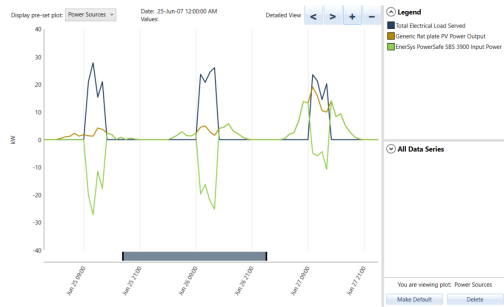


Figure 5 - Power flow comparison on three random days of June

In figure 6, the summary of the optimized electrical data is shown, which shows the electricity production, consumption and the renewable fraction constraints. It can be seen that the system is environmental friendly as all the energy used is from the renewable energy sources.

Production	kWh/yr	%
Generic flat plate PV	35,467	100
Total	35,467	100

Consumption	kWh/yr	%
AC Primary Load	29,179	100
DC Primary Load	0	0
Deferrable Load	0	0
Total	29,179	100

Quantity	kWh/yr	%
Excess Electricity	7,864	22.2
Unmet Electric Load	20.6	0.0704
Capacity Shortage	27.4	0.0939

Quantity	Value	Units
Renewable Fraction	100	%
Max. Renew. Penetration	281	%

Figure 6 - Electrical Summary

Figure 7 shows cash flow analysis for 25 years according to the type of cost, including replacement, salvage, operating and capital. The cost in all systems was selected from the internet and is put in Canadian Dollars for economic analysis.

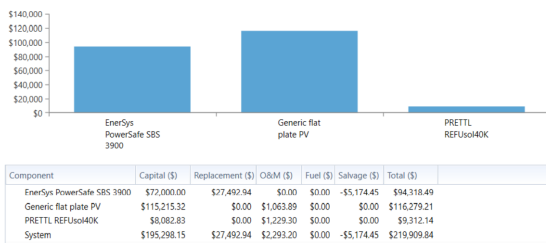


Figure 7 - Cost Analysis

VI. SYSTEM DYNAMIC MODELLING AND ANALYSIS

A. System Modelling in MATLAB Simulink:

MATLAB is a programming software that is used worldwide for developing algorithms and creating models of systems.

Simulink is a sub program of MATLAB which allows the user to create multidisciplinary systems with the power of writing and integrating the code and finding the response of that system. The figure 8 shows the complete electrical model of the system in which photovoltaic source is used followed by boost converter for charging the battery. The battery is connected with the inverter and the motor load is considered as resistor for simplification in model.

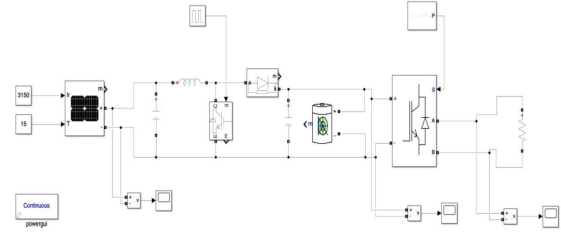


Figure 8 – Complete System Model in Matlab Simulink

B. System Dynamic Modelling Results:

The system dynamic response was analyzed using the Simulink. The figure 9 shows the photovoltaic voltage response and it can be seen that in start the transient behavior of PV increases the voltage to 100 volts whereas after reaching the steady state it stabilizes to actual 65V.

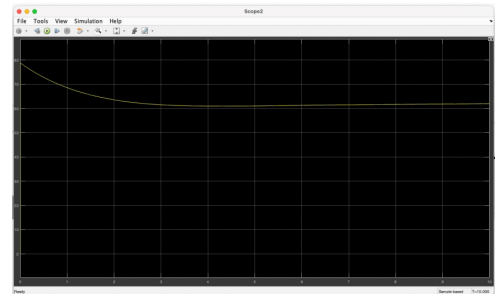


Figure 9 - PV Voltage Response

The boost charging response is shown in figure 10. It can be seen that the voltage has fast switching spikes as duty cycle is used for increasing the voltage level for battery charging.

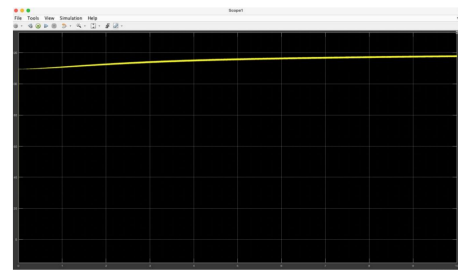


Figure 10 – Battery Boost Charge

Figure 11 shows the response of voltage across the load. It can be seen that the voltage waveform has both positive and negative cycles and DC waveform is converted to AC waveform using the inverter. The waveform is not pure sine

wave as filtering circuit was not implemented across the load for simplification purpose of model.

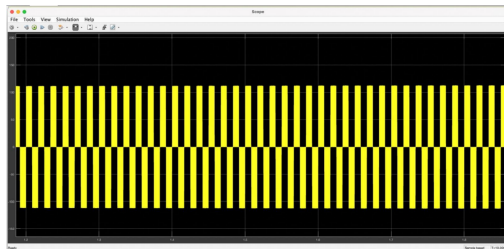


Figure 11 - Load Voltage Response

VII. SYSTEM PROTECTION CONTROL DESIGN

The protection control of the water filtration system aims at increasing the reliability in operation in case of any fault/loss of any individual system/component. The proposed protection system will have valves to control water flow and an over-current relay to protect motors, flow rate sensors, pressure sensors and level switches. The figure 12 depicts the proposed PV-powered water filtration system's protection system process and instrument diagram using several field instrumentation. The table 2 elaborates the instruments used in order to implement the protection control of the system. The complete protection control strategy is elaborated further and 6 major protection schemes are explained.

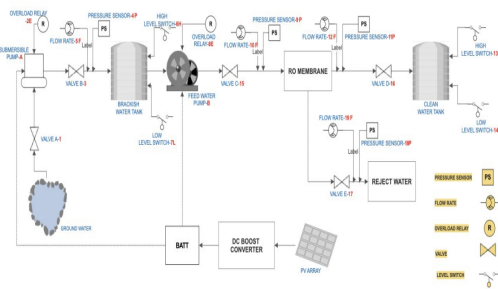


Figure 12 - Protection System Process and Instruments Diagram

Table 2 - Field Instrument List

Identifier	Description
1	Valve-A Open/Close
2E	Overload relay for Pump-A
3	Valve-B Open/Close
4P	Pump-A Pressure sensor
5F	Pump-A Flow rate sensor
6H	Brackish water tank High-level limit switch
7L	Brackish water tank Low-level limit switch
8E	Overload relay for Pump-B
9P	Pump-B Pressure sensor
10F	Pump-B Flow rate sensor
11P	RO Membrane clean water outlet pressure sensor
12F	RO Membrane clean water outlet Flow rate sensor
13H	Clean water tank High-level limit switch
14L	Clean water tank Low-level limit switch

15	Valve-C Open/Close
16	Valve-D Open/Close
17	Valve-E Open/Close
18P	Reject water outlet Pressure sensor
19F	Reject water outlet Flow sensor

A. Pump – A Protection Control

If pump-A has faults like over current, thermal, phase loss and undercurrent the overload relay 2E gets activated and trips the pump-A.

B. Pump – B Protection Control

If the overload relay 8E senses any abnormal conditions in the running parameters, it trips the pump-B.

C. Clean Water Tank Protection Control

Two level switches, 13H and 14L, monitor the clean water tank level for high and low levels. If the clean water tank level reaches the high-level set point and holds for more than 15 seconds, the level switch contact initiates a signal to trip the complete system.

D. Brackish Water Tank Protection Control

Two level switches are also installed on the brackish water tank to monitor the liquid levels in the tank. The high-level switch, 6H, initiates a partial shutdown of the system by first sending a stop signal to pump-A followed by closing the valve-B. In this condition, the liquid inflow to the brackish water stops, but the rest of the system remains in operation. The system normalizes when the 6H level switch resets back to its normal condition.

E. Reverse Osmosis Membrane Protection Control

This protection function considers the differential pressure across the membrane. The inputs are taken from the pressure sensor, 11P at the outlet of the RO membrane and the pressure sensor, 9P at the inlet of the RO membrane. If the differential pressure reaches a defined set point, the emergency shutdown is activated and the whole system is shut down. The filtration system can be restored after checking the healthiness of the RO membrane. This condition indicates the healthiness of the membrane.

F. Reject Water Outlet Protection Control

This function monitors the productivity and efficiency of the system by monitoring the flow rates of 12F and 19F, the flow rate sensors at the RO membrane outlet and the reject water inlet, respectively. If the system's efficiency falls below 50%, an emergency shutdown signal is initiated to shut down the complete system.

V. CONCLUSION

Based on the optimized results from the HOMER Pro Software, the most economical design of a solar-powered based drinking water reverse osmosis system is designed, which provides enough drinking water to the whole

community of the Black Tickle-Domino. The system is based on solar power, which means no carbon dioxide emission in the environment. The results depict that 32.9 KW of maximum power from the solar panels and battery arrangement of 72 batteries of 4 strings will be sufficient for the smooth operation of the water filtration system. The system dynamic response concludes that with minor changes the system will be able to operate in stable zone. Further the detailed protection control scheme of the system ensures that the system operation remains reliable irrespective of any faults that can occur. Considering the advantages of the system and the need of the residents of the community, the initial investment cost is high but justified, and the final design turns out to be the most feasible solution for providing drinking water to the whole community.

REFERENCES

[1] H2 –Whoa: The Water Crisis in Newfoundland; Chad Pelly; The Overcast – Newfoundland Alternative Newspaper; <https://theovercast.ca/h2-whoa-the-water-crisis-in-newfoundland/> (Accessed 16th Jan 2022)

[2] Joshua Barrett and Maura Hanrahan; Case Study for the NL Rural Drinking Water Project; Black Tickle-Domino, Labrador; Memorial University of Newfoundland; <http://nlwater.ruralresilience.ca/wp-content/uploads/2013/04/Black-Tickle-Short-Community-Profile-FINAL.pdf> (Accessed 16th Jan 2022)

[3] Atanu S., Maura H. and Amy Hudson; Water Quality in Aboriginal Communities in Labrador; The Harris Centre – Memorial University; https://www.mun.ca/harriscentre/reports/Sarkar_Water_12_13_Final.pdf (Accessed 16th Jan 2022)

[4] In the World: Small Mexican village produces clean water with solar powered system; Jennifer Chu; MIT News; <https://news.mit.edu/2013/clean-water-solar-powered-system-0911> (Assessed 17th Jan 2022)

[5] Muhammad Wajid Saleem, Asad Abbas, Muhammad Asim, Ghulam Moeen Uddin, Tariq Nawaz Chaudhary and Asad Ullah, Design and cost estimation of solar powered reverse osmosis desalination system, *Advances in Mechanical Engineering* 2021, Vol. 13(6) 1–11 The Author(s) 2021 DOI: 10.1177/16878140211029090

[6] Abdul Ghafoora, Anjum Munir , Tauseef Ahmed , Muhammad Nauman , Waseem Amjad , Azlan Zahid, Investigation of hybrid solar-driven desalination system employing reverse osmosis process, *Desalination and Water Treatment*, 178 (2020) 32–40 February, doi: 10.5004/dwt.2020.24996

[7] M.A. Abdelkareem, M. El Haj Assad, E.T. Sayed, B. Soudan, Recent progress in the use of renewable energy sources to power water desalination plants, *Desalination*, 435 (2018) 97–113

[8] E.S. Ali, A.S. Alsaman, K. Harby, A.A. Askalany, M.R. Diab, S.M.E. Yakoot, Recycling brine water of reverse osmosis desalination employing adsorption desalination: a theoretical simulation, *Desalination*, 408 (2017) 13–24.

[9] C. Charcosset, A review of membrane processes and renewable energies for desalination, *Desalination*, 245 (2009) 214–231.

[10] G.E. Ahmad, J. Schmid, Feasibility study of brackish water desalination in the Egyptian deserts and rural regions using PV systems, *Energy Convers. Manage.*, 43 (2002) 2641–2649.

[11] W. Gocht, A. Sommerfeld, R. Rautenbach, Th. Melin, L. Eilers, A. Neskakis, D. Herold, V. Horstmann, M. Kabariti, A. Muhaidat, Decentralized desalination of brackish water by a directly coupled reverse-osmosis-photovoltaic-system - a pilot plant study in Jordan, *Renewable Energy*, 14 (1998) 287–292

[12] B.S. Richards, A.I. Schäfer, Design considerations for a solarpowered desalination system for remote communities in Australia, *Desalination*, 144 (2002) 193–199.

[13] E. Tzen, K. Perrakis, P. Baltas, Design of a stand-alone PV-desalination system for rural areas, *Desalination*, 119 (1998) 327–334.

[14] S.A. Kalogirou, Effect of fuel cost on the price of desalination water: a case for renewables, *Desalination*, 138 (2001) 137–144.

[15] M.F.A. Goosen, H. Mahmoudi, N. Ghaffour, Today’s and future challenges in applications of renewable energy technologies for desalination, *Crit. Rev. Env. Sci. Technol.*, 44 (2014) 929–999

[16] What is Reverse Osmosis System and how does it work?, *Fresh Water Systems*, <https://www.freshwatersystems.com/blogs/blog/what-is-reverse-osmosis> (Accessed 18th Jan 2022)

[17] M. Hanrahan, “Black Tickle--Domino, Labrador: Case Study for the NL Rural Drinking Water Project A community case--study report for the Exploring Solutions for Sustainable Rural Drinking Water Systems Project.” Accessed: Feb. 20, 2022. [Online].

[18] M. Hanrahan and A. Hudson, “Water insecurity in the Indigenous communities in Labrador View project Perception of Occupational Exposure of Noise and Its Impact on Fish Harvester’s Health in Newfoundland and Labrador’s Fleet: A Mixed-Method Study View project,” Article in *Water Quality Research Journal of Canada*, 2015, doi: 10.2166/wqrjc.2015.010

[19] A. Sarkar, M. Hanrahan and A. Hudson, “Water insecurity in Canadian Indigenous communities: some inconvenient truths,” *Rural and Remote Health* Oct 2015, doi: 10.22605/rrh3354.

[20] “Sun Insolation Hours per Day in Canadian Cities”; <https://www.solar-store.com/Insolation%20Chart.pdf> (Accessed 21st Feb 2022)

[21] “Water Pump”; Tomei Water Solutions; <https://www.tomeiwatersolutions.com/en/home/water-pumps/calpeda/centrifugal/centrifugal-water-pump-calpeda-nm-2-s-a-0-6kw-0-8hp-3-phase-400v-heavy-duty.2.5.145.gp.32976.uw>; (Accessed 21st Feb. 2022)

[22] “Skid mounted reverse osmosis system”; <https://www.environmental-expert.com/products/aquatech-model-40-gpm-skid-mounted-reverse-osmosis-system-193033>; (Accessed 21st Feb. 2022)

Reconfigurable Star-Delta VBR Induction Machine Model for Predicting Soft-Starting Transients

Sheraz Baig

Department of Electrical and
Computer Engineering
University of British Columbia
Vancouver, Canada
sherazbaig@ece.ubc.ca

Taleb Vahabzadeh

Department of Electrical and
Computer Engineering
University of British Columbia
Vancouver, Canada
talebv@ece.ubc.ca

Seyyedmilad Ebrahimi

Department of Electrical and
Computer Engineering
University of British Columbia
Vancouver, Canada
ebrahimi@ece.ubc.ca

Juri Jatskevich

Department of Electrical and
Computer Engineering
University of British Columbia
Vancouver, Canada
jurij@ece.ubc.ca

Abstract—Induction machines are extensively utilized in many commercial and industrial applications. Simulations of such machines to analyze their starting and operational performance require numerically accurate and efficient models. The conventional qd models, although simple to implement, require snubber circuits for interfacing with an external network, which reduces the accuracy and makes the model computationally expensive. This paper extends the prior work and presents a reconfigurable star-delta constant-parameter voltage-behind-reactance (CPVBR) model of a three-phase squirrel-cage induction machine considering the low-frequency deep-rotor-bar phenomenon. The eigenvalue analysis and computer studies demonstrate that the proposed model yields superior computational performance while providing an efficient machine-network interface as compared to the established qd model. It is envisioned that the new model can be useful for efficient simulation of power systems including induction machines with star-delta starters.

Keywords—Constant-parameter voltage-behind-reactance (CP-VBR) model, delta-connection, deep-rotor-bar effect, Induction machine (IM), qd model, star-delta starting method, wye/star-connection.

I. INTRODUCTION

Induction machines (IMs), due to their robustness, reliability, low cost, simple design, and self-starting capability are extensively utilized in many commercial and industrial applications, such as pumps, mills, crushers, conveyor belts, centrifugal machines, drilling machines, etc. [1]–[4]. Depending on performance requirements and applications, various stator windings configurations of induction machines may be used [5], [6], wherein the star-connected and delta-connected stator winding configurations are commonly used.

The direct-on-line (DOL) starting of induction machines has several issues [7], [8], including high starting inrush currents, pulsating torque, large real and reactive power to counteract the moment of inertia at rest, increased stresses on power system equipment, and voltage dips across the system buses. All of these parameters return to their normal values once the steady-state is reached.

Conventionally, several starting methods of induction motor have been proposed in the literature which can include DOL starting, star-delta starting, auto-transformer starting, AC

voltage regulators, variable frequency drives (VFDs), microcontroller-based starters, [9], [10] etc. The choice of each motor starting method generally depends on specific applications, speed and torque requirements, cost, space, and weight restrictions. For example, the star-delta starting method provides an effective and economic solution to improve both starting and operational performance of the induction machine. The advantages of this method include lower starting currents and lower starting torque pulsations. Also, depending on the motor loading and switching time, a star-delta starter with an open transition can cause voltage dips, current spikes, and torque surges. To overcome this limitation, an alternative approach namely a star-delta starter with closed transition can be useful which uses one additional contactor and a few power resistors/reactances to reduce the current and torque spikes during the transient.

Simulation studies utilizing numerically efficient and accurate induction machine models in electromagnetic transient (EMT) simulation programs have become indispensable for the design, analysis, operation, and control of power systems. Traditionally, in EMT simulation programs, induction machines are modeled using $qd0$ equivalent circuit [11]. In the classical qd model, both stator and rotor variables are transformed into qd coordinates, thus resulting in decoupled and rotor position-independent inductances. However, the indirect interface of the qd model with external inductive or switching circuits using resistive or capacitive snubber circuits enforces additional numerical error in the solution and leads to numerical stiffness and simulation inefficiency [12].

To overcome the limitations of qd model, the constant-parameter voltage-behind-reactance (CP-VBR) formulations of the induction machine model; namely VBR-I, II, III models, have been proposed in [13]. The VBR formulations represent the stator interfacing circuit in abc phase coordinates as three-phase controlled voltage sources behind three-phase RL branches and the rotor subsystem in qd coordinates with rotor flux linkages as state-variables. As a result, the VBR models become non-stiff, numerically accurate, and efficient because they enable direct interconnection of the stator interfacing circuit with arbitrary inductive or power-electronic circuits without snubbers, thus making them useful for power system simulation studies. According to [13], the VBR-I and VBR-II models are suitable to represent both wye/star (Y) and delta (Δ) machine stator winding configurations. The VBR-III model has decoupled RL

branches and is suitable to represent the wye (Y) machine stator winding configuration only.

To employ the advantageous properties of the CPVBR models, it can be preferable to switch between the star- and delta-connected stator windings under different conditions of starting and operational performance of the induction motor. Therefore, this paper presents a reconfigurable star-delta CPVBR model considering the low-frequency deep-rotor-bar phenomenon for a three-phase squirrel-cage induction motor. It is shown that the proposed Y-Δ CPVBR model provides superior numerical performance compared to the alternative/conventional $qd0$ model when considering the star-delta starting transients.

II. VBR FORMULATION WITH VARIABLE ROTOR RESISTANCE

This paper assumes a three-phase squirrel-cage induction machine with either wye-connected or delta-connected stator windings. For simplicity, magnetic saturation is not considered. All rotor parameters are referred to the stator side, and motor sign convention is used. The CPVBR formulations of the induction machine proposed in [13], are based on the classical $qd0$ equivalent circuit model in an arbitrary reference frame [11] as shown in Fig. 1. The interested reader can find a more detailed derivation of the CPVBR models in [13].

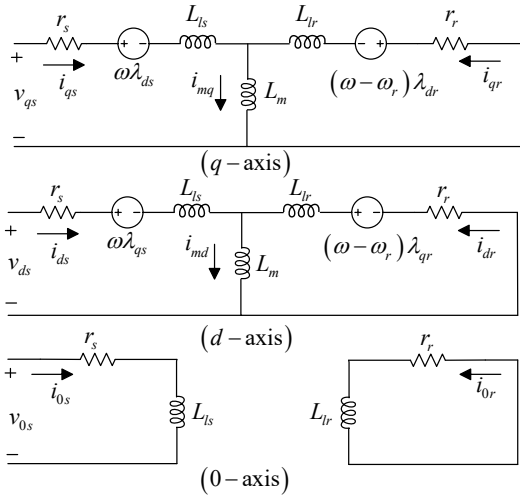


Fig. 1. Classical $qd0$ equivalent circuit in arbitrary reference frame for a three-phase squirrel-cage induction machine.

It is important to note that during the starting electromechanical transients, the rotor resistance may change significantly with the changes in slip and frequency of the rotor currents due to the deep-rotor-bar effect in the squirrel-cage rotor structure. To account for this effect, the single equivalent rotor resistance can be made speed-dependent using a simple linear approximation method proposed in [14]. According to [14], the speed-dependent rotor resistance can be written as

$$r_r(\omega_r) = r_{r1} + s(r_{r2} - r_{r1}), \quad (1)$$

where, r_{r1} and r_{r2} correspond to low-slip ($s \approx 0$) and stand-still ($s = 1$) rotor resistance values, respectively. The slip s is expressed as

$$s = \frac{n_s - n_r}{n_s} = \frac{\omega_s - \omega_r}{\omega_s}. \quad (2)$$

Herein, the VBR-II model [13] is considered as it remains valid for both wye and delta stator windings connections. To better predict the low-frequency deep-rotor-bar phenomenon, the VBR-II model is modified by assuming the speed-dependent variable rotor resistance $r_r(\omega_r)$ in the model equations. Based on the $qd0$ model equivalent circuit of Fig. 1 and considering the low-frequency deep-rotor-bar effect, the rotor flux linkages λ_{qdr} chosen as state-variables, can be expressed as

$$p\lambda_{qr} = -r_r(\omega_r)i_{qr} - (\omega - \omega_r)\lambda_{dr}, \quad (3)$$

$$p\lambda_{dr} = -r_r(\omega_r)i_{dr} + (\omega - \omega_r)\lambda_{qr}. \quad (4)$$

where p is the Heaviside's derivative operator d/dt , $r_r(\omega_r)$ is the variable rotor resistance as a function of rotor speed, ω is the electrical angular frequency of the arbitrary reference frame, and ω_r is the electrical angular frequency of the rotor. The rotor currents can be expressed in terms of rotor flux linkages and magnetizing flux linkages as [11]

$$i_{qr} = \frac{1}{L_{lr}}(\lambda_{qr} - \lambda_{mq}), \quad (5)$$

$$i_{dr} = \frac{1}{L_{lr}}(\lambda_{dr} - \lambda_{md}). \quad (6)$$

where L_{lr} denotes rotor leakage inductance. Replacing the rotor currents from (5)–(6) in (3)–(4), the final rotor state equation becomes

$$p\lambda_{qr} = -\frac{r_r(\omega_r)}{L_{lr}}(\lambda_{qr} - \lambda_{mq}) - (\omega - \omega_r)\lambda_{dr}, \quad (7)$$

$$p\lambda_{dr} = -\frac{r_r(\omega_r)}{L_{lr}}(\lambda_{dr} - \lambda_{md}) + (\omega - \omega_r)\lambda_{qr}. \quad (8)$$

In (7)–(8), the q - and d -axes mutual flux linkages are

$$\lambda_{mq} = L_m'' \left(i_{qs} + \frac{\lambda_{qr}}{L_{lr}} \right), \quad (9)$$

$$\lambda_{md} = L_m'' \left(i_{ds} + \frac{\lambda_{dr}}{L_{lr}} \right). \quad (10)$$

where L_m'' is defined as [13]

$$L_m'' = (L_m^{-1} + L_{lr}^{-1})^{-1}. \quad (11)$$

The stator flux linkages can be written in terms of stator currents and rotor flux linkages as

$$\lambda_{qs} = (L_{ls} + L_m'')i_{qs} + \frac{L_m''}{L_{lr}}\lambda_{qr}, \quad (12)$$

$$\lambda_{ds} = (L_{ls} + L_m'')i_{ds} + \frac{L_m''}{L_{lr}}\lambda_{dr}, \quad (13)$$

$$\lambda_{0s} = L_{ls}i_{0s}. \quad (14)$$

where L_{ls} denotes stator leakage inductance. Substituting the stator flux linkages from (12)–(14) into stator voltages equations in [13] gives,

$$\begin{bmatrix} v_{qs} \\ v_{ds} \\ v_{0s} \end{bmatrix} = r_s \begin{bmatrix} i_{qs} \\ i_{ds} \\ i_{0s} \end{bmatrix} + \mathbf{L}_{qd0}'' p \begin{bmatrix} i_{qs} \\ i_{ds} \\ i_{0s} \end{bmatrix} + \begin{bmatrix} \omega(L_{ls} + L_m'')i_{ds} \\ -\omega(L_{ls} + L_m'')i_{qs} \\ 0 \end{bmatrix} + \begin{bmatrix} e_q'' \\ e_d'' \\ 0 \end{bmatrix}. \quad (15)$$

where \mathbf{L}_{qd0}'' is defined as

$$\mathbf{L}_{qd0}'' = \begin{bmatrix} L_{ls} + L_m'' & 0 & 0 \\ 0 & L_{ls} + L_m'' & 0 \\ 0 & 0 & L_{ls} \end{bmatrix}. \quad (16)$$

In the CPVBR-II model, the sub-transient back-emf voltages in q - and d - axes are calculated as

$$e_{qs}'' = \frac{\omega_r L_m''}{L_{lr}} \lambda_{dr} + \frac{L_m'' r_r (\omega_r)}{L_{lr}^2} \left(\frac{L_m''}{L_{lr}} - 1 \right) \lambda_{qr} + \frac{L_m''^2 r_r (\omega_r)}{L_{lr}^2} i_{qs}, \quad (17)$$

$$e_{ds}'' = \frac{-\omega_r L_m''}{L_{lr}} \lambda_{qr} + \frac{L_m'' r_r (\omega_r)}{L_{lr}^2} \left(\frac{L_m''}{L_{lr}} - 1 \right) \lambda_{dr} + \frac{L_m''^2 r_r (\omega_r)}{L_{lr}^2} i_{ds}, \quad (18)$$

Transformation of (15) into abc phase coordinates by applying the inverse Park transformation matrix $\mathbf{K}_s^{-1}(\theta)$ gives the constant-parameter stator interfacing circuit model as follows

$$\mathbf{v}_{abc} = r_s \mathbf{i}_{abc} + \mathbf{L}_{abc}'' p \mathbf{i}_{abc} + \mathbf{e}_{abc}'', \quad (19)$$

where the transformed sub-transient back-emf voltage sources in abc coordinates are defined as

$$\mathbf{e}_{abc}'' = \mathbf{K}_s^{-1}(\theta) \cdot [e_{qs}'' \quad e_{ds}'' \quad 0]^T. \quad (20)$$

In (19), the constant-parameter inductance matrix \mathbf{L}_{abc}'' of the stator interfacing circuit is calculated as

$$\mathbf{L}_{abc}'' = \mathbf{K}_s^{-1}(\theta) \mathbf{L}_{qd0s}'' \mathbf{K}_s(\theta) = \begin{bmatrix} L_S & L_M & L_M \\ L_M & L_S & L_M \\ L_M & L_M & L_S \end{bmatrix}, \quad (21)$$

In (21), the self and mutual inductances of the stator interfacing circuit are defined as

$$L_S = L_{ls} + \frac{2}{3} L_m'', \quad (22)$$

$$L_M = -\frac{1}{3} L_m''. \quad (23)$$

The electromagnetic torque and electrical rotor speed equations can also be written as [13]

$$T_e = \left(\frac{3P}{4} \right) (\lambda_{md} i_{qs} - \lambda_{mq} i_{ds}), \quad (24)$$

$$p\omega_r = \frac{P}{2J} (T_e - T_m). \quad (25)$$

Finally, the CPVBR-II model [13] constitutes a rotor state-space model (7)–(10), (17)–(18), stator interfacing circuit (19), and mechanical subsystem (24)–(25). The implementation of the CPVBR-II model and its interfacing circuit for both wye- and delta- connected stator windings is shown in Fig. 2, where three-phase controlled voltage sources behind three-phase RL branches representing stator interfacing circuit are used to enable its direct interface with an arbitrary external network without shunt snubbers. The stator currents in CPVBR-II model are selected as inputs to the rotor and mechanical subsystem to calculate the sub-transient back-emf voltages in abc coordinates, as inputs to controlled voltage sources. This model utilizes two three-phase switches “1, 2” to re-configure the machine stator windings connections into either Y or Δ in the run-time. For example, closing of the three-phase switch

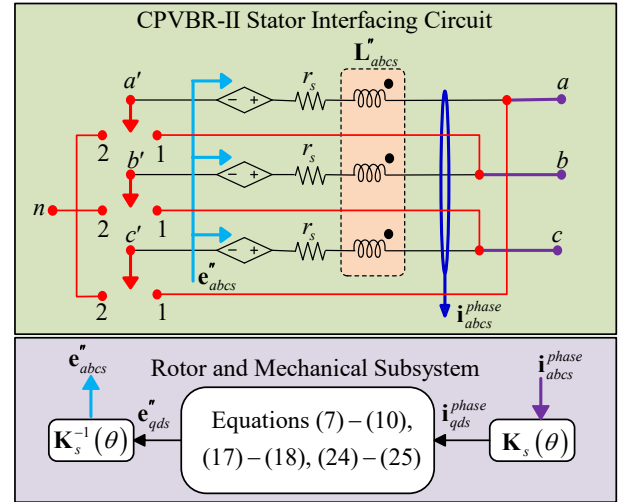


Fig. 2. Implementation of CPVBR-II model of a three-phase squirrel-cage induction machine for reconfigurable wye- and delta-connected stator windings.

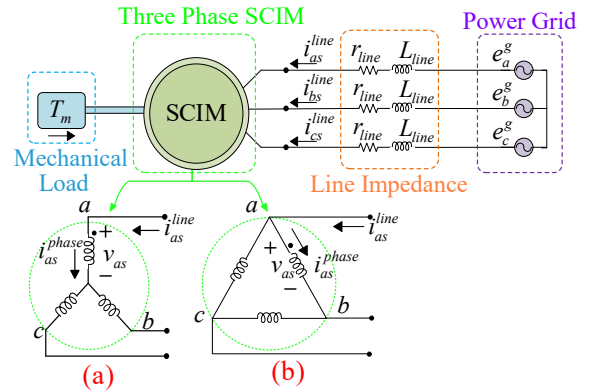


Fig. 3. The test system consisting of a three-phase squirrel-cage induction machine connected to a power grid through power cables: (a) Wye-connected machine, and (b) Delta-connected machine.

“2” connects the stator winding terminals (a' , b' , c') to terminal points (2), consequently resulting in Y-connected IM configuration and vice versa.

III. SIMULATION STUDIES

To validate the simulation results and numerical performance of the considered models, a simple case system depicted in Fig. 3 is considered. Herein, a 460V 50hp squirrel-cage induction motor with either wye- or delta-connected stator windings, is connected to the power grid (represented by ideal voltage sources \mathbf{e}_{abc}^g) through power cables (represented by r_{line} and L_{line}). The system parameters are summarized in Appendix.

For the purpose of accuracy and efficiency comparison, both the classical qd and the proposed CPVBR models considering the deep-rotor-bar phenomenon, have been implemented and simulated in MATLAB-Simulink software. The models were solved using *ode3* fixed time-step solver (with different time-steps), and also *ode45* and *ode23tb* variable time-step solvers (with absolute and relative tolerances set to 10^{-4}). Synchronous reference is used in all subject models. Due to the qd model's indirect interface with external inductive or switching circuits,

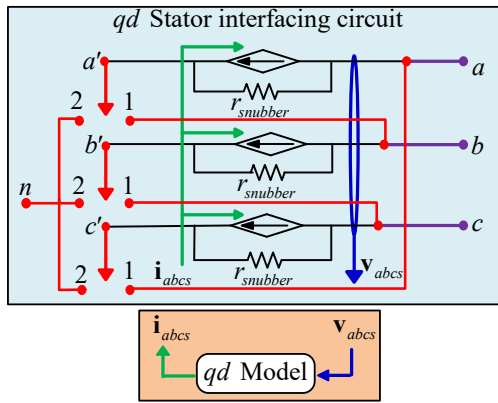


Fig. 4. Implementation of *qd* model of a three-phase squirrel-cage induction machine for wye-connected and delta-connected stator windings.

the shunt resistive snubbers of 982Ω (corresponding to 100pu) are considered to limit the solution error to 1% for the *qd* model. The interfacing circuit of *qd* model for both wye- and delta-connected stator windings is shown in Fig. 4, where three-phase controlled current sources connected in parallel with snubber resistors are used to obtain its compatible interface with external inductive or switching power networks. For consistency, all simulations have been run on a PC with Intel® Core™ i7-10750H @ 3GHz processor.

A. DOL Starting of Delta-connected Motor

In this study, the squirrel-cage induction motor with delta-connected stator windings is assumed to start from zero initial conditions under nominal operating voltages. The configuration of this study is shown in Fig. 3(b). A no-load start-up transient study followed by a load torque change is simulated for 4 s. At $t = 2$ s, the step change in load torque from zero to 198 Nm is applied to the machine, and the simulation is run until 4 s. Fig. 5 presents the transient responses of several variables obtained by the subject models. As it can be observed in Fig. 5, the CPVBR-II model and the *qd* model are consistent, and both accurately predict the system’s transient and steady-state response. Also, for better clarity and comparison purposes, the magnified plots of selected variables from Fig. 5 are depicted in Fig. 6.

In the considered study, the machine’s stator windings are energized from a three-phase voltage source with the phase-phase voltage of 460V, so the rotor of the machine begins to accelerate from a standstill to reach a steady-state speed. As it was observed in Figs. 5–6, the starting inrush current and torque pulsations are significant, which consequently causes voltage dip across the machine terminals. During the start-up, the voltage dip level reaches to about 15% and it recovers approximately at $t = 0.8$ s when the motor reaches steady-state. Then, as a result of a step-change in load torque at $t = 2$ s, the machine establishes a new operating condition as illustrated in Figs. 5–6. Finally, it is deduced from the simulation results that starting squirrel-cage IMs in delta-connected stator windings will give poor starting performance in terms of high starting inrush currents, increased voltage dips, and increased stresses on power utility equipment during their start-up, which may lead to power system instability.

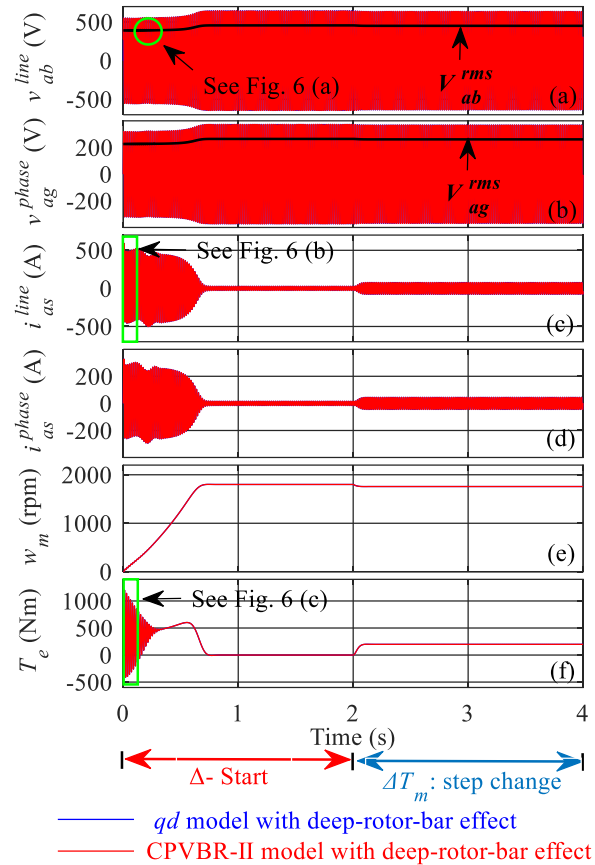


Fig. 5. Transient response of system variables of Δ -connected IM as obtained by the subject models during start-up followed by a load change: (a) $v_{ab, line}$, (b) $v_{ag, phase}$, (c) $i_{as, line}$, (d) $i_{as, phase}$, (e) w_m , and (f) T_e

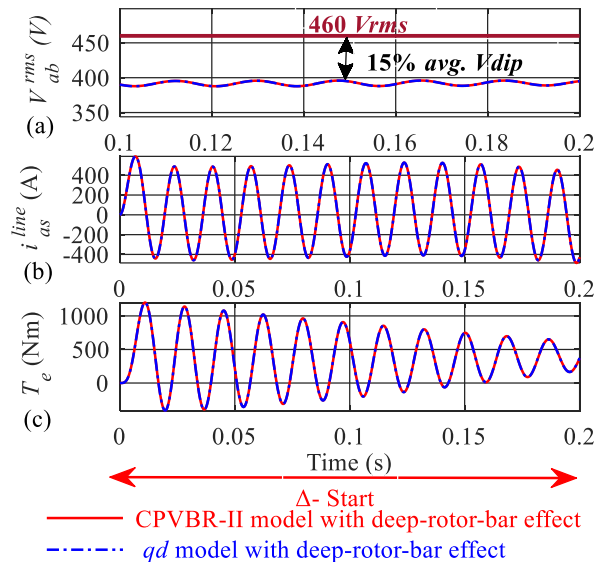


Fig. 6. Magnified plots of selected variables during start-up of Δ -connected IM: (a) RMS line voltage $v_{ab, rms}$, (b) line current $i_{as, line}$, and (c) electromagnetic torque T_e

It is worth mentioning that the new CPVBR-II model with low-frequency deep-rotor-bar-effect has been verified against the *qd* model with the low-frequency rotor-bar-effect. The inclusion of the deep-bar-rotor effect in the subject models

produces more accurate simulation results during starting transient and steady-state periods, as has been demonstrated in [14]. While the models without deep-rotor-bar effects, will consequently result in incorrect estimation of stator-currents and steady-state speed [14].

B. Wye/Delta Starting of SCIM

In this study, the wye-delta starting of the squirrel-cage induction motor is demonstrated. This method allows the motor to operate in wye configuration as shown in Fig. 3(a), at reduced operating voltages (generally by $1/\sqrt{3}$). This leads to reduced power and inrush currents during start-up. Then, after some time, the motor is switched into delta-configuration. A no-load start-up transient study followed by wye-delta switching and load torque step change is simulated for 4 s. At $t = 2$ s, the stator winding configuration is switched from wye- to delta-connection resulting in switching transients. Then, at $t = 3$ s, a step-change in load torque from zero to 198 Nm is applied to the machine and continued to run until 4 s. The transient responses of selected variables obtained by the subject models are demonstrated in Fig. 7. For comparison purposes, the magnified views of selected variables from Fig. 7 are also shown in Fig. 8.

During this study, the machine’s stator phase windings are initially energized from a voltage of $460/\sqrt{3}$ V, due to which

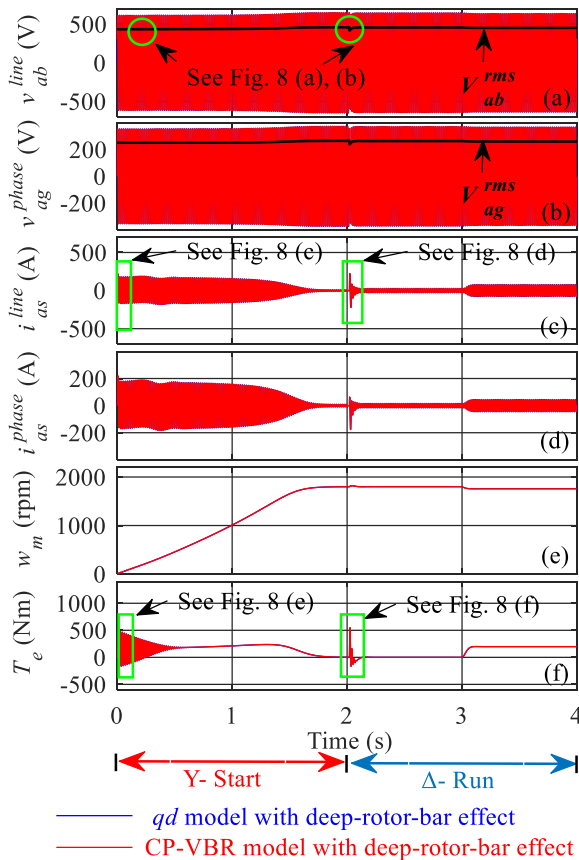


Fig. 7. Transient response of system variables of Y-Δ starter-based IM as obtained by the subject models during start-up followed by Y-Δ switching and load change: (a) $v_{ab, line}$, (b) $v_{ag, phases}$, (c) $i_{as, lines}$, (d) $i_{as, phases}$, (e) w_m , and (f) T_e

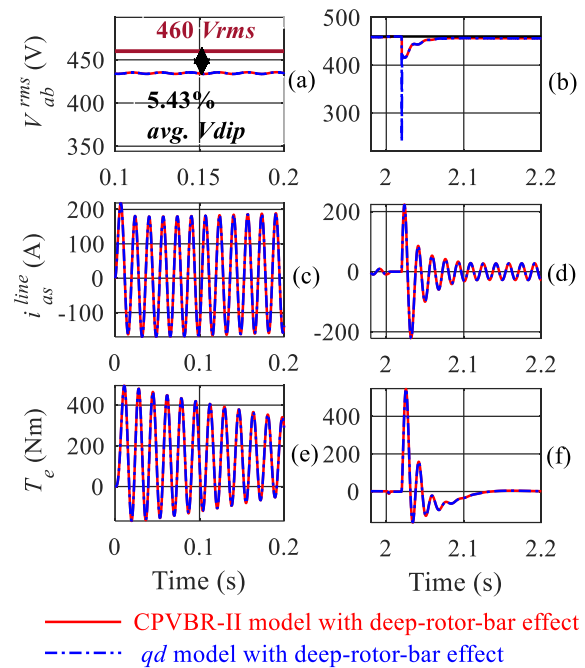


Fig. 8. Magnified views of selected variables under: (1) wye- connected IM starting condition and (2) Y-Δ switching transition

motor starts to accelerate from standstill to steady-state. As observed from Figs. 7–8, the machine draws lower starting inrush currents, produces lower torque pulsations, and has a slower dynamic response as compared to the delta-connected study in Figs. 5–6. This happens due to a reduction in operating voltages across the stator phase winding terminals. Consequently, it improves the motor starting by reducing starting inrush currents flowing through stator windings, and also, reducing the voltage dips across the machine terminals. For our case, the voltage dip is now reduced to about 5.43%, which is in the appropriate range. This voltage dip is cleared after $t = 1.8$ s once the steady-state condition is reached.

After reaching the steady-state operating condition, the stator winding configuration is switched from star to delta causing switching transients in the system which vanishes quickly. It is worth noting that this open transition star-delta switching causes voltage dips, currents spikes, and torque surges, which sometimes may become much worse depending on the switching time and loading condition, consequently leading to failure of power system operation. Therefore, for safe operation, a small switchover transition delay of 20ms is introduced due to which star-connection switches open at $t = 2$ s and delta connection switches close at $t = 2.02$ s resulting in switching transients and voltage dip of 9.78% for CPVBR-II model which is within an acceptable range and cleared in very short time duration. Then, at $t = 3$ s, a load torque of 198 Nm is applied to the machine causing it to establish a new operating condition.

It is also seen in Figs. 7–8, that the proposed CPVBR model and the qd model produced consistent results. However, the qd model produces a deviation in predicting voltage dips due to interfacing error introduced by 1% of current flowing through the snubbers.

C. Computational Performance

Although the considered models produce consistent and matching simulation results, their computational properties are quite different. The computational performance of the subject models simulated using both fixed- (*ode3* solver) and variable-time-step solvers (non-stiff *ode45* and stiff *ode23tb* solvers) are summarized in Tables I, and II. It is observed in Table I, II that CPVBR possesses better-scaled eigenvalues and a lower stiffness ratio as compared to the *qd* model. As seen in Table I from results obtained using the non-stiff *ode45* solver, the CPVBR model is computationally more efficient than the *qd* model in terms of faster simulation speed (3.4 s vs 264.2 s), fewer steps (13177 vs 3122687) and larger average time-step size (258 μs vs 84.6 μs).

Since *qd* model becomes numerically stiff due to shunt snubbers, so it can be observed in Table I that when it is simulated with the stiff *ode23tb* solver, the computational burden is reduced significantly by allowing smaller steps (38418) and CPU time (7.6 s) for the same 4 s transient study. However, the CPVBR model using *ode23tb* solver, still works faster than the *qd* model in terms of much smaller steps (16543 vs 38418) and CPU time (3.5 s vs 7.6 s) due to its lower numerical stiffness.

As observed in Table II from the results obtained using fixed time-step solver *ode3*, the *qd* model is only capable of producing results at smaller time steps. However, it yields a faster simulation speed (90.8s vs 333.9s) when run using fixed time-step solvers at a very small time-step as compared to the CPVBR model. The CPVBR model works slower due to the use of snubber resistances in the switches for both subject models for fixed-time step study. The CPVBR model still has the advantage of running with larger time steps using fixed time-step solvers. This is particularly advantageous for large-scale power system studies where it is desirable to run the simulation with larger time-steps.

Finally, it is deduced that the proposed CPVBR model possesses excellent numerical properties and simulation performance, while accurately predicting the waveforms for several variables. The *qd* model produces a deviation in simulation results due to interfacing error and small currents

TABLE I.

COMPUTATIONAL PERFORMANCE COMPARISONS OF THE SUBJECT MODELS FOR THE 4-SECOND WYE-DELTA STARTING TRANSIENT CASE-STUDY USING NON-STIFF ODE45 AND STIFF ODE23TB VARIABLE TIME-STEP SOLVERS

Solver	Model	Largest eigen value	Steps	CPU time per step (μs)	CPU time (s)
<i>ode45</i>	<i>qd</i> model (snubber)	-3.9×10^6	3122687	84.6	264.2
<i>ode23tb</i>			38418	197.8	7.6
<i>ode45</i>	CPVBR-II model	$-6.4 \pm 377i$	13177	258	3.4
<i>ode23tb</i>			16543	211.6	3.5

TABLE II.

COMPUTATIONAL PERFORMANCE COMPARISONS OF THE SUBJECT MODELS FOR THE 4-SECOND WYE-DELTA STARTING TRANSIENT CASE-STUDY USING ODE3 FIXED TIME-STEP SOLVER

Model	Time step (Δt)	Largest eigenvalue	Steps	CPU time per step (μs)	CPU time (s)
<i>qd</i> model (with snubbers)	1μs	-1.09×10^5	4×10^6	22.7	90.8
	100μs	-2.44×10^3	–	–	cannot run
CPVBR-II model	1μs	$-6.42 \pm 377i$	4×10^6	83.5	333.9
	100μs	$-6.38 \pm 377i$	4×10^4	85.5	3.42

flowing through the snubbers leading to reduced numerical efficiency and accuracy.

IV. CONCLUSION

This paper presented a reconfigurable star-delta starter CPVBR model for a three-phase squirrel-cage induction machine. The new model also considers the low-frequency deep-rotor-bar phenomenon by utilizing linear approximation for a speed-dependent equivalent rotor resistance. The new model accurately predicts the transient and steady-state responses under star-delta switching transitions, and also remains valid for both wye- and delta-connected configurations. The computer studies validate the improved model and demonstrate that it offers superior numerical performance even at larger time-steps as compared to the conventional *qd* model. It is envisioned that this model can be extensively utilized in the many EMT simulators for efficient and accurate simulations of large-scale power systems employing several induction machines with star-delta starters.

APPENDIX

Three-phase Squirrel-cage Y/Δ IM parameters:

Rated power: 50 hp, rated voltage: 460 V, poles: 4,

speed: 1705 rpm, $J = 1.662 \text{ kg}\cdot\text{m}^2$, $r_s = 0.261 \Omega$,

$r_{r1} = 0.342 \Omega$, $r_{r2} = 0.684 \Omega$, $X_{ls} = 0.906 \Omega$,

$X_{lr} = 0.906 \Omega$, $X_m = 39.24 \Omega$.

Power grid: $V_{line} = 460 \text{ V}$, $f = 60 \text{ Hz}$.

Line impedance: $R_{line} = 53.8 \text{ m}\Omega$, $L_{line} = 0.2813 \text{ mH}$.

REFERENCES

- [1] A. T. de Almeida, F. J. T. E. Ferreira, and G. Baoming, "Beyond induction motors—technology trends to move up efficiency," *IEEE Trans. Indus. Applications*, vol. 50, no. 3, pp. 2103–2114, Jun. 2014.
- [2] G. Tanuku and P. Pillay, "Emulation of an induction machine for unbalanced grid faults," *IEEE Trans. Indus. Applications*, vol. 57, no. 5, pp. 4625–4635, Oct. 2021.
- [3] Z. Yin, Q. Hou, Y. Zhang, and Y. Zhang, "A sensorless predictive torque control for induction motor using ultra-local model," in *Proc. 24th Int. Conf. on Electrical Machines and Systems (ICEMS)*, 2021, pp. 136–140.
- [4] K. Kumar, M. Marchesoni, Z. Maule, M. Passalacqua, F. Soso, and L. Vaccaro, "Currents and torque oscillations mitigation in high power induction motor drives," in *Proc. IEEE 15th Int. Conf. Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, 2021, pp. 1–5.

- [5] M. S. Toulabi, L. Wang, L. Bieber, S. Filizadeh, and J. Jatskevich, "A universal high-frequency induction machine model and characterization method for arbitrary stator winding connections," *IEEE Trans. Energy Convers.*, vol. 34, no. 3, pp. 1164–1177, Sep. 2019.
- [6] S. Decker, C. Rollbühler, M. Brodatzki, F. Rehm, A. Liske, and M. Hiller, "Comparison of losses in small star- and delta-connected permanent magnet synchronous machines," in *Proc. 23rd European Conf. Power Electron. and Applications (EPE'21 ECCE Europe)*, 2021, pp.1–10.
- [7] N. S. Behzad and M. Negnevitsky, "Soft and fast starting induction motors using controllable resistive type fault current limiter," in *Proc. IEEE Power Energy Society General Meeting*, 2015, pp. 1–5.
- [8] L. Gumilar, W. S. Nugroho, and M. Sholeh, "Effect of induction motor starting on the power quality of synchronous generator," in *Proc. Int. Conf. Electrical and Information Technology (IEIT)*, 2021, pp. 263–268.
- [9] J. Larabee, B. Pellegrino, and B. Flick, "Induction motor starting methods and issues," in *Proc. Conference Papers Industry Applications Society 52nd Annual Petroleum and Chemical Industry Conference*, 2005, pp. 217–222.
- [10] G. Zigirkas and J. Kalomiros, "An embedded fuzzy controller for the soft-starting of low-voltage induction motors," in *Proc. IEEE 8th Int. Conf. Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2015, pp. 22–27.
- [11] P. Krause, O. Wasynczuk, S. D. Sudhoff, and S. Pekarek, "Symmetrical Induction Machines," in *Analysis of Electric Machinery and Drive Systems*, 2013, pp. 215–270.
- [12] N. Amiri, S. Ebrahimi, and J. Jatskevich, "Efficient simulation of wind generation systems using voltage-behind-reactance model of doubly-fed induction generators and average-value model of switching converters," in *Proc. IEEE First Ukraine Conf. Electrical and Computer Engineering (UKRCON)*, 2017, pp. 605–610.
- [13] L. Wang, J. Jatskevich, and S. D. Pekarek, "Modeling of induction machines using a voltage-behind-reactance formulation," *IEEE Trans. Energy Convers.*, vol. 23, no. 2, pp. 382–392, Jun. 2008.
- [14] S. C. Foroosh, L. Wang, and J. Jatskevich, "A simple induction machine model for predicting low frequency dynamics," in *Proc. Canadian Conf. Electrical and Computer Engineering*, 2008, pp. 1655–1660.

Comparative Analysis of Deep Learning and Machine Learning Techniques for Power System Fault type Classification and Location Prediction

Sivaramarao Bodda
 Dept. of EECS
 Indian Institute of Technology
 Bhilai, India
 sivaramaraob@iitbhilai.ac.in

Anjali Thawait
 Dept. of EECS
 Indian Institute of Technology
 Bhilai, India
 anjalit@iitbhilai.ac.in

Prashant Agnihotri
 Dept. of EECS
 Indian Institute of Technology
 Bhilai, India
 pagnihotri@iitbhilai.ac.in

Abstract—Power system fault type classification and location prediction is critical in assessing the reliability of the power system, and later restoring it to a stable operating point followed by a fault. State of the art methods include sequence component, impedance measurement from the origin of the fault, and traveling wave based methods for fault type detection and classification problem. It is important to identify and classify the fault as quickly as possible for restoring power system stability to normal operation. Machine Learning and Deep Learning methods allow the analysis of large data of fault voltages and currents by using fast and efficient algorithms. These methods require large amount of data, however, with the recent advances in the field of power system, data acquisition using smart meters and Phasor Measurement Units (PMU), huge amount of system-wide data can be made available to analyze the problem of fault type classification and location prediction. This paper presents a comparative study of Stochastic Gradient Descent (SGD) based Deep Neural Network (DNN) and Machine Learning (ML) applied to power system fault type and location prediction problem. DNN architecture uses 10 hidden layers, each layer having 60 units with hyperbolic tangent as activation function, and a combination of Support Vector Machine (SVM) and Principal Component Analysis (PCA) method are considered. Comparative results in terms of time taken to run the algorithms, and accuracy of the results obtained are presented for a 3 machine 9 bus system. Results indicate that SVM method is an optimal choice for fault classification with high accuracy for location prediction and low computational requirements compared with DNN.

Index Terms—Deep Neural Network; Support Vector Machine; Machine Learning; Principal Component Analysis; Hybrid classifier; Kernel; Optimizer; Hidden layer

I. INTRODUCTION

An increase in the deployment of devices such as smart meters, and phasor measurement units in the power system to monitor and collect the system wide data is critical to improve the stability and reliability of the system. The data collected from these devices is generally of high dimension and need to be analyzed. Machine learning methods are good choice for analyzing this large amount of data as they are fast and accurate. Deep Learning (DL) is subset of ML and

they can approximate highly non-linear complex functions by increasing depth of the network. But as the network is more and more deep, computational burden will increase. On the contrary ML methods involve less computational burden compared to DL methods but they may suffer accuracy for highly non-linear problems. To choose an optimal algorithm among ML and DL, a comparative study is carried out for a fault classification problem of a small scale 3 machine 9 bus power system. The most widely used ML algorithm for analysis of such high dimensional dataset is SVM and PCA also known as dimensionality reduction technique. These are closely related and mostly used on a dataset containing high dimensional data matrix and converts it to some useful dataset containing less number of features (Principal Components in PCA), while at the same time preserving the variance in the original dataset. Another method to analyze the high dimensional dataset is to use the DL based techniques. Both ML and DL based techniques have certain advantages and disadvantages. For example, ML based techniques requires lesser amount of data and time to train whereas the DL require higher amount of data and time to train. However, DL based method does not require any feature extraction procedure which is generally required for ML based methods. State of the art research has proposed many methods for fault classification problem [1-6]. An S-transform based Decision Tree (DT)-Fuzzy rule approach is proposed for fault classification in [7]. In this method, S-transform is used for extracting features, DT for primary classification and then based on obtained DT decision boundaries, fuzzy membership function is developed for final classification. Machine learning methods like bagging, boosting, radial basis function and naïve Bayesian are used for power system fault type and location prediction in [8]. As these methods are validated on Single Machine Infinite Bus (SMIB) system, their applicability can't be generalized to multi-machine systems. SVM based fault classification methods are proposed in [9-11]. In [9], authors proposed combination of two SVM algorithms for fault classification. SVM-1 will determine the involvement of particular phase where as SVM-2 will determine the involvement of ground in

the fault. This method is complex as it involve running of two algorithms and also dimensionality reduction technique like PCA is not used. In [10-11], authors proposed the combination of wavelet transform and SVM where in the former method uses fault current features and the later method uses phase angles as input features. PMU measurements based fault detection method considering the effect of communication network delays on fault detection time is proposed in [12]. Symmetrical component technique along with PCA is used in [13] which uses quarter cycle of phase current for fault detection. With the increase of computational power and availability of large data, application of DL methods in power system gained importance and a DL based power system fault classification and location prediction is proposed in [14-16].

This paper compares the performance of both the ML (combination of SVM and PCA based technique) and DL techniques to address a very important problem of fault classification and location detection in power system network. A benchmark 3 machine 9 bus test system is designed and data is collected which mainly consists of the three phase voltages and currents for various types of fault and their locations. Both the DL and ML methods are then trained on this dataset. Later, a comparative analysis is presented to show the performance of both these methods to identify the fault type and location using precision, recall, F1-score, and time taken. The rest of the paper is divided in the following sections. Section II describes the SVM and PCA methods for Power System (PS) fault classification and location prediction; section III describes the stochastic gradient descent based Deep Learning method. Section IV describes the simulation results of 3 machine 9 bus test system and the procedure for data collection. It further describes and compares the results obtained from both the ML and DL techniques on fault type classification and fault location identification followed by conclusion in Section V.

II. SVM AND PCA BASED METHODS FOR PS FAULT CLASSIFICATION AND LOCATION PREDICTION

One of the most influential approaches to supervised learning is SVM, since it provides decision boundaries with large margins; it tends to have lower generalization errors and is less prone to over fitting. Another advantage of SVM is, we need only a small number of samples. It can easily be kernelized (Linear, RBF, Poly, Gaussian, etc.) to solve non-linear classification problems which allows us to learn models using convex optimization techniques that are guaranteed to converge efficiently. Although the PCA and SVM are commonly used in pattern recognition, an effective methodology using both the PCA and SVM for fault classification remains unexplored.

This study proposes a novel framework combining both of these algorithms and coming up with a classifier system which is much more versatile, reliable and is effective in fault classification. It employs a feature-based scheme that integrates PCA with an SVM for effectively detecting patterns in the system. The detailed explanation of the algorithm as well as the implementation technique is explained in the subsequent

sections. Support vectors is one of the most widely used clustering algorithms for unsupervised or unlabeled data. Since the dataset obtained by modeling three machines and nine bus system is highly nonlinear, the Radial Basis Function (RBF) kernel is used for representing the decision boundary in hyper plane for nonlinear data matrix. Taking accuracy into account, RBF kernel gives better performance rate compared to linear or polynomial kernels. RBF classifier is a class of functions which can basically be implemented with linear as well as non linear models with any kind of network considering single layered or multilayered. It is classified as non linear network where the basis function can be changed or adjusted, basically used for linearly non-separable features. An RBF function makes it easier to create the complex decision boundaries having high dimension (even infinite dimensional) features. RBF kernel looks similar to Gaussian kernel except for $\frac{1}{2\sigma^2}$ replaced by γ . For a non-linear feature matrix RBF kernel considering two data points $x^{(i)}, x^{(j)}$ can be represented as

$$\phi : x \rightarrow \varphi(x) \quad (1)$$

$$K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi = \exp(-\gamma \|x^{(i)} - x^{(j)}\|^2), \gamma > 0 \quad (2)$$

Where $x^{(i)}, x^{(j)}$ are the two data points

$\|x^{(i)} - x^{(j)}\|$ represents the Euclidean distance between $x^{(i)}, x^{(j)}$

equation(1) is mapping into higher dimensions.

γ , gamma represents the spread of the kernel function.

Since many functions are available in the Scikit-learn library, the kernel functions need not to be estimated manually. The value of $\gamma (> 0)$, represents the spread of the kernel function and hence the decision boundary. Having lower values for γ may result in under fitting while keeping higher values can cause over fitting. C (Regularization parameter) is defined as a parameter for SVM learner which trades off correct classification of training examples against maximization of the decision function's margin, having small values of C depicts high bias and low variance (Under fitting) while the larger values represents low bias and high variance (Over fitting). Thus, an optimal value of γ and C is determined using grid search algorithm which allows us to select from a proper range of values for each hyper parameter to increase the solution in hyper plane. Using a hybrid combination of both of these algorithms can therefore improve the overall performance of the discussed fault classifier algorithm. The number of features to be extracted as PCs can be determined by various methods. The proposed scheme evaluates the cumulative variance so as to attain nearly 80% of the variability in the data to determine the components. The procedure involved in PCA includes the estimation of covariance matrix $X^T X$ where, X is the $m \times n$ data matrix. The covariance matrix can be calculated as

$$Cov(X) = \frac{1}{n} (X - \bar{X})(X - \bar{X})^T \quad (3)$$

Where, \bar{X} is the mean of n dimensional dataset. It also includes the calculation of Eigen vectors and Eigen values representing

the direction and magnitude of the features in the multidimensional plot. The magnitude therefore represents the importance of respective vectors in the final estimation of classifier algorithm. The system is trained with the obtained labelled dataset to recognize the fault patterns. The complete schematic of the proposed algorithm is given in the Fig.1. Scikit learn is used, importing all the useful libraries, since it provides built-in library functions for the defined algorithms. Instead of simply using test and train dataset, 4-fold cross validation

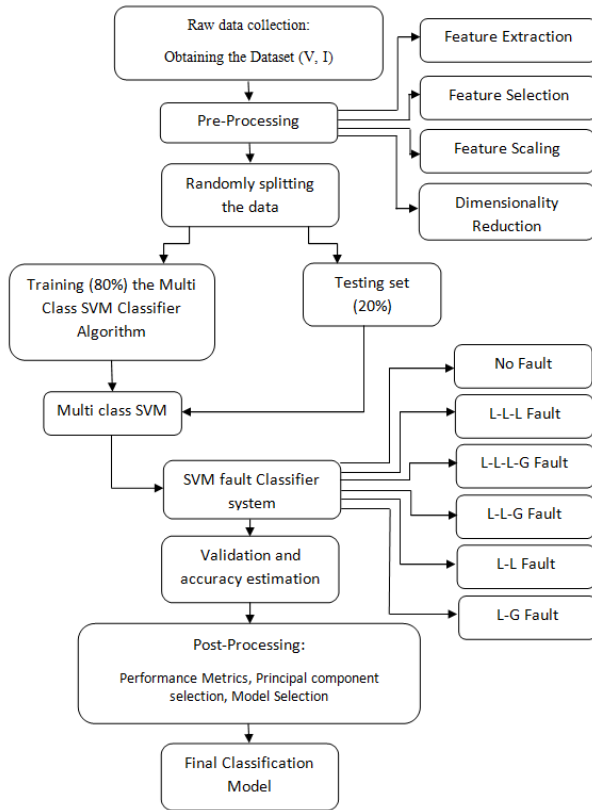


Fig. 1. Proposed Hybrid ML classifier algorithm.

is used resulting in better training and classification of the unforeseen dataset which therefore increases the effectiveness of the classifier algorithm with limited data. The basic aim of k-fold cross validation is to randomly divide the complete data into k different non overlapping subsets where each of these is used to train and rest is used to validate the classifier system and simultaneously recording the accuracy scores.

III. SGD BASED DL METHOD FOR PS FAULT CLASSIFICATION AND LOCATION

DL system performance can be enhanced mostly by three approaches. Firstly, by improving the structure of the model. Secondly, by improving the model initialization, and thirdly by trying more powerful learning algorithm. This paper chose the third option to enhance the model performance. Gradient descent methods are core of deep neural network learning algorithms. These are mainly two types;

- Batch gradient methods

- Stochastic gradient methods.

Batch gradient descent methods are very slow and require more memory as we need to calculate the gradients of whole dataset to perform just one update towards the optimal solution. SGD based optimizers are of core practical importance in the field of DL. SGD proved itself as an efficient and effective optimizer in many machine learning techniques and recent advances in deep learning [17-18].

A. Root Mean Square Propagation (RMS Prop)

‘RMS Prop’ is an adaptive learning rate method proposed by Geoff Hinton. It uses the sign of the gradient and divides the learning rate(α) by an exponentially decaying average of squared gradients. Following are the update rules of weight (W) and bias (b) parameters in ‘RMS Prop’ algorithm;

$$S_{dw}^t = \beta S_{dw}^{t-1} + (1 - \beta)dW^2 \quad (4)$$

$$S_{db}^t = \beta S_{db}^{t-1} + (1 - \beta)db^2 \quad (5)$$

$$W^t = W^{t-1} - \alpha \frac{dW}{\sqrt{S_{dw}^t}} \quad (6)$$

$$b^t = b^{t-1} - \alpha \frac{db}{\sqrt{S_{db}^t}} \quad (7)$$

Where $S_{dw}^t, S_{db}^t, S_{dw}^{t-1}, S_{db}^{t-1}$ are average of squared gradients of weight and bias terms respectively at iteration t and (t-1). dW, db , are change in weight and bias terms and α, β are hyper parameters. The default values of α, β are 0.001 and 0.9 respectively. $W^t, b^t, W^{t-1}, b^{t-1}$ are updated weight and bias terms at iteration t and (t-1) respectively

IV. EXPERIMENTAL RESULTS

A. Test System & Data Collection

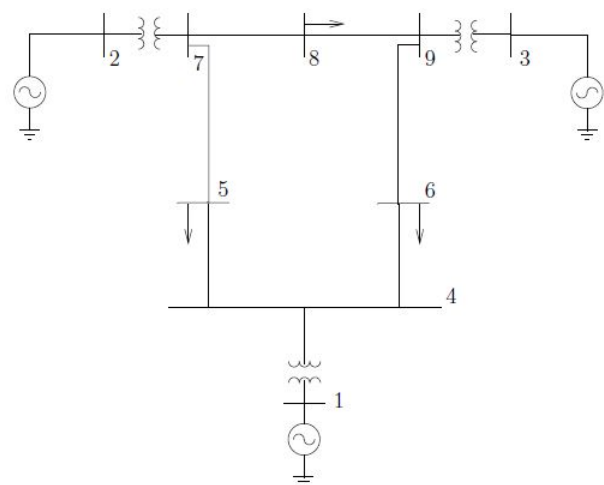


Fig. 2. IEEE 9 Bus system simulated in MATLAB-SIMULINK.

IEEE 3 machine 9 bus system is taken as test system for the application of proposed method. MATLAB & SIMULINK

is used to simulate the test system shown in Fig.2 and the respective data is collected. Each fault case is simulated using a 3 phase fault block in SIMULINK library and all three phase voltages and currents are recorded at both from-Bus and to-Bus. A sample size of 100 points is collected for each simulation experiment and the data is accumulated for all fault case experiments. The collected data is then processed with python scripts using PANDAS library. Transmission line parameters used for the simulation experiment are shown in the Table I. There are 3 generators, each having terminal voltage of 17.16kV,18.45kV, and 14.145kV respectively at 50Hz. Transmission voltage level is 230kV which is also taken as base voltage and 100MVA is taken as base MVA for per unit conversion of parameters. All fault cases are mainly classified in the following four categories:

- L-G Fault
- L-L Fault
- L-L-G Fault
- L-L-L Fault

TABLE I
TRANSMISSION LINE PARAMETERS OF IEEE 9 BUS SYSTEM

Line		R(p.u/m)	X(p.u/m)	B(p.u/m)
From Bus	To bus			
4	5	0.0100	0.0680	0.1760
4	6	0.0170	0.0920	0.1580
5	7	0.0320	0.1610	0.3060
6	9	0.0390	0.1738	0.3580
7	8	0.0085	0.0576	0.1490
8	9	0.0119	0.1008	0.2090

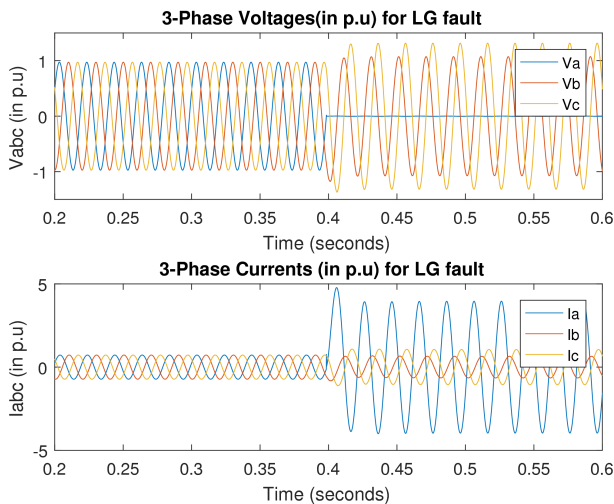


Fig. 3. Line to Ground (LG) fault simulated at Bus 6.

For location prediction problem, the transmission line 4-6 is segmented in 10 equal parts. Various fault experiments are carried out with different fault resistances $R_f = 0.1, 0.1, 1, 10, 100$ ohms. Fig.3 illustrates the voltage and current waveforms when a single phase (L-G) fault is applied at bus 6. A window of 1 cycle data starting from the fault inception time is taken

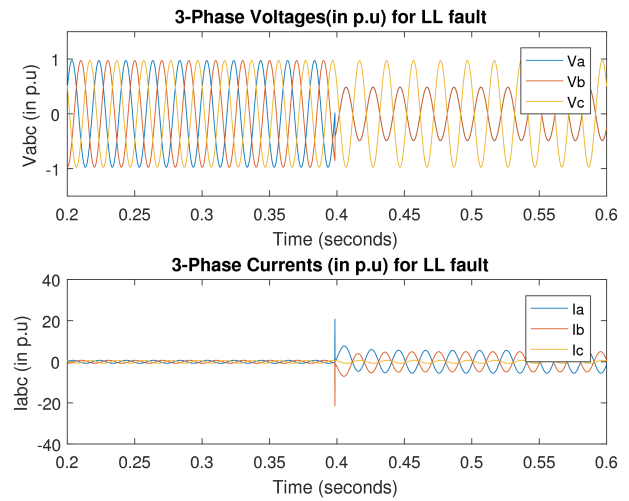


Fig. 4. Line to Line (LL) fault simulated at Bus 6.

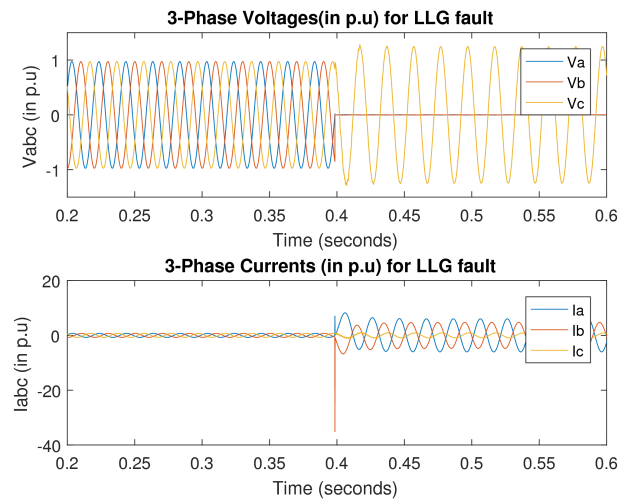


Fig. 5. Double Line to Ground (LLG) fault simulated at Bus 6.

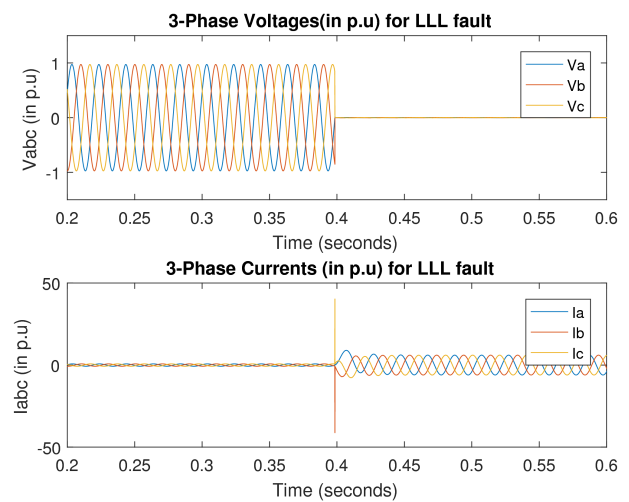


Fig. 6. Triple Line (LLL) fault simulated at Bus 6.

for preparing the dataset. Similarly Fig.4 represents voltage and current waveforms during line-line fault at bus 6. Fig.5 illustrates the current and voltage waveforms when line-line-ground (LL-G) fault is applied. Fig.6 illustrate the current and voltage waveforms during triple line (L-L-L) fault. From Fig.6, it can be observed that even though the triple line fault occurrence is very rare, it impacts the system severely.

B. Training of ML and DL models

The data collected from MATLAB-Simulink is processed and labeled using python script. The processed dataset is then shuffled to avoid biasing problem. Total dataset is split in two parts, 70 % for training and 30 % for testing. Training and testing of model are performed on a CPU with i5 processor, 8GB RAM. Jupyter Notebook is used as python script editor and it also includes Google’s TensorFlow library for working on DNN, ML algorithms.

C. Fault Type Prediction–SVM and PCA

The proposed hybrid algorithm uses PCA as a metric for ranking features in the data space based on the variance prior to classifying faults for each of the ten class pairs by analyzing the Eigen values and Eigen vector distribution. SVM is a robust technique in full feature dimension and is here incorporated with PCA to utilize the efficiency of both the techniques. The best parameters for C and gamma for SVM with radial basis function kernel is obtained using Sklearn and is applied by fixing the number of principal components. The parameters are chosen so as to avoid over fitting and

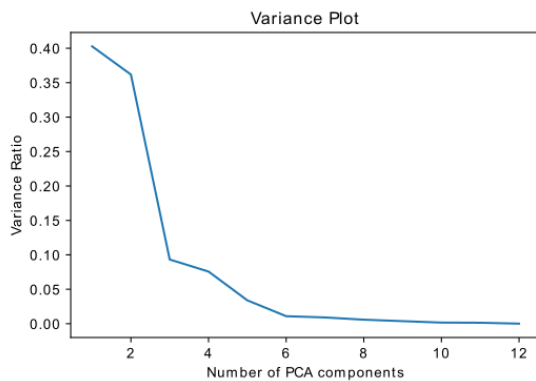


Fig. 7. Variance plot- Fault Type(ML).

less prediction errors. Also, the process is repeated with varying components. Each of these calculations is done on 4-fold cross-validation with the specified classifier system. Fig.7 shows the variance plot with respect to the varying Principal Components (PCs) for determining the fault type in electrical system. It can be observed that there is no significant variation as we move from 6 to 7 PCs. Therefore, only 6 PCs is enough to retain the variance of the complete dataset. The obtained confusion matrix as well as the precision, recall, F1-score and support is given is Fig.8 and Table II. For example, the asymmetric fault type from line C to line A, it is predicted as

high as 62 times correctly while for 6 times it is predicted as CAG fault. And therefore the recall is 1.0 while the precision came out to be 0.8. The F1 score became 0.9.

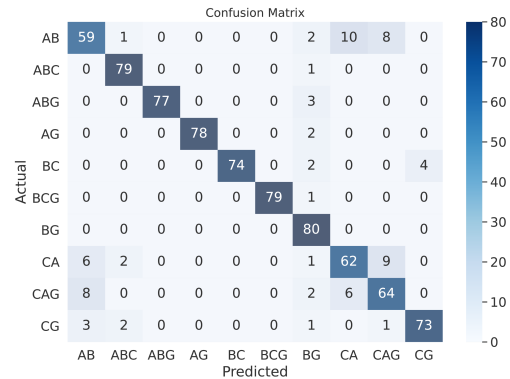


Fig. 8. Confusion Matrix- Fault Type (SVM with PCA).

TABLE II
CLASSIFICATION SCORE-FAULT TYPE(ML)

	Precision	Recall	F1-Score
AB	0.8	0.69	0.74
ABG	1	0.91	0.95
BC	1	1	1
BCG	1	0.97	0.99
CA	0.8	1	0.89
CAG	0.97	0.96	0.97
AG	0.99	0.95	0.97
BG	0.76	0.78	0.77
CG	0.72	0.79	0.75
ABC	1	0.93	0.96
Average	0.9	0.9	0.9

The accuracy rate with varying PCs can be observed from the Table III. The efficiency of the hybrid algorithm in predicting the type of faults with only 5 PC is nearly 89%, which gives a sense that 5 of the components account for all the multivariate variability. The first component accounts for 40%variability while the second component accounts for 36%variability of the overall data space. The average time for processing the hybrid algorithm to predict fault type is 0.95 seconds, which is obtained after sampling the voltages and currents from the corresponding bus within two cycles for each of the fault resistance values.

D. Fault Location prediction –SVM and PCA

The complete electrical system is divided into ten equal parts based on location of faults and is classified into various classes which are used as the dataset to train, validate and test the system using hybrid classifier algorithm. The features are extracted using PCA and the classification is done for different fault case to train and test these PCs over SVM based classifier. The accuracy is computed considering different number of components based on the PCA and simultaneously applying SVM. Fig.9 shows the variance plot with respect

TABLE III
TYPE PREDICTION REPORT USING SVM AND PCA

S.No	PCs	Variance ratio	Accuracy(%)
1	n=1	4.027e-01	21.125
2	n=2	3.617e-01	42.625
3	n=3	9.318e-02	73.840
4	n=4	7.510e-02	85.843
5	n=5	3.389e-02	89.125
6	n=6	1.089e-02	89.562
7	n=7	9.093e-03	89.750
8	n=8	5.847e-03	90.625
9	n=9	3.725e-03	90.843
10	n=10	1.638e-03	90.875
11	n=11	1.392e-03	90.750
12	n=12	1.098e-04	90.750

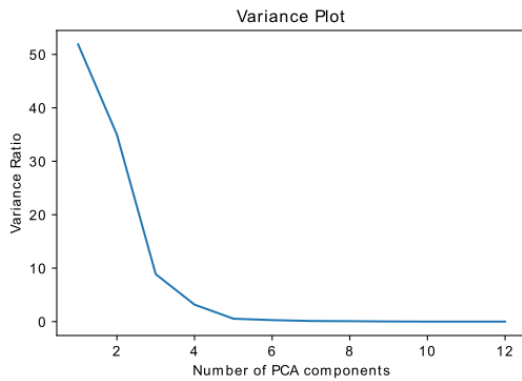


Fig. 9. Variance plot- Fault Location.(SVM and PCA).

to the varying principal components for determining the fault location. It is observed that the variance decreases as we go on moving from one principal component to the last one depending upon the importance of the given feature. The

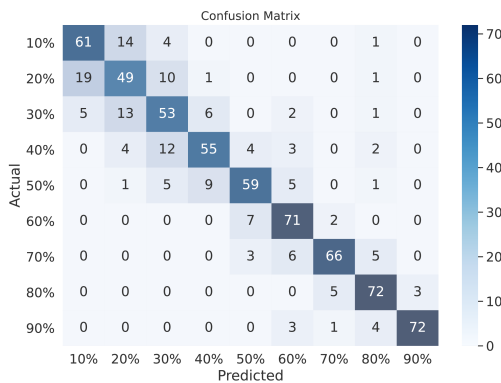


Fig. 10. Confusion Matrix- Fault Location (SVM with PCA)..

variance and proportion of variance contributed by each of the given feature is used to determine the number of PCs to be retaining the properties of the dataset. The principal components are selected based on the significant cumulative variance proportion. It can be clearly seen that the variance

nearly saturates with 5 to 6 components. And hence, only 6 PCs are sufficient to analyze the fault location with minimum variation in prediction efficiency. Henceforth, from the Table IV, the accuracy attained with six principal components is seen to be nearly 76%. Comparably, considering 9 PCs would enable the accuracy as high as 85%. The variance ratios and the accuracy rate are tested with varying number of PCs to analyze the versatility of the classifier system. The proposed framework, which employs a feature-based PCA algorithm and incorporates SVM classification, provides the accuracy rate of nearly 86%. Also, predicting the location of fault using feature based scheme with SVM takes about 1.4 seconds. Table.V depicts the accuracy attained using machine learning classifier algorithm. Fig.10 shows the confusion matrix of classifier for location prediction task.

TABLE IV
LOCATION PREDICTION REPORT USING ML

S.No	PCs	Variance ratio	Accuracy(%)
1	n=1	5.187e-01	15.312
2	n=2	3.498e-01	30.381
3	n=3	8.878e-02	53.854
4	n=4	3.176e-03	69.409
5	n=5	5.503e-03	72.847
6	n=6	2.866e-03	76.493
7	n=7	1.220e-03	79.618
8	n=8	8.926e-04	83.506
9	n=9	3.091e-05	85.555
10	n=10	4.697e-05	85.552
11	n=11	1.508e-10	85.520
12	n=12	2.523e-11	85.552

TABLE V
TESTING ACCURACY OF TWO CLASSIFIERS

	Type Classification	Location Classification
Accuracy	90.0	86.0

E. Fault Type prediction –DNN

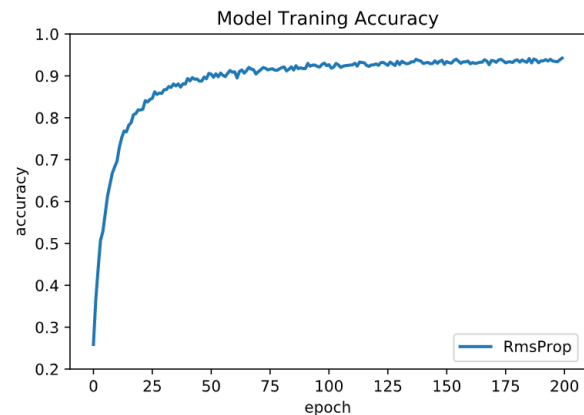


Fig. 11. Training accuracy of RMS Prop DL model-Fault type.

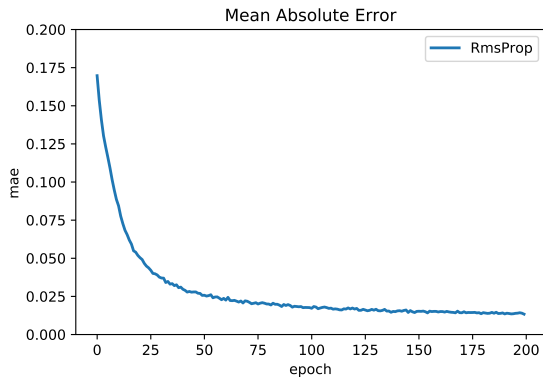


Fig. 12. MAE of RMS Prop DL model-Fault type.

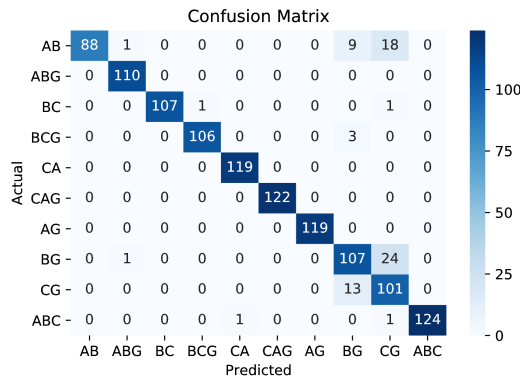


Fig. 13. Confusion matrix- Fault type(DL).

Fault types are mainly classified into 10 types and are encoded as binary numbers to classify using softmax classifier. The DNN model is trained for 200 iterations on training dataset and then tested with test dataset. Fig.11 shows the model training accuracies of RMSProp DL learning algorithm against number of iterations. Mean Absolute Error (MAE) is taken as the performance metric for the model. From Fig.12 it is observed that after 175 epochs, the MAE is not considerably reducing therefore the number of epochs is limited to 200. Fig.13 shows the confusion matrix of DL based classifier for fault type prediction task. It can be seen that for 116 test points, 88 times the double line fault between A & B phases is correctly predicted but for 18 times, double line fault between A and B phases is incorrectly predicted as line to ground fault between C phase and ground. Therefore, in Table VI, the recal corresponding to AB fault suffers.

F. Fault Location prediction – DNN

Fault locations are categorized in 9 parts starting from 10% of the length of transmission line to 90 % of the line. The DNN model is trained for 500 iterations on training dataset and then tested with the test dataset. Fig.14 , Fig.15 shows the accuracy and MAE plots for fault location model. Fig.16 shows the confusion matrix for fault location prediction. It

TABLE VI
TYPE CLASSIFICATION REPORT -DL

Class	Precision	Recal	F1-Score
AB	1.00	0.76	0.86
ABG	0.98	1.0	0.99
BC	1.0	0.98	0.99
BCG	0.99	0.97	0.98
CA	0.99	1.0	1.0
CAG	1.0	1.0	1.0
AG	1.0	1.0	1.0
BG	0.81	0.81	0.81
CG	0.70	0.89	0.78
ABC	1.0	0.98	0.99
Average	0.95	0.94	0.94

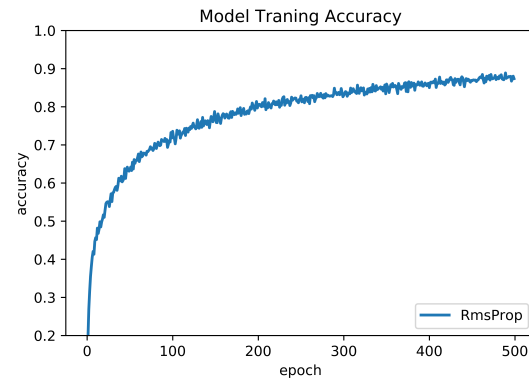


Fig. 14. Training accuracy of RMS Prop DL model-Fault location.

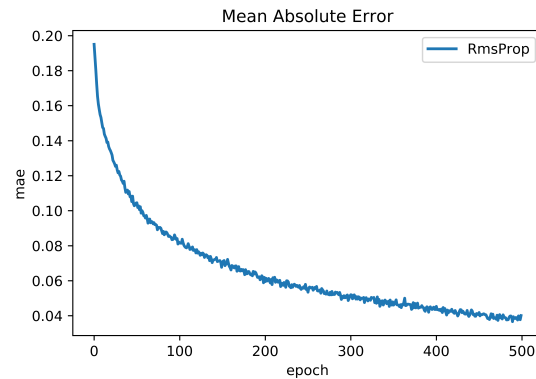


Fig. 15. MAE of RMS Prop DL model-Fault location.

can be seen that for 85 test points, fault at 0.4 times the length of the line is correctly predicted but for 22 times, it is incorrectly predicted as fault at 0.3 times the length of the transmission line. Therefore in Table VII, the precision corresponding to 40% fault suffers. Similar conclusion can be drawn for the 20% fault location as well. But for the remaining fault locations the precision is close to 0.8 and 0.9. From Table VIII, it is observed that for fault type classification problem the model accuracy is 94 % with mean absolute error 0.0118 whereas the location prediction model accuracy is 82 % with mean absolute error 0.0483.

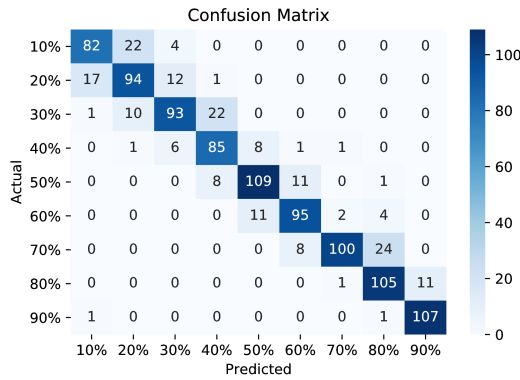


Fig. 16. Confusion matrix- Fault location(DL).

TABLE VII
LOCATION PREDICTION REPORT-DL

S.No	Precision	Recall	F1-Score
10%	0.81	0.76	0.79
20%	0.74	0.76	0.75
30%	0.81	0.74	0.77
40%	0.73	0.83	0.78
50%	0.85	0.84	0.85
60%	0.83	0.85	0.84
70%	0.96	0.76	0.85
80%	0.78	0.90	0.83
90%	0.91	0.98	0.94
Average	0.83	0.82	0.82

TABLE VIII
TESTING ACCURACY & MAE OF TWO CLASSIFIERS- DNN

	Type Classification	Location Classification
Accuracy	93.79	82.15
MAE	0.0118	0.0483

G. Comparison of ML and DNN for PS Fault Type & Fault Location prediction

From Table IX, it is observed that fault type prediction accuracy is slightly higher for DL method when compared to that of with ML method whereas the location prediction accuracy is high for ML method when compared to that of with DL method. However, for the given test system, the model training times are significantly low for ML methods compared to the DL method, therefore they can easily be deployed onto real time fault detecting circuits.

TABLE IX
COMPARISON OF ML & DL METHODS

	DL Method RMSProp (Accuracy in %)		ML Method SVM with PCA(Accuracy in %)	
	Type Prediction	Location Prediction	Type Prediction	Location Prediction
	94	82	90	86
Model Training Time	105 sec (200 epoch)	230 sec (500 epoch)	0.95 sec	1.4 sec

V. CONCLUSION

Analysis and results demonstrated that machine learning method dominates with high accuracy of location prediction and low model training time. As ML method has low execution time, it can be easily deployed on to an embedded target for real time fault detection. Also location prediction is more critical in fault-clearing process when compared to the type prediction. Therefore in small scale electrical system, the proposed ML method can predict the fault with a much higher accuracy rate and provides an efficient fault selection technique which is proved to be faster, reliable and is more efficient than the traditional selector modules.

ACKNOWLEDGMENT

The authors are thankful to IIT Bhilai for providing research facility in the campus.

REFERENCES

- [1] Lai, T., Snider, L., Lo, E., et al.: 'High-impedance fault detection using discrete wavelet transform and frequency range and RMS conversion', IEEE Trans. Power Deliv., 2005, 20, pp. 397-407.
- [2] Das, B., Reddy, J.V.: 'Fuzzy-logic-based fault classification scheme for digital distance protection', IEEE Trans. Power Deliv., 2005, 20, pp. 609-616.
- [3] Kezunovic, M.: 'Smart fault location for smart grids', IEEE Trans. Smart Grid, 2011, 2, pp. 11-22
- [4] Shaik, A.G., Pulipaka, R.R.V.: 'A new wavelet based fault detection, classification and location in transmission lines', Int. J. Electr. Power Energy Syst., 2015, 64, pp. 35-40
- [5] S. R. Samantaray, P. K. Dash, and G. Panda, "Fault classification and location using HS-transform and radial basis function neural network," Electr. Power Syst. Res., vol. 76, no. 9-10, pp. 897-905, 2006.
- [6] K. Chen, C. Huang and J. He, "Fault detection, classification and location for transmission lines and distribution systems: a review on the methods," in High Voltage, vol. 1, no. 1, pp. 25-33, 4 2016.
- [7] S. R. Samantaray, "A systematic fuzzy rule based approach for fault classification in transmission lines," Appl. Soft Comput. J., vol. 13, no. 2, pp. 928-938, 2013.
- [8] A. N. Hasan, P. S. P. Eboule and B. Twala, "The use of machine learning techniques to classify power transmission line fault types and locations," 2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) & 2017 Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP), Brasov, 2017, pp. 221-226
- [9] S. R. Samantaray, P. K. Dash and G. Panda, "Fault Classification and Ground detection using Support Vector Machine," TENCON 2006 - 2006 IEEE Region 10 Conference, Hong Kong, 2006, pp. 1-3
- [10] V. Malathi and N.S. Marimuthu, "Multi-class support vector machine approach for fault classification in power transmission". IEEE Conference Publications (2008), pp. 67-71.
- [11] Omar A.S. Youssef,"An optimised fault classification technique based on support-vector-machines". IEEE Conference Publications (2009), pp. 1-8.
- [12] H. A. Tokel, R. A. Halaseh, G. Alirezaei and R. Mathar, "A new approach for machine learning-based fault detection and classification in power systems," 2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, 2018, pp.1-5.
- [13] Qais Alsafasfeh, Ikhlas Abdel-Qader and Ahmad Harb "Symmetrical Pattern and PCA Based Framework for Fault Detection and Classification in Power Systems". IEEE Conference Publications (2010), pp. 1-6.
- [14] S. Bodda and P. Agnihotri, "Deep Learning based AC Line Fault Classifier and Locator for Power System," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019, pp. 1-5.
- [15] Pullabhatla Srikanth, Chiranjib Koley, An intelligent algorithm for autorecognition of power system faults using superlets, Sustainable Energy, Grids and Networks, Volume 26, 2021, 100450, ISSN 2352-4677

- [16] Ladislav Zjavka, Power quality multi-step predictions with the gradually increasing selected input parameters using machine-learning and regression, Sustainable Energy, Grids and Networks, Volume 26, 2021, 100442, ISSN 2352-4677
- [17] X. Du, Y. Cai, S. Wang and L. Zhang, "Overview of deep learning," 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, 2016, pp. 159-164.
- [18] S. Ruder, "An overview of gradient descent optimization algorithms," CoRR, vol. abs/1609.04747, 2016.

Machine Learning Approach to Predict Road Accidents in the United States

Sri Siddhartha Reddy
sgudemup@gmu.edu

Yen Ling Chao
ychao4@masonlive.gmu.edu

Lakshmi Praneetha Kotikalapudi
lkotikal@gmu.edu

Abstract—Transportation facilities are becoming more developed as society develops, and people’s travel demand is increasing, but so are the traffic safety issues that arise as a result. And car accidents are a major issue all over the world. The cost of traffic fatalities and driver injuries has a significant impact on society. The use of machine learning techniques in the field of traffic accidents is becoming increasingly popular. Machine learning classifiers are used instead of traditional data mining techniques to produce better results and accuracy. As a result, this project conducts research on existing work related to accident prediction using machine learning. We will use crash data and weather data to train machine learning models to predict crash severity and reduce crashes.

Index Terms—Accident, Machine Learning, Predict, United States

I. INTRODUCTION

The Association for Safe International Road Travel showed around 37000 persons die in automobile accidents each year, with another 2.35 million wounded or incapacitated by car accidents. Children under the age of 15 were responsible for 1600 deaths, while approximately 8000 persons between the ages of 16 and 20 were killed in automobile accidents [1]. As a result, vehicle accidents have become a societal hazard and one of the four leading causes of mortality in the metropolitan population. Every year, traffic accidents cost citizens hundreds of billions of dollars. Most of the losses were caused by a few big incidents. The purpose of major accident prevention is to avoid potentially hazardous road conditions. We can take appropriate actions and better manage financial and human resources if we can identify the primary reasons for catastrophic accidents. This case study’s data is a nationwide accident data collection encompassing 49 states in the United States. From 2016 to 2020, there will be 1.5 million road accidents.

A. Problem Statement:

The goal of this paper is to conduct a statistical analysis of the data, investigate the states with the most accidents, investigate when accidents are most likely to occur and the weather conditions at the time of accidents, and create a visual display: summarize and analyze the information, tell the overall situation of accidents in the United States, and discover the factors influencing the occurrence and severity of accidents; finally, the severity of the accident is predicted and evaluated.

B. Significance:

This research studies the current state of road accidents in the United States and offers recommendations for reducing the number and severity of accidents based on several data parameters. We shall examine this data from many perspectives. First, examine the distribution of accidents in the United States based on geography. Second, from a time perspective, examine the period of concentrated accidents and the volatility of total accidents in recent years. Next, the weather dimension is used to examine the diverse effects of various meteorological elements on severity. Finally, analyze the POI variables that may be easily increased and decreased to raise and decrease the accident rate from the POI dimension, and give appropriate adjustment suggestions. As a result, in this study, we will utilize Python’s panda module to evaluate and comprehend if current road accidents in the United States are rising or decreasing year by year, as well as which period is the peak period during which traffic accidents are most likely to occur. A simultaneous accident is a severe type of accident. Most of the accidents happen in either good or bad weather. We can clarify the causes of road

accidents and minimize the accident rate based on the findings of these data analyses.

C. The Reason For Using Big Data Solution:

Road safety has become a key focus of current social concerns as the number of traffic accidents has increased. The location of the accident, the time of day, the driver's emotions, the weather, and other unpredictable and complicated elements all play a role in road accidents [2]. As a result, the correlation between data from various components of the car accident data must be investigated. Understanding the factors that impact vehicle accidents and forecasting car accidents through data analysis can help reduce the number of fatalities and injuries. As a result, fast and accurate data collection, rapid analysis, and the creation of appropriate models have become critical requirements for big data traffic accident investigation.

II. LITERATURE REVIEW

Researchers used meta-analysis to assess the impact of bad weather on accident rates, demonstrating that weather has an impact on traffic safety. Extreme weather in the United States, according to the study, will have an impact on injuries and automotive accidents, notably on wet and snowy days, with the average percentage of collision rate on rainy and snowy days being 71 percent and 84 percent, respectively [3]. The severity of traffic accidents was assessed and forecasted using K-means clustering. This study compared the performance of two machine learning applications, random forest (RF) and Bayesian additive region trees (Bart), to investigate alternative machine learning approaches for predicting the severity of traffic accidents. It is discovered that, when compared to the prediction of the Bayesian additive region trees (Bart) model, the random forest (RF) model with meteorological circumstances has a greater prediction probability in estimating the severity of a traffic collision. The variable importance technique reveals that the mode of traffic accidents and weather circumstances are two major factors that might give significant information in the modeling and estimating processes [4].

Their study analyzes data from 50 states to discuss traffic accident patterns and causes, as

well as what can be done to prevent them, for US government agencies and the general public. Several variables that related to the severity of the accident were examined and analyzed using logistic regression [5].

Their analysis includes the number of accidents by year, number of accidents by state, the best time to travel by month, day, and hour, accident-prone areas in each state, factors that cause accidents such as weather, wind flow, temperature, location, and so on, deaths in each state, age groups of fatalities, drivers involved in accidents, drivers age groups, vehicles involved in accidents, and drivers who have consumed alcohol. Tableau was used to create the analysis platform. [6]

Another study tries to address this issue and go deeper into the elements that contribute to the rise in the number of car accidents. The data used in this study was collected constantly in the United States from 2016 to 2020 from traffic accident incidents captured by the Department of Transportation, law enforcement agencies, and traffic cameras. Two models were used to anticipate the impact of car accidents on traffic, with an emphasis on the primary causes of traffic accidents. The primary two factors determining car accident rates, according to the findings, are traffic induced by work rush hour and population density. [7]

Another research study was to present a model that can be used to explain why different countries have different rates of road deaths. As potential variables, national infrastructure, transportation, and socioeconomic indicators from international databases were explored. Stepwise regression analyses were used to create the model. [8]

R language demonstrates how data is related by analyzing traffic statistics and graphs. Locations of the accidents and the accident thermal chart were obtained after data preprocessing and data selection using R language Remap package remap and remapH functions. In addition, to model the data, they used decision trees, linear regression, and the random forest approach. We can check the model's correctness and obtain the most accurate model based on the actual findings, which will help in predicting the model's accuracy with comparable data in the future. After validating the model and examining the data properties and relationship

between the variables, the ultimate purpose of data analysis is to identify the most correct model. [9]

III. AIMS AND OBJECTIVES

Traffic accidents are the greatest cause of death worldwide, taking the lives of millions of people each year due to their regularity. As a result, technology that predicts traffic accidents or accident-prone areas may be able to save lives. Nowadays, there is an increasing emphasis on traffic accident data mining and analysis, which can improve in-depth investigation and reduce traffic-related deaths.

We will use Python pandas to analyze car accidents in the United States in this project. Real-time data, accident sites, casualty analysis, driving speed, traffic conditions, road structure, and weather may all be used to anticipate accidents. Therefore, we can predict accidents based on a variety of factors. We'll consider factors such as road conditions, speed limits, and the state in which the accident occurred. The analysis of historical accident data will assist in determining the likely causal relationship between these factors and road accidents, enabling the creation of accident predictors to reduce the risk of injury caused by accidents. As a consequence, utilizing this data collection, a machine learning model is built and applied that can accurately anticipate when and where accidents will occur, reducing the number of automobile accidents.

IV. METHODOLOGY

The vehicle accident dataset was obtained from the Sobhan Moosavi website, which has 1.5 million entries [10]. To undertake data analysis, the dataset must be preprocessed and has to be cleaned up by eliminating null values and filling in blanks, which will aid in data normalization. Data insights may be obtained for analysis and better decision-making using a variety of machine learning methods. Charts, graphs, and ordered tables can be used to visualize the data. One of the most significant jobs in road accident analysis is to predict the severity level of the event using various classifier methods such as logistic regression, decision trees, and random forest.



Fig. 1

V. PROPOSED APPROACH TO SOLVE PROBLEM

A. Focus:

Accidents in the United States may be utilized for several purposes, including real-time accident prediction, investigating the location of accident hotspots, fatality analysis and deriving causal principles to anticipate accidents, and researching the potential influence of climate on accidents. The analysis's goal is as follows:

- 1) Which state has the highest number of accidents in the United States?
- 2) What time of day is the most prone to accidents, and what is the pattern of accident change?
- 3) What connection exists between the accident and the weather?

B. Stages:

The project is broken into five stages:

1) *Problem Definition:* Every year, traffic accidents account for a high share of serious injuries reported. However, establishing the conditions that cause these occurrences is frequently difficult, making it more difficult for local law enforcement to handle the frequency and severity of traffic accidents. Many questions, however, remain unsolved. As a result, it is critical to forecast and investigate traffic accidents that we want to know which elements are the most important in causing automobile accidents and to understand which factors of the accident will have a bigger influence on traffic accidents.

2) *Data Collectin:* The practice of gathering and measuring data, information, or any variable of interest in a regulated and defined method that allows the collector to answer or test hypotheses and assess the findings of a specific collection is known as data collection. As a result, we gathered the accident dataset from Sobhan Moosavi's website which include 47 columns and 1.5 million records.

3) *Data Cleaning*: Data cleaning is an important phase in the data analysis process, and the quality of the findings is strongly related to the model impact and the conclusion. In practice, data cleansing often consumes 50% - 80% of the time allotted to the analysis process. Data cleaning can help to enhance data quality, reduce interference, and acquire useful and trustworthy data to aid decision-making.

4) *Data Visualization*: Data visualization is a way of communicating information more simply and effectively by conveying facts through data, primarily using graphical approaches. The goal of visualization is to present facts more naturally, making the data more objective and compelling. And data visualization is significant because graphs and charts may be more effectively represented, people, want to see different graphs, and it is simpler to recall. In a sea of data and information, it might be difficult to uncover connections, but graphs and charts can deliver information in seconds, and it gives people all the information they need at a look.

5) *Machine Learning*: One of the most significant study areas for machine learning algorithms is prediction. Using machine learning techniques, researchers may create traffic accident, prediction models. Among several models, the traffic accident prediction model is the most challenging to develop. Because the elements that influence automobile accidents are complex and changing. Because our goal is to forecast traffic accidents.

VI. PRELIMINARY RESULTS

Exploratory data analysis has been done on the dataset where cleaning and preparation of the data and then analyze it with different plots and visualizations. In the data preparation, the dataset file has been loaded using pandas, and cleaning is done by fixing missing or incorrect values. Some of the columns that we analyzed were State, start time, Start Lat, Start Long, Temperature, and Weather condition to analyze the severity of the accidents state-wise weather-wise. Some of the results produced on analysis are below. From that, we can see that during thunder weather conditions, accident severity is highest.

From the map, we see that California is recorded to be the highest accident-prone state

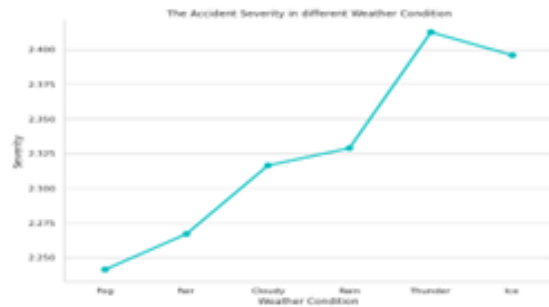


Fig. 2

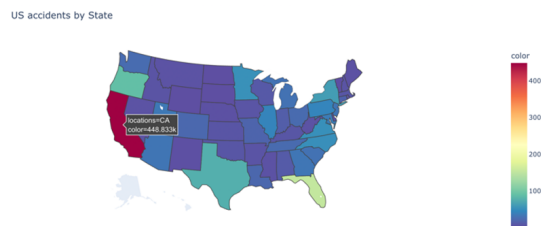


Fig. 3

According to this graph, the top ten states have the highest rate of auto accidents. As can be seen, the top three states with the most accidents are California, Florida, and Oregon, all of which are at the top of the US in terms of population and GDP. The expansion of traffic will be fueled by the economy's prosperity, and the likelihood of traffic accidents will rise as well.

The higher the number of accidents in the morning peak, the greater the traffic flow; and as the evening peak comes to an end, the serious accident rate continues to rise, reaching a peak at 17:00 and then gradually decreasing; it is speculated that this is due to fatigue driving after normal working hours. As a result, fatigue is more likely than low

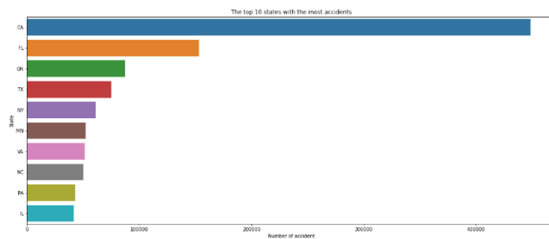


Fig. 4

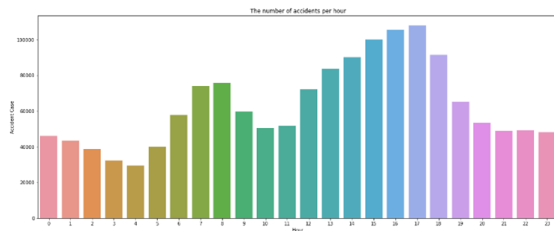


Fig. 5

Severity & Distance of accidents

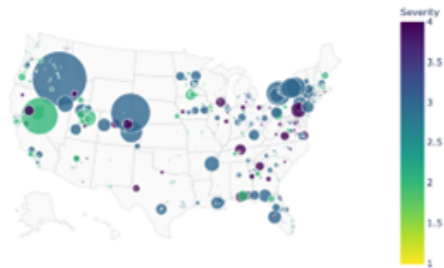


Fig. 6

visibility to cause serious accidents.

According to the findings, the number of car accidents in 2020 will be significantly higher than in 2016-2019, with many accidents occurring in the second half of 2020.

On analyzing the severity & distance of the accidents, we can see that severity 3 is spread more across the states although there are fewer states with severity 4 their distance is not widely spread.

The pie chart depicts the severity of the accident, with 1 being the least serious and 4 being the most serious. According to the findings, the majority of traffic accidents have a severity of 2. According to annual accident severity statistics, the severity of traffic accidents in 2020 is 2, and the severity of most traffic accidents in 2019 is also 2.

Considering the climate conditions such as temperature and pressure, on analyzing them, can infer that all the severity 2 is more when the pressure is between 29.5 and 30.5 and found that other severities are recorded to be more at that particular pressure point.

Severity of accidents

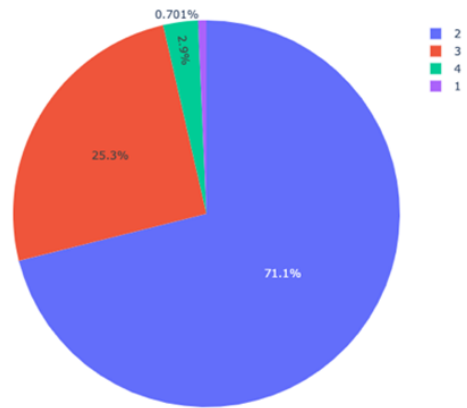


Fig. 7

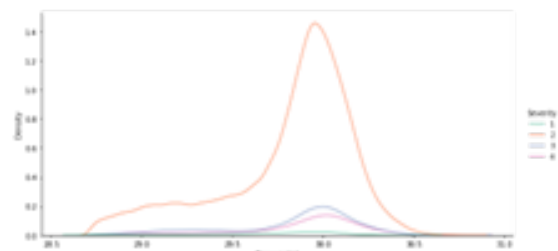


Fig. 8

Considering temperature as a factor, severity 2 is highest when the temperature is recorded between 40F and 80F.

Other references [?], [11]–[17]

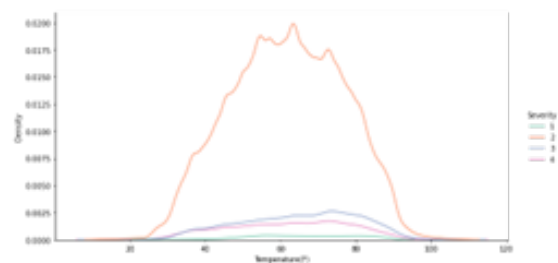


Fig. 9

TABLE I: Project timeline

	TASK	TO BE COMPLETED BEFORE
1	Project Proposal	23 February 2022
2	Preprocessing the data	4 March 2022
3	Analyzing the data and visualizing the findings	9 March 2022
4	Preparing draft for project milestone 1	16 March 2022
5	Finalizing Project Milestone 1	16 March 2022
6	Training, evaluating, and validating the models	28 March 2022
7	Reviewing and documenting the findings	13 April 2022
8	Preparation for project presentation	18 April 2022
9	Project paper	29 April 2022
10	Final proof reading and submission	6 May 2022

REFERENCES

[1] T. Sridharan. How many car accidents occur in the u.s. each year? (November 3, 2020). [Online]. Available: <https://www.1800thelaw2.com/resources/vehicle-accident/how-many-accidents-us/>

[2] C. Chen, "Analysis and forecast of traffic accident big data," in *ITM Web of Conferences*, vol. 12. EDP Sciences, 2017, p. 04029.

[3] L. Qiu and W. A. Nixon, "Effects of adverse weather on traffic crashes: systematic review and meta-analysis," *Transportation Research Record*, vol. 2055, no. 1, pp. 139–146, 2008.

[4] A. R. Mondal, M. A. E. Bhuiyan, and F. Yang, "Advancement of weather-related crash prediction model using nonparametric machine learning algorithms," *SN Applied Sciences*, vol. 2, no. 8, pp. 1–11, 2020.

[5] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 729–741, 2002.

[6] G. Li, T. Huang, H. Yan, and X. Huang, "Grey residual error model of highway traffic accident forecast [j]," *Journal of Traffic and Transportation Engineering*, vol. 9, no. 5, pp. 88–93, 2009.

[7] elsevier. Journal. Accident analysis & prevention. (n.d.). Retrieved March 20, 2022. [Online]. Available: <https://www.sciencedirect.com/journal/accident-analysis-and-prevention>

[8] M. Aljaban, "Analysis of car accidents causes in the usa," 2021. [Online]. Available: <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12218&context=theses>

[9] C. J. Bester, "Explaining national road fatalities," *Accident Analysis & Prevention*, vol. 33, no. 5, pp. 663–672, 2001.

[10] S. Moosavi. Us-accidents: A countrywide traffic accident dataset. (January 2021). [Online]. Available: https://smoosavi.org/datasets/us_accidents

[11] Y. Qian, X. Zhang, G. Fei, Q. Sun, X. Li, L. Stallones, and H. Xiang, "Forecasting deaths of road traffic injuries

in china using an artificial neural network," *Traffic injury prevention*, vol. 21, no. 6, pp. 407–412, 2020.

[12] S. Saha, P. Schramm, A. Nolan, and J. Hess, "Adverse weather conditions and fatal motor vehicle crashes in the united states, 1994-2012," *Environmental health*, vol. 15, no. 1, pp. 1–9, 2016.

[13] M. Feng, J. Zheng, J. Ren, and Y. Liu, "Towards big data analytics and mining for uk traffic accident analysis, visualization & prediction," in *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, 2020, pp. 225–229.

[14] H. Manner and L. Wünsch-Ziegler, "Analyzing the severity of accidents on the german autobahn," *Accident Analysis & Prevention*, vol. 57, pp. 40–48, 2013.

[15] Z. Christoforou, S. Cohen, and M. G. Karlaftis, "Vehicle occupant injury severity on highways: An empirical investigation," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1606–1620, 2010.

[16] P. A. Nandurde and N. V. Dharwadkar, "Analyzing road accident data using machine learning paradigms," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2017, pp. 604–610. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8058251>

[17] N. Tyagi, N. Chauhan, A. Sighal, and R. Khan, "Sanjyot—we save life using big data-apache spark," *EAI Endorsed Transactions on Cloud Systems*, vol. 6, no. 19, p. 166660, 2020.

Dynamic simulation of a microgrid system for a university community in Nigeria

Stephen Ogbikaya
 Department of Electrical and Computer Engineering
 Memorial University of Newfoundland
 St. John's, Canada
 sogbikaya@mun.ca

M. Tariq Iqbal
 Department of Electrical and Computer Engineering
 Memorial University of Newfoundland
 St. John's, Canada
 tariq@mun.ca

Abstract— In this research, the optimal PV size of the system was first determined with the aid of OpenSolar, PVWatts and REopt software for an annual energy consumption of 969,000kWh as compared to that obtained from HOMER Pro software. The result obtained was then used to design and simulate the dynamic model of the campus microgrid system in MATLAB/Simulink environment. The designed system in Simulink consists of PV size of 675.2kW comprised of 1350 modules each of 500W, with 25 connected in series and 54 in parallel, a utility grid of 50Hz, 6MVA, 33 kV transmission network with a power transformer of 2.5MVA, 33/11 kV and a distribution transformer of 750 kVA, 11/0.415 kV and a back-up generator of 1.5 MVA rating. The system dynamics was considered under three conditions which includes PV + grid mode, PV + generator mode and only generator mode. Simulated results include transient responses observed when the utility grid and the generator were connected to the network separately through their respective toggle switches. Waveform of power, voltage and current through the varied electric load of the system for various conditions are presented in this paper.

Index Terms— Dynamic simulation, MATLAB/Simulink software, microgrid, PV sizing, renewable energy.

I. INTRODUCTION

A microgrid consist of a network of electricity consumers with a local source of supply usually attached to a centralized national grid but it can function separately. In this case the consumers include residential homes, stores, streetlights, health centers, churches and even schools. Microgrids are more efficient, low-cost, clean energy, resilient, and improve the operation and stability of grid system. They can be classified into remote, grid-connected and networked type. In this research, the dynamics of a microgrid system for a university community in Nigeria that is grid connected is considered.

II. LITERATURE REVIEW

Research [1] work explain that for future power systems, the deficiency in power supply can be solved by a microgrid system which consists of renewable power sources such as

wind, solar and hydro. In this research, a microgrid system with wind and solar power sources was used to solve the issues relating to operation, control and stability of the system. With the aid of MATLAB/Simulink, the system was modeled and simulated based on the renewable power generation units to know the relevant technical issues associated with the operation of a microgrid system.

The paper [2] reveals that microgrid application at different voltage levels for power system networks are in the increase. To increase energy efficiency, reduce electricity bills and reduce the problems with power delivery to customers, community microgrid systems are encouraged. In that work, a simulation-based electricity analysis scheme for a community microgrid with the aid of proposed modelling methodology, simulation mechanisms, and a power balancing control strategy in MATLAB environment was developed. That paper also presents an effective simulation mechanism using MATLAB/Simulink software for the electricity analysis in microgrid systems which can be flexibly modified to comply with different simulation requirements when faced with various system topologies.

In paper [3], a solar PV powered DC microgrid that was proposed and designed in Nigeria at Umuokpo Amumara community with a household population of 800 was presented. The component sizes selected always meets the load demand of the community with energy requirement of 3.16MWh/day. A battery capacity of 21,944Ah was sized as the battery storage to meet the energy required by the community for one day without any renewable energy source. In this research, the dynamic model of the microgrid was simulated in MATLAB/Simulink to observe the system's dynamic response in view of the power quality, load impact, and battery storage charging. Results obtained from simulation indicates that the stand-alone DC microgrid can meet the daily electric energy of the system with relatively good voltage stability.

Study [4] stated that the main aim of microgrid (MG) system was to integrate microsources and load into a controlled system to supply electric energy to the end

consumers. In this work the dynamic characteristics of a grid connected system associated with power conditioning system (PCS) to regulate its power was discussed. The system investigated consists of grid, microsources, lumped static loads and other components. To cater for the dynamic behaviour of the system, a detailed model of the system was developed in MATLAB/Simulink in which two operational modes were investigated which includes: four-quadrant operation of PCS and use of PCS to control the power of MG. Simulation results indicate that MG can operate satisfactorily in these operational modes.

Paper [5] presents the dynamic operation of a low-voltage microgrid system with various distributed energy sources and the system consists of a low-voltage microgrid with a 30 kVA micro-hydro generator, a 30 kVA diesel engine generator, a 30 kVA gas engine generator, and a 15 kVA micro-wind turbine generator including the loads. In this research, the individual components of the system were modeled and integrated to form a microgrid system whose dynamic simulation was simulated in MATLAB/Simulink environment. Results obtained provide information on the dynamic characteristics of AC low-voltage microgrid systems and help in the development of microgrid system in Taiwan.

In this paper [6] the development of a microsources booster converter and PWM inverter in MATLAB/Simulink and combining them to form a microgrid was presented. The system was designed such that it can operate both in islanding mode and in grid mode. The purpose of this work was to lay foundations which allow further investigation and development of more complex microgrid models to know the behaviour of microgrid systems.

In paper [7], the design of a network-based scheme for inverter-based sources was studied. This scheme provides proper current control when connected in grid and voltage control during islanding mode. In that work, the algorithm for international islanding detection and synchronization controller needed for the grid connection was developed. When the dynamic modeling and simulation of system was conducted in Simulink, results reveal that the controllers provide the microgrid with a deterministic and reliable connection to the grid.

In this work [8] behaviour of hybrid AC/DC microgrid system was analyzed when it was grid tied. The microgrid in this case was developed by photovoltaic system, wind turbine generator and battery. For proper coordination from AC sub-grid to DC sub-grid, a control mechanism was developed for the converters. The system was simulated with the results obtained in MATLAB/Simulink. The results obtained indicate that the hybrid grid system provides efficient power with high quality and more reliable power to the consumers which may be feasible for small isolated industrial plants with both PV systems and wind turbine generator as the major power supply.

In research [9], the modeling and performance of a

microturbine generating (MTG) system in grid connected and islanding mode was analyzed with the aid of MATLAB/Simulink software. Based on the various conditions considered in that study, simulated results show that MTG system follows the load with variation of fuel flow, power and temperature of the system. It was also observed that the MTG system contributes clean, reliable and cost-effective energy for future distributed generation (DG) systems.

In paper [10], the design and operation of a microgrid (MG) for the main campus of the Technical Institution Hawija was considered. That microgrid design includes a battery energy storage system (BESS), photovoltaic (PV) generation system, and controllable loads. The efficacy of their intended design was simulated in MATLAB/Simulink. Simulated results indicate that the distribution system obtained is more robust, reliable and resilient against weather disaster or technical issues.

In this research, the dynamic simulation of a microgrid system for a university community in Nigeria is presented. The system consists of a PV size of 675.2 kW comprising of 1350 modules each of 500W, with 25 connected in series and 54 in parallel with a utility grid system and a diesel generator set is incorporated in case of emergency. The system was then designed and simulated in MATLAB/Simulink environment to test its dynamics. This work is novel in a sense that it provides a Nigeria case study, include design, dynamic modeling and simulation.

III. METHODOLOGY

Based on the campus annual electric load (kWh), a microgrid for the university community was designed in HOMER Pro software to obtain the PV size (kW) of the system. Further design from OpenSolar, PVWatts and REopt softwares were done based on the same annual energy consumption (kWh) of the selected site to determine the optimal PV size (kW) of the system as compared to that obtained from the HOMER Pro software. Based on the optimal PV size selected, the system was then simulated in MATLAB/Simulink software to determine its dynamics.

IV. SYSTEM DESIGN

The university campus ($7^{\circ} 8'8.25''N$, $6^{\circ}18'28.13''E$) at Kilometer 7, Auchi-Abuja Road, Iyamho-Uzairue, Edo State, Nigeria is located as shown in Figure 1 and the annual energy consumption of the site selected was determined as 969,000kWh based on the monthly energy consumption of the campus as shown in Table 1.



Fig. 1 Overview of university campus

TABLE I
SUMMARY OF MONTHLY ENERGY CONSUMPTION OF THE
SELECTED SITE FROM OCTOBER 2020 TO SEPTEMBER 2021

Month	Previous meter reading (kWh)	Present meter reading (kWh)	Energy consumption (kWh)
January	3635000	3655000	20000
February	3655000	3689000	34000
March	3689000	3781000	92000
April	3781000	3836000	55000
May	3836000	3893000	57000
June	3893000	3986000	93000
July	3986000	4097000	111000
August	4097000	4254000	157000
September	4254000	4324000	70000
October	4324000	4411000	87000
November	4411000	4516000	105000
December	4516000	4604000	88000

Annual energy consumption 969000

The load profile of the designed system based on HOMER Pro software is shown in Figure 2 and the solar irradiance of this site is shown in Figure 3. The average solar irradiance per annum of this location is 5.10 kWh/m²/day.

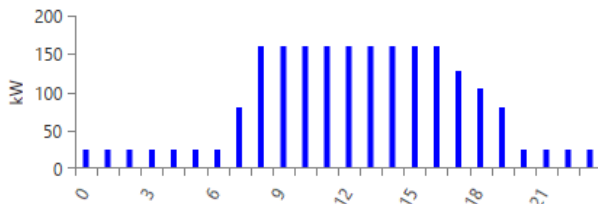


Fig. 2 Load profile of university campus

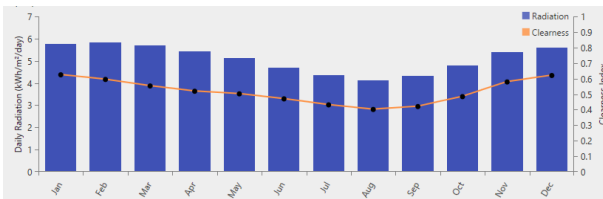


Fig. 3 Monthly solar irradiance of university campus

Based on this annual energy consumption and selected components, a microgrid system of the site was designed. This microgrid consist of solar panel of 0.5kW, the grid system, diesel generator of 1.5MVA rating and inverter sinexcel 500 with the aid of HOMER Pro Energy software to accommodate the annual energy consumption (kWh). The schematic diagram of the microgrid system obtained is shown in Figure 4.

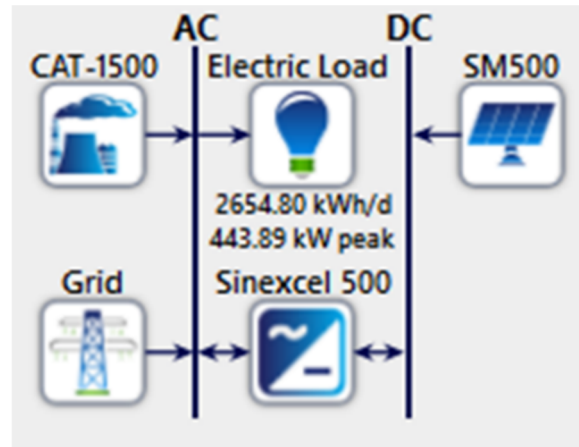
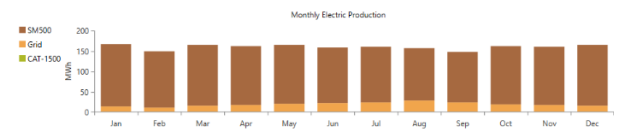
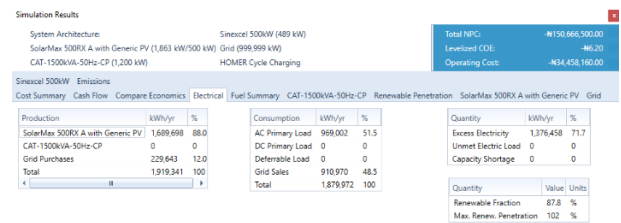


Fig. 4 Schematic diagram of microgrid system

The result obtained from the HOMER Pro software indicates that the system would need 1,868kW PV size for system installation to meet the 87.8% consumption of the



proposed microgrid system as shown in Figure 5.

Fig. 5 Simulated result from HOMER Pro showing system PV size

To determine the optimal PV size of the microgrid system of the site with the same annual energy consumption (kWh) for dynamic simulation, PVWatts, OpenSolar and REopt were used to design the system.

A. PVWatts simulation

Based on the site location and energy demand of the campus, PVWatts was used to design and simulate the system to determine the PV size that would be required. Result obtained indicates that the system would require 670.5kW PV size for the system installation this is depicted in Figure 6 and Figure 7.

PVWatts Calculator

RESULTS

967,113 kWh/Year*

Month	Solar Radiation (kWh / m ² / day)	AC Energy (kWh)	Value (\$)
January	5.53	85,879	11,164
February	5.80	81,845	10,640
March	5.96	92,340	12,004
April	5.51	83,657	10,875
May	5.05	78,874	10,254
June	4.65	71,768	9,330
July	4.51	72,440	9,417
August	4.66	74,950	9,744
September	5.01	77,664	10,096
October	5.31	83,995	10,919
November	5.42	81,916	10,649
December	5.25	81,785	10,632
Annual	5.22	967,113	\$ 125,724

Location and Station Identification

Requested Location	Edo State University Uzairue, Auchi, Nigeria	
Weather Data Source	(INTL) ACCRA/KOTOKA INTL, GHANA	436 mi

Fig.6 Simulation from PVWatts showing system annual energy consumption

Requested Location	Edo State University Uzairue, Auchi, Nigeria	
Weather Data Source	(INTL) ACCRA/KOTOKA INTL, GHANA	436 mi
Latitude	5.6° N	
Longitude	0.17° W	
PV System Specifications (Commercial)		
DC System Size	670.5 kW	
Module Type	Standard	
Array Type	Fixed (open rack)	
Array Tilt	8°	
Array Azimuth	180°	
System Losses	14.08%	
Inverter Efficiency	96%	
DC to AC Size Ratio	1.2	
Economics		
Average Retail Electricity Rate	0.130 \$/kWh	
Performance Metrics		
Capacity Factor	16.5%	

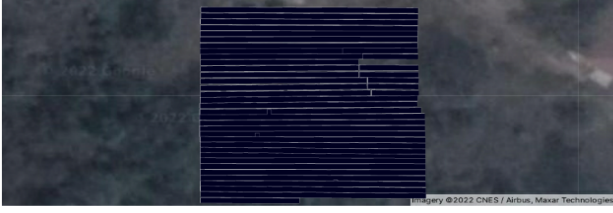
Fig. 7 Simulation from PVWatts showing system PV size

B. OpenSolar Simulation

Figure 8 show that 683.645kW PV size would be required for the same energy demand of the campus when OpenSolar software was used to simulate the system.

Recommended System Option

683.645 kW System Size	\$50,389 Estimated Annual Electricity Bill Savings	\$1,504,019 Total System Price	\$1,504,019 Net System Price
---------------------------	--	-----------------------------------	---------------------------------



Your Solution

Solar Panels

LG Electronics Inc.

683.645 kW Total Solar Power

1873 x 365 Watt Panels (LG365Q1C-A5)

969,063 kWh per year

Fig. 8 Simulation from OpenSolar software showing system PV size

C. REopt Simulation

When REopt software was used to design the system as shown in Figure 9, the PV size recommended for this design was 658kW as represented in Figure 10.

Technologies Selected	
	PV
Site and Utility	
Site Location	Auchi, Nigeria (7.066864499999999, 6.274773400000001)
PV & wind space available	Land
Annual energy charge (\$/kWh)	\$0.13
Annual demand charge (\$/kW/month)	\$1.05
Load Profile	
Typical electric load profile type	simulated campus
Campus total electric energy consumption (kWh)	969,000
Building #1	SecondarySchool (100% of total energy consumption)

Fig. 9 Simulation from REopt showing system annual energy consumption

Results for Your Site



Your site at Auchi Nigeria evaluated on February 13, 2022

These results from REopt summarize the economic viability of PV, wind, battery storage, and/or CHP at your site. You can edit your inputs to see how changes to your energy strategies affect the results.

Your recommended solar installation size

658 kW

PV size

Measured in kilowatts (kW) of direct current (DC), this recommended size minimizes the life cycle cost of energy at your site.

This optimized size may not be commercially available. The user is responsible for finding a commercial product that is closest in size to this optimized size.

Fig. 10 Simulation from REopt software showing system PV size

The resulting PV sizes obtained from these software as compared to that of the HOMER Pro software is summarized as shown on the comparison table in Table 2. We believe

TABLE 2
COMPARISON TABLE SHOWING THE SIMULATED RESULTS OBTAINED FROM THE DIFFERENT SOFTWARE USED TO DETERMINE THE SYSTEM PV SIZE FOR THE MICROGRID SYSTEM OF THE UNIVERSITY CAMPUS

S/N	Software Used	Annual Energy Consumption (kWh)	System PV Size (kW)
1	HOMER Pro	969,000	1868.00
2	OpenSolar	969,063	683.65
3	PVWatts	967,113	670.50
4	REopt	969,000	658.00

results of REopt from NREL are more reliable and we used that for the next section.

IV. DYNAMIC SIMULATION

The dynamic system design of the university campus microgrid was then carried out with the aid of MATLAB/Simulink software. The system consists of a PV size of 675.2kW comprised of 96 cell modules each of 500W, with 25 connected in series and 54 in parallel. To ensure that the PV gives maximum power, MPPT is applied to the PV. Figure 11 shows the simulation results of I-V and P-V curves of the PV output.

The inverter of the system was selected based on the PV size as 700kW. Since the microgrid system is an on-grid system, a utility grid of 50Hz, 6MVA, 33 kV transmission network with a power transformer of 2.5MVA, 33/11 kV and a distribution transformer of 750 kVA, 11/0.415 kV was incorporated to the network through breaker 2, which is controlled by toggle switch 2. Also incorporated into the network is a permanent type 1.5MVA rating generator through breaker 3, that is controlled by toggle switch 3. This stands as a backup when the PV and grid system fails. The electric load of the university campus is also connected and varied through breaker 4 which is controlled by toggle switch 4 to network. Figure 12 shows the diagram of the dynamic simulation of the entire system in MATLAB/Simulink environment.

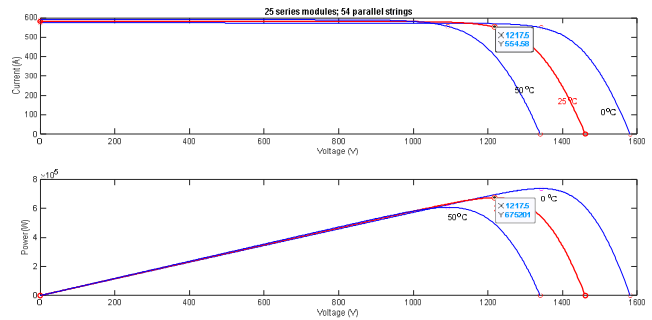


Fig. 11 Irradiation, temperature, power, voltage and duty cycle signal of PV module

V. SIMULATION RESULTS

The Irradiation, Temperature, Power, Voltage, and duty cycle graphs of the PV module are shown in Figure 13. Simulated results shows that the duty cycle varies according to the solar resources and the MPPT is able to track the variation of the solar input.

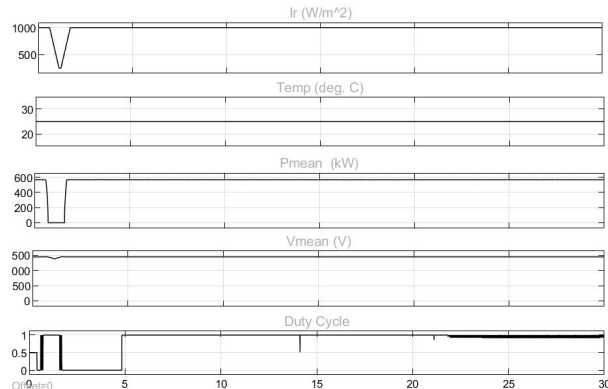


Fig. 13 Irradiation, temperature, power, voltage and duty cycle signal of PV module for 30 seconds

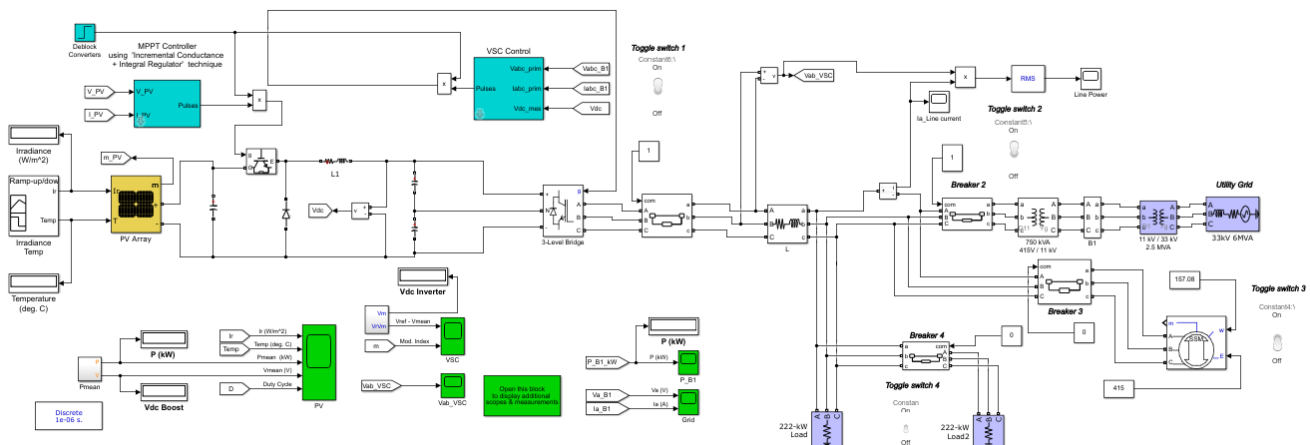


Fig. 12 Dynamic simulation of campus microgrid in MATLAB/Simulink software

Dynamic modeling and simulation of systems shows how the system behaves under proposed conditions using Simulink. In this case, the dynamic simulation shows the behaviour of the campus microgrid system when the utility grid and generator are connected to the network separately. For the purpose of analysis, the system dynamics are considered in three cases.

- i. PV + Grid mode i.e normal operating condition of the proposed system.
- ii. PV + Generator mode i.e when the grid fails.
- iii. Generator only i.e when the generator supplies energy to the entire load of the system.

The behaviour of the system based on the three cases stated above are illustrated as depicted in the respective graphs.

i. PV + grid mode

This mode is achieved when the toggle switches 1 and 2 are in closed state as shown in Figure 12. This is the normal operation of the system. In this case, the PV modules and the grid system supplies energy to the electric load of the network. Figure 14 shows the power curve and Figure 15 and Figure 16 shows the voltage and current through the load during this mode. The dynamic behaviour of the network is summarized as follows. The PV and grid generate energy to accommodate the load of the system which is divided into two by breaker 4 controlled by toggle switch 4. The variation in the ac power, voltage and current curves between the offset time $t = 1.15\text{secs}$ to $t = 1.53\text{secs}$ was when the balance half of the load was connected to the network through the toggle switch 4.

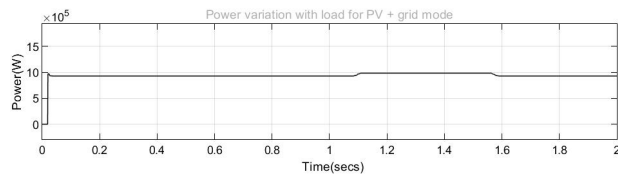


Fig. 14 Power variation with load for PV + grid mode

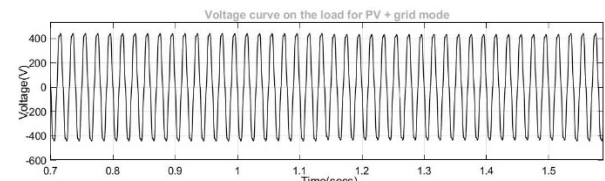


Fig. 15 Voltage curve on the load for PV + grid mode for a changing load

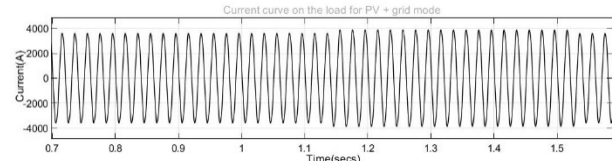


Fig. 16 Current curve on the load for PV + grid mode for a changing load

ii PV + Generator mode

PV + generator mode is achieved when the toggle switches 1 and 3 are in the closed state while breaker 2 is open as shown in Figure 12. In this case, the switches were closed throughout the simulation while the electric load was varied through breaker 4. The transient behaviour of the system in response to this new state is depicted in Figure 17, Figure 18 and Figure 19 which shows the power curve and the voltage and current curves across the electric load. Load varies from 222kW to 272kW at $t = 0.95\text{secs}$ and $t = 1.43\text{secs}$.

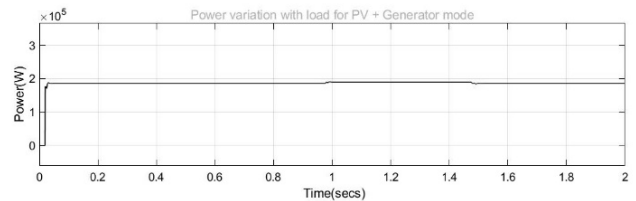


Fig.17 Power variation with load for PV + Generator mode

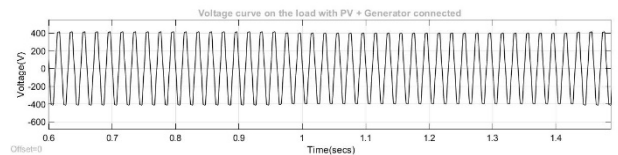


Fig. 18 Voltage curve on the load for PV + Generator mode for a changing load

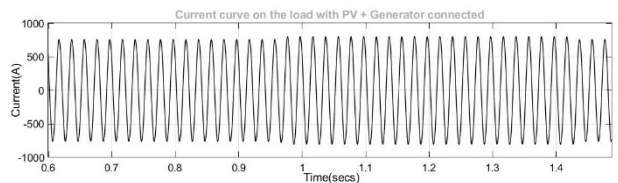


Fig. 19 Current curve on the load PV + Generator mode for a changing load

iii Generator only

The generator is introduced into the network as an emergency backup in case the grid system fails when the PV system is not generating energy to supply the system load. Toggle switch 3 controls the generator circuit in the network. When it is ON, the generator is connected to the network and when it is OFF, the generator is out of the system. Figure 20, Figure 21 and Figure 22 shows the graphs of the power and voltage and current through the load when the electric load of the system is varied through toggle switch 4. This transient response on the graphs indicates the behaviour of the system when only the generator is connected to the network. At $t = 1.0\text{sec}$ and $t = 1.32\text{secs}$, load increased from 222kW to 444kW.

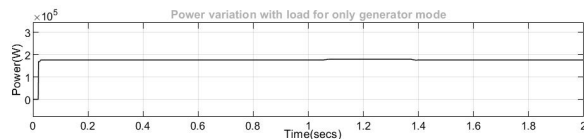


Fig. 20 Power variation with load for only generator mode

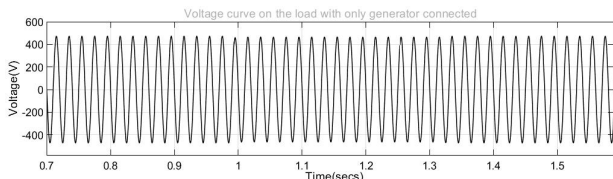


Fig. 21 Voltage curve on the load for only generator mode for a changing load

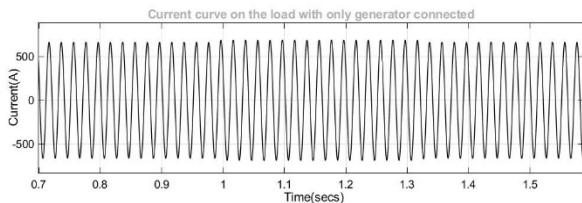


Fig. 22 Current curve on the load for only generator mode for a changing load

VI. CONCLUSION

In this research, the dynamic simulation of a microgrid system for a university community in Nigeria is presented. The system designed was simulated in MATLAB/Simulink environment. The system consists of a PV size of 675.2 kW comprising of 96 cell modules each of 500W, with 25 connected in series and 54 in parallel. Also, a utility grid system and a diesel generator set in case of emergency were connected through toggle switches 2 and 3. Simulated results indicates that the system realized has acceptable dynamics as it responds appropriately when a new state was introduced to the network by varying the electric load of the network for three different cases. This was shown by the transient responses indicated on the system when the toggle switch 4 was turned ON and OFF to vary the electric load during simulation to meet the load demand of the university campus. We recommend Nigerian university community to find funds for implementation of such a system to have a reliable low cost electricity.

ACKNOWLEDGMENT

Authors would like to thank School of Graduate Studies (SGS), Memorial University of Newfoundland, National Science and Engineering Research Council Canada for funding this research.

REFERENCES

[1] M.A. Fouad, M.A. Badr and M.M. Ibrahim “Modeling of Micro-Grid System Components using MATLAB/Simulink”, *Global Scientific Journals*, Volume 5, pp. 163 – 177, 2017.
 [2] Y. J. Liu, S. I. Chen, Y. R. Chang and Y. D. Lee, “Development of a Modelling and Simulation Method for Residential Electricity Consumption Analysis in a Community Microgrid System” *Applied Sciences*, pp. 1 -19, 2017.

[3] C. Ndukwe and T. Iqbal “Sizing and dynamic modelling and simulation of a standalone PV based DC microgrid with battery storage system for a remote community in Nigeria” *Journal of Energy Systems*, Volume 3, pp. 67 – 85, 2019.
 [4] M. J. Chen, Y. C. Wu and Y. X. Huang, “Dynamic behavior of a grid-connected microgrid with power conditioning system”, *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, Volume 27, pp. 318 -333, 2014.
 [5] W. T. Huang, K. C. Yao, C. C. Wu and S. W. Wang, “Dynamic Simulation and Analysis of a Low-Voltage Micro-Grid”, *2012 International Conference on Computing, Measurement, Control and Sensor Network*, pp. 245 – 248, 2012.
 [6] R. Chowdhury and T. Boruah, “Design of a Micro-Grid System in MATLAB/Simulink”, *International Journal of Innovative Research in Science, Engineering and Technology*, Volume 4, pp. 5262 – 5269, 2015.
 [7] R. J. Vijayan, S. Ch and R. Roy, “Dynamic Modeling of Microgrid for Grid Connected and Intentional Islanding Operation”, *IEEE*, pp. 1 – 6, 2012.
 [8] A. Sheikh and P. Chavan, “Design & Simulation of Micro grid”, *International Refereed Journal of Engineering and Science (IRJES)*, Volume 5, pp. 7 – 15, 2016.
 [9] S. K. Nayak and D. N. Gaonkar, “Modeling and Performance Analysis of Microturbine Generation System in Grid Connected/Islanding Operation”, *International Journal of Renewable Energy Research*, Volume 2, pp. 750 – 757, 2012.
 [10] Z. H. Ali, Z. H. Saleh, R. W. Daoud and A. H. Ahmed, “Design and simulation of a microgrid for TIH campus”, *Indonesian Journal of Electrical Engineering and Computer Science*, Volume 19, pp. 729 – 736, 2020.

DESIGN OF A COMPUTER STEREO VISION BASED ROAD MARKING SYSTEM

1st SURYAKANT

ECE

NITTTR

Chandigarh, India.

suryakantsurya1981@gmail.com

2nd AMOD KUMAR

ECE

NITTTR

Chandigarh, India.

csioamod@yahoo.com

3rd GARIMA SAINI

ECE

NITTTR

Chandigarh, India.

garima@nitttrchd.ac.in

Abstract—Existing methods to put white and yellow marking on the road are based on manual measurement of the road dimensions and manual marking of guide lines. It is a time consuming, cumbersome and labour intensive process. This was the motivation for this work to develop a system which can automatically apply road markings according to the width of the road. In this paper, a stereo-vision based road marking system is presented. Computer stereo-vision has been used to measure the width of the road which helps us decide the number of lanes on the road. To generate the disparity map, a semi-global matching algorithm has been used, incorporating the advantages of both local and global matching algorithms. Efficient algorithms have been used for camera calibration and to estimate the road width. For deciding the number of lanes, the standards laid down by Indian Road Congress (IRC) have been followed..

Index Terms—Road markings, stereo vision, disparity map, semi-global matching, computer vision, image processing, automation

I. INTRODUCTION

Computer vision is a popular field of science and technology. With the surge in usage of artificial intelligence in last two decades, extensive work was done by researchers in this field. From the medical field to industrial robotics and space exploration, computer vision has played an important role. Computer vision is the field that deals in understanding the features of the objects present in the workspace by computer which provide useful information about the environment. Computer vision has many techniques for analysing and understanding the captured image data. Application of these techniques depends upon the problem in hand.

There are two methods to capture the environment data in image format: Mono-Vision and Stereo-Vision. The Mono-Vision method incorporates a single camera to capture the image and extract 2D information from the image. Mono-Vision technique has many applications like object detection (e.g. cancer cell detection, plant disease detection etc.), object recognition and object classification. Stereo-Vision method incorporates two cameras that are placed a known distance apart. In this method, two cameras capture the images of the same scene of the environment simultaneously. This method

reconstructs the 3D information by correlating the two images taken from the pair of cameras.

According to a World Health Organisation report of 2018, every year around 13 million people die in traffic accidents while 20 to 50 million people become disabled in these accidents [1]. Roads are the economic arteries for a nation; they allow the flow of services throughout the whole country which helps in rising of the nation economically. These are the most important infrastructure in the present era. However, they deteriorate with time. Deterioration of road includes the decaying of road marking lines. Poorly maintained road markings cause many problems like reduced traffic mobility, increased number of accidents, traffic congestions etc.

Now-a-days, we are moving towards automation in every industry, organization and service. Every new vehicle comes with some ADAS (Advanced driver-assistance system) features like parking assist system, cruise control system and lane centering system. In case of self-driving cars, appreciable research is going on to increase accuracy and ease of operation. Many companies have launched self-driving cars in the developed countries. All these technologies rely on the road markings to guide or assist the movement of vehicles. Therefore, good quality, accurate road markings are very important.

In this paper, a road marking system is designed using computer vision. Using stereo vision, the width of road is measured and a center line is drawn on road if width of road is more than the two lane road according to the Indian Road Congress (IRC) protocol.

II. RELATED WORK

There are many methods to apply road markings which make use of different approaches. They can be broadly classified into two categories – manual and automatic. In manual methods, an extended bar [2] [3] [4] and periscope [5] are used which align the paint sprayer with guideline markings. Manual processes require complete attention of the manual human operators. Automatic methods require only marginal attention from the machine operator. While manual approach may work in reviving marking of old roads where previous

marking has faded, it is highly unsuitable and undesirable for new roads. In automatic methods for reviving the old faded markings, people have used different types of detectors to detect the old line-markings as guide marks for applying the new paint. These types of detectors are combination of cameras and laser scanner [6] [7], electromagnetic radiation [9] [10] [11], UV light scanners [11] and video cameras [12] [13] [14]. In other automatic systems, people have used GPS technology to record and locate the marks to paint the road markings [15] [16]. All these systems have some drawbacks. Manual system has human error while the radiation-based system is affected by illumination. These systems fail if the old road markings are eroded. GPS based systems cannot work in an area where there is no availability of the Internet or some obstruction is there to the signal. All these systems can work only on the roads which have been marked earlier which are still not in bad condition. They cannot work for newly constructed roads.

III. METHODOLOGY

The aim of this research is to develop a method for obtaining the width of the road and calculate the position of road markings. For applying Road marking, dimensions of the road need to be measured. On a two-lane road, centre line of road is drawn to define the area of road access to the motorist, order the traffic and maintain a smooth flow. Centreline defines that each half of the road will provide access to one-sided driving in opposite direction. For example, in Indian road scenario, traffic has to be on the left side of the centre line while driving.

To determine the width of road, depth map has to be generated using stereo-matching algorithm. Following steps are required for implementation of marking lines:

- Camera calibration
- Image rectification
- Road segmentation
- Disparity map generation
- Calculation of road width
- Determination of marking position

These tasks are discussed in sequence in next subsections.

A. Camera Calibration

In general, two cameras from the same manufacturer are never identical and stereo camera pair cannot produce two horizontally aligned images. Camera calibration is the process of estimating the intrinsic and extrinsic parameters of camera with the help of images captured by the camera. These parameters help in rectification of images (make epipolar lines parallel to the x-axes) to prevent distortion of the images. The cameras are calibrated by repeatedly shooting a calibration pattern with known points for which their relative positions in space are known. Generally, a chessboard with known number of squares is used as a calibration pattern. In this work, 82 image pairs of a chessboard pattern containing 7×9 boxes, each box of 35mm × 35mm size, were captured. Camera calibration

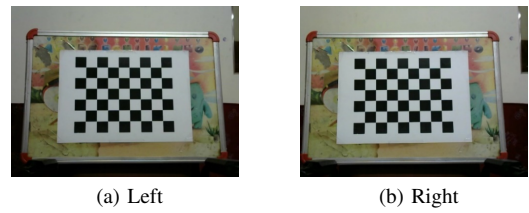


Fig. 1: Images of calibration pattern (a) from left camera (b) from right camera

was done by using Heikkila, J, and O. Silven [17] method by using MATLAB Stereo Camera Calibrator application.

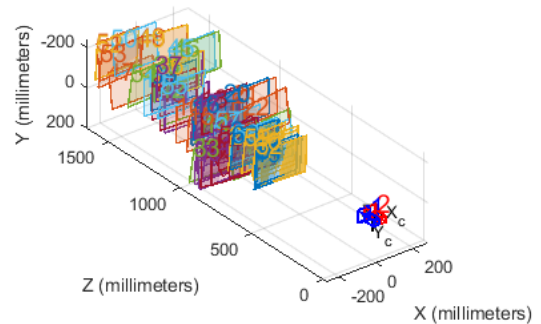


Fig. 2: Camera centric view of calibration showing camera pair and all images for calibration

Figure 3 shows the reprojection error for all the captured images of calibration pattern. The reprojection error is a geometric error corresponding to the image distance between a projected point and a measured one. It is used to quantify how close a 3D recreated point is to the point's true projection [18].

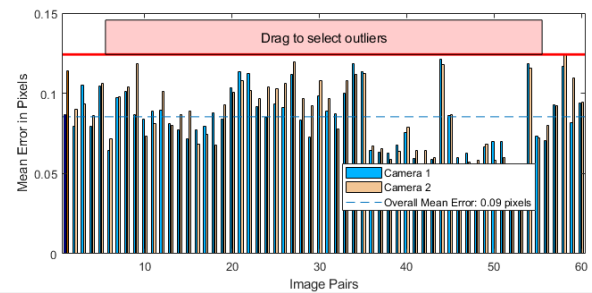


Fig. 3: Reprojection errors of all calibration image pairs

B. Image Rectification

Images taken from both the cameras simultaneously are of 480×640×3 pixel resolution with intensity of pixels as 8-bit unsigned integer data type. As two cameras can never be identical and cannot align perfectly along horizontal axis,

image rectification is required. Rectification is the projection of two or more images on the same plane so that the image lines correspond to the epipolar lines [19]. After rectification, all pixels of the right image will be in one line with the corresponding pixels of the left image. Figure 4 shows the rectified images of calibration pattern and figure 5 shows the rectified images of road.

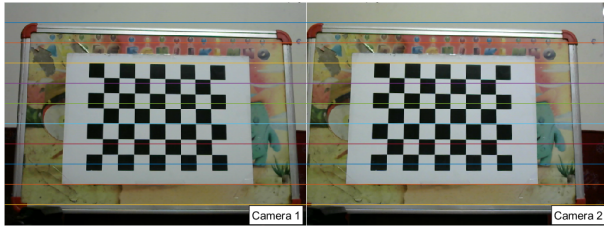


Fig. 4: Rectified images

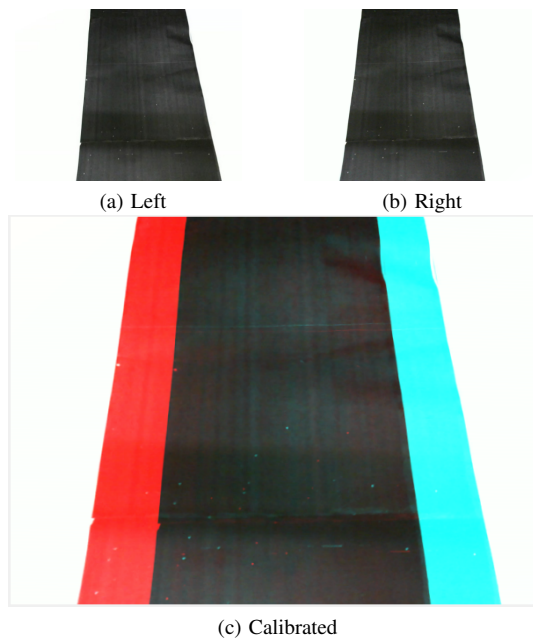


Fig. 5: Road image (a) from Left camera (b) from Right camera (c) After rectification

C. Road Segmentation

In order to determine the width of the road, it is required to get the location of road edges in 3D depth map. To achieve this, all the edge points of the road are to be extracted. By doing road segmentation, exact boundary of the road may be obtained. However, the workspace contains many objects and directly applying edge detection algorithm would give edges of every object whereas only road edges are needed. Therefore, the image is made binary and morphological operation ‘Erosion’ is applied to fill the gaps. To get the edge points of road, Sobel operation is applied along x-axis [20].

The road segmented images of both left and right camera are used for disparity map calculations. As a result, disparity map value for every pixel of edge is obtained. In figure 7, disparity map has been shown and in figure 8, corresponding 3D map of the images appears. It may be seen that 3D map for every pixel of road boundary was obtained.

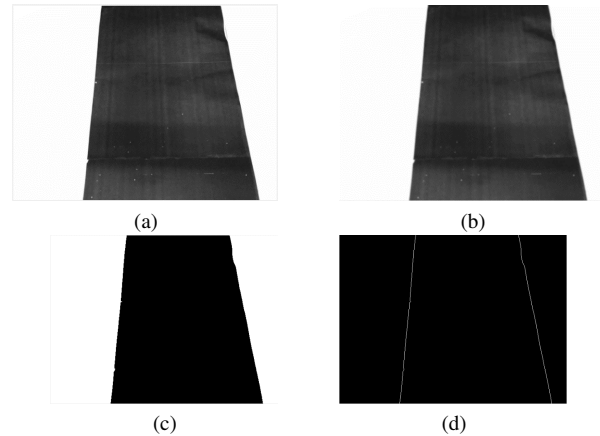


Fig. 6: Road segmentation process (a) Grayscale image, (b) Gaussian filtered image, (c) Erosion operation result on binary image and (d) Sobel filter along x-axis to extract edge points of road

D. Disparity Map Generation

As road edges are the objects of interest, getting disparity for edges is more important as compared to the rest of the region of image. Disparity map of an image is the matrix each element of which stores information about the number of pixels between the projection of an object point on the left and right cameras. The principle behind the disparity map is that for each pixel of the left image with coordinates (x_0, y_0) there is a pixel of the right image. It is assumed that the pixel of the right image will have coordinates (x_0-d, y_0) where d is the offset (disparity).

There are 2 types of approaches for finding a Disparity map - Local and Global. Local map has low computational complexity, but it is less accurate. This method calculates the disparity of each pixel individually. Global map has high computational complexity, but with low execution speed. They, unlike local methods, look for the best disparity map immediately for the entire image. Such algorithms are quite difficult to implement.

To generate the disparity map, Semi-Global Matching algorithm [21] was used in this work. This method incorporates the advantages of both local and global method.

E. Calculation of road width

The distance to the object varies inversely with the disparity. Points with larger disparity values represent points that are closer to the camera, and smaller disparities represent points

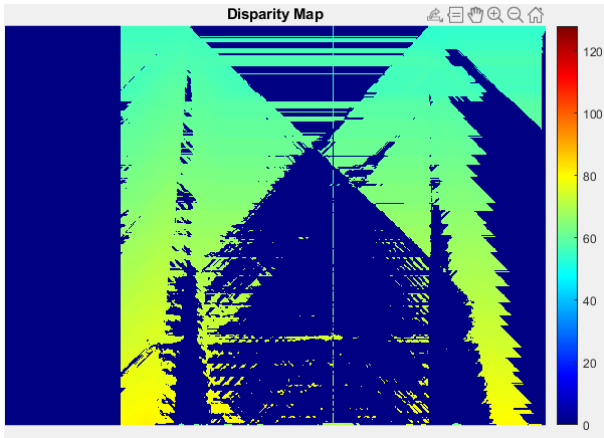


Fig. 7: Rectified images

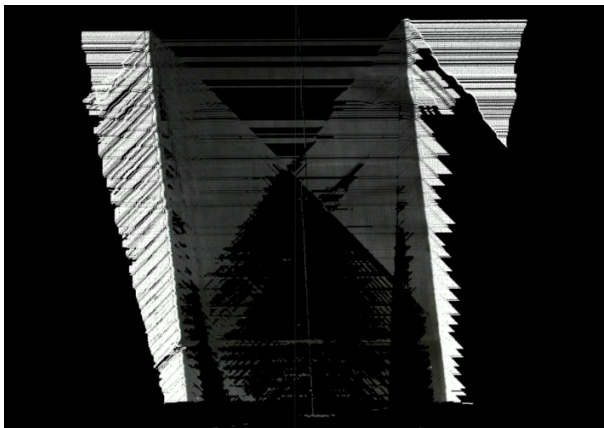


Fig. 8: Rectified images

that are farther away from the camera. The accuracy of the distance measurement depends on the resolution of the camera images, the optical properties of the lens, and the distance between the optical axes of both cameras. For distance measurement, first, 3D point cloud is constructed with the help of disparity map and stereo parameters. Then 3D Cartesian coordinates for the road boundary points are extracted from 3D point cloud of the scene. After getting the 3D Cartesian coordinates, Euclidean distance between the left boundary point P1 and right boundary point P2 is calculated from

$$Distance(P1, P2) = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)} \quad (1)$$

Here, the problem is how to decide the points of boundary between which width of the road is to be measured. For this, the whole image is divided into small segments horizontally with the size of 100 vertical pixels. In every segment, we take a point from left side and calculate the distance to every point of right boundary. The minimum distance points are decided as the corresponding boundary points.

By using Equation (1) for all left and right boundary points of selected segment calculate the distance

$$D(m, n) = Distance((X_{Left}(m), Y_{Left}(m)), (X_{Left}(n), Y_{Right}(n)), \text{ for } 0 < m < 100, 0 < n < 100 \quad (2)$$

Find the points of left and right boundary, in which the distance is minimum

$$(A, B) = Index(min(D(m, n))) \quad (3)$$

From Equation (3), we got the index value of left and right point

$$P1 = ((X_{Left}(A), Y_{Left}(A))) \quad (4)$$

$$P2 = ((X_{Right}(A), Y_{Right}(B))) \quad (5)$$

Width of Road is calculated between point P1 and P2

$$WidthofRoad = Distance(P1, P2) \quad (6)$$

When width of road is in the range of two-lane road width, then position of the center line is given by

$$(Cx, Cy) = (X_{Left}(A) - X_{Right}(B)), (Y_{Left}(A) - Y_{Right}(B)) \quad (7)$$

IV. RESULTS AND DISCUSSION

The proposed method was implemented using two Logitech C270 HD webcams. A testing setup was built for testing the performance of developed system. The results for some distance measurement using Equation (1) tests are presented with respect to actual distance. Figure 9 shows final output in which width of road, center line, 3D coordinates of left and right boundaries and boundary lines are shown.

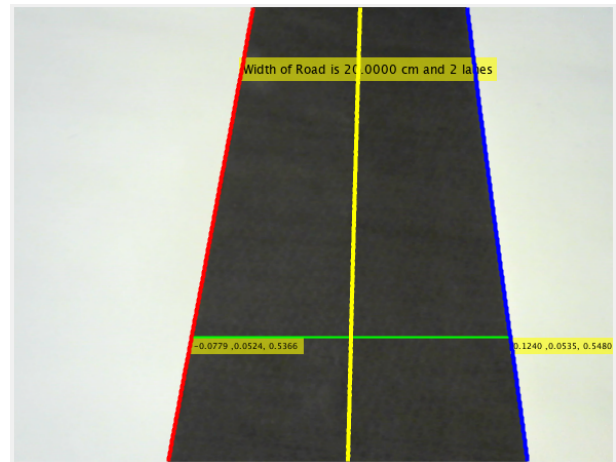


Fig. 9: Final output with road width and centerline

A. Measurement Validation

Three simulated stretches of the road having 10 cm, 15 cm and 20 cm width have been used for the analysis of the developed measurement system. The maximum width values obtained are 10.30 cm, 15.40 cm, 20.30 cm and minimum values are 9.70 cm, 14.90 cm and 19.80 cm respectively for the above stretches. Average of all obtained values are 10.00 cm, 15.00 cm, 20.00 cm respectively. Error in measurement may be calculated from:

$$Error = Measuredvalue - Actualvalue \quad (8)$$

TABLE I: Width measurement result for different road Widths

Actual width (cm)	Measured Width(cm)	Error from Average (cm)
10 cm	10.10	+0.10
	10.00	0.00
	10.00	0.00
	9.80	-0.20
	9.90	-0.10
15 cm	15.20	+0.20
	15.10	+0.10
	14.90	-0.10
	15.00	0.00
	15.20	+0.20
20 cm	19.90	-0.10
	20.00	0.00
	20.10	+0.10
	19.90	-0.10
	19.80	-0.20

TABLE II: Width measurement Deviations from Average calculated Widths

Actual width (cm)	Deviation(cm)
10 cm	Maximum 10.30
	Minimum 9.70
	Average 10.01
15 cm	Maximum 15.20
	Minimum 14.90
	Average 15.00
20 cm	Maximum 20.30
	19.80
	Average 20.00

So, for 10 cm, 15 cm, 20 cm road width, by using Equation(8) maximum positive errors are +0.3 cm, +0.4 cm, +0.3 cm and maximum negative errors are -0.3 cm, -0.1 cm, -0.20 cm respectively. From this, it can be concluded that maximum deviation in the measurement system is +0.4 cm and -0.3cm. The results show that there is ±0.4 cm error in width measurement and it remains the same for every width range from 10 cm to 20 cm. The deviation column shows the maximum or minimum values provided by the system for a particular width and the average tells the mean value of all measurement results for a particular stretch of road of fixed width. It may be observed that the average width is equal to the actual width which indicates that the deviation of the values has equal proportion in negative and positive direction.

B. Practical Scenario

For measurement validation, road was made having straight boundaries which is not practically possible. In reality, roads have irregular boundaries for the asphalt layer. A road with rough boundary was, therefore, simulated for testing. The following images depict the rough road scenario.

In Fig 10, marker position is shown on a straight road having an irregular boundary. Horizontal measuring line and a dot on the center showing the actual marker position can clearly be seen.

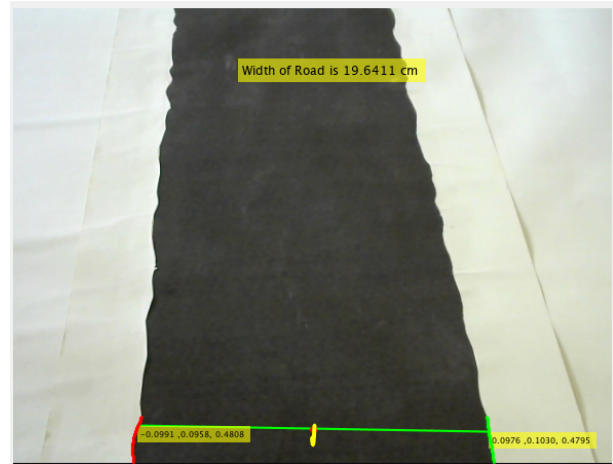


Fig. 10: Straight Road with rough boundary

In Fig 11, marking is shown as we move forward. For visualisation purposes, old markings are shown. The top line with a dot defines the current marker position.

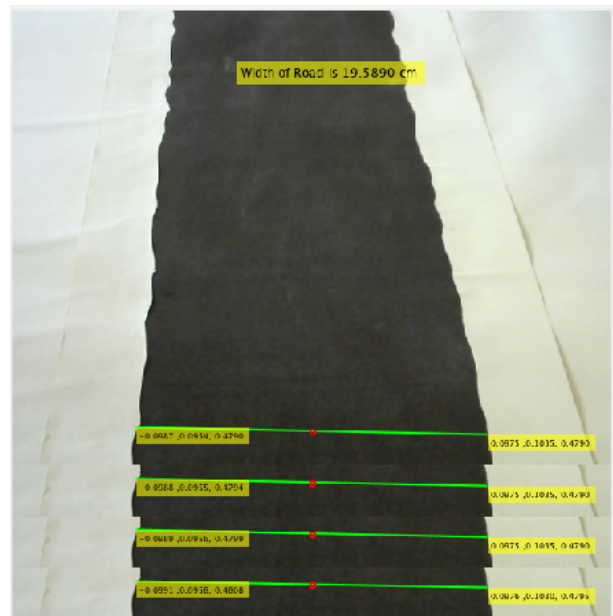


Fig. 11: Moving forward on straight road

In Fig 12 and 13, marking at curved roads are shown. In Fig 13, it can be visualized how marking is done on a curved road.

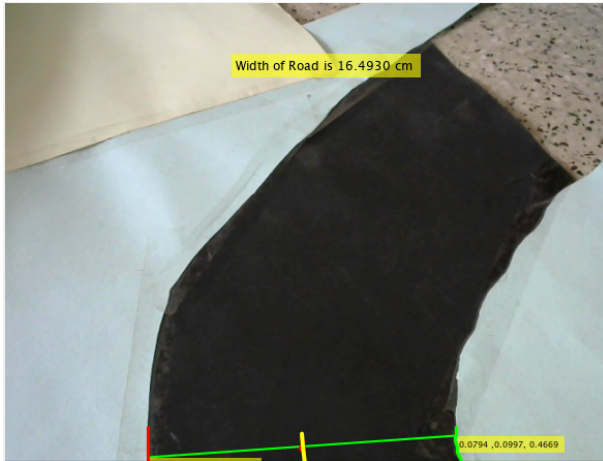


Fig. 12: Marking on curved road

C. CONCLUSION

An automatic road marking system using stereovision method is proposed. Image thresholding technique has been used for segmentation of the road from environment and for getting the road boundary points. Semi-global matching algorithm was used for disparity map generation. Euclidean distance was used to calculate the distance between two points in 3D Cartesian coordinates. The distance measurement has ± 0.4 cm as the maximum error. The designed system was tested in lab for different stretches of road on in lab generated road Environment. No need to place the camera at a fixed position and angle. Improved performance is possible through the use of more modern cameras with high resolution, or a single stereo camera. Limitations of the system: • In Real environment we have lots of colors for different object to thresholding-based segmentation can not work there, so we can use a deep learning based approach to segment the area of interest.

D. FUTURE SCOPE

This system is designed only for two lane roads. System can be developed for more numbers of lane roads by maintaining the government road authority standards for that particular place. Deep learning model can be used for better accuracy.

REFERENCES

[1] <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
 [2] J. C. Lee, J. Hwa, "Traffic lane painting device projection type of traffic lane painting" KR101036179B1, May 23, 2011.
 [3] P. A. Harrison, "Striping lay out machine" US6702516B1, Mar 09, 2004.
 [4] J. C. Schroeder, "Self-aligning mechanical pointer" US6811351B1, Nov 2, 2004.
 [5] G. D. Jurcisin, "Line striper apparatus with optical sighting means" US5302207, Apr 12, 1994.

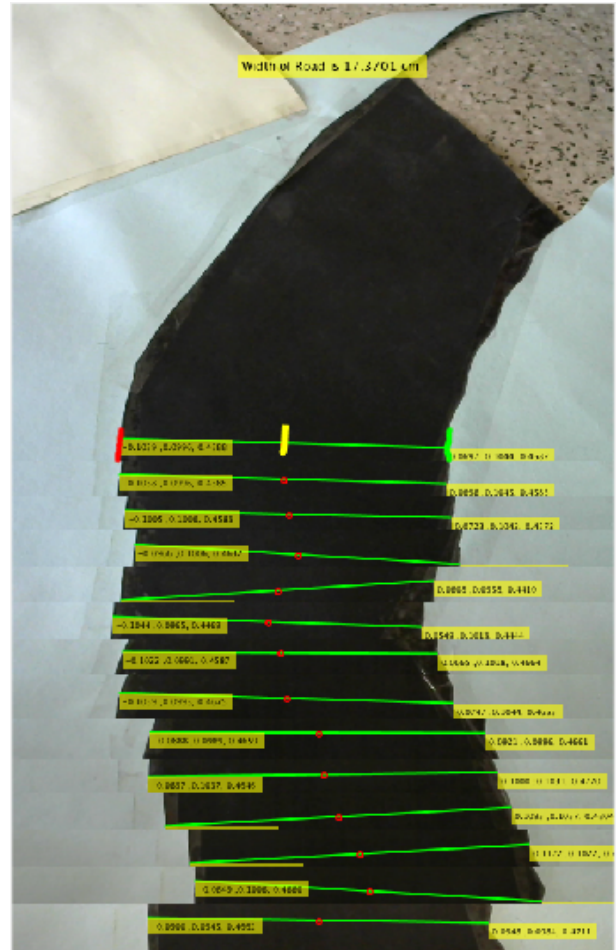


Fig. 13: Moving forward on curved road

[6] D. Charnenka, V. Arkesau, "Device for automatic re-stripping of horizontal road markings" US 10113277 B2, Oct 30, 2018.
 [7] D. D. Dolinar, W. R. Haller, "Roadway maintenance striping apparatus" US20160356005A1, Dec 8, 2016.
 [8] I. P. McGuffie, J. Nicholls, L. D. Philpotts, J. P. Ballard, R. J. Davis, S. N. R. Swatton, K. J. Palmer, A. W. Walker, "Line marking apparatus" US20090205566A1, Aug 20, 2009.
 [9] I. P. McGuffie, "Line marking apparatus, system and method" US20140106066A1, Apr 17, 2014.
 [10] R. W. Vanneman, T. A. Treichler, "Paint-stripping laser guidance system and related tech-nology" US20150330039A1, Nov 19, 2015.
 [11] J. B. S. Wilson, R. J. Milligan, A. J. Loughron, "Line marking apparatus" EP0129551B1, Feb 02, 1989.
 [12] W. H. Gmbh, "Method of re-painting road markings" DE19511893C1, May 30, 1996.
 [13] N. D. McNutt, "Line striper" US20140205744A1, July 24, 2014.
 [14] G. Ernest, L. Hamon, "Device for locating the central axis of a road" FR2609068A1, July 01, 1988.
 [15] D. D. Dolinar, W. R. Haller, M. W. Smith, C. C. Stahl, "Enhanced roadway mark locator, Inspection apparatus, and marker" US20160209511A1, Jul 21, 2016.
 [16] C. D. H. Manning, "Global positioning system controlled paint sprayer" US6074693A, June 13, 2000.
 [17] Heikkila, J, and O. Silven. "A Four-step Camera Calibration Procedure with Implicit Image Correction." IEEE International Conference on Computer Vision and Pattern Recognition, 1997.
 [18] <https://en.wikipedia.org/wiki/Reprojectionerror>.

- [19] V. Yu, V. A. Deart and A. V. Mankov i, "Pilyugin Statisticheskiye kharakteristiki trafika sovremennogo provaydera dostupa v Internet", T-Comm, 2008.
- [20] <https://www.mathworks.com/help/images/ref/edge.html>
- [21] Hirschmuller, H. "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 807-814. San Diego, CA: IEEE, 2005.
- [22] Y. Mustafah, R. Noor, H. Hasbi and A. Azma, "Stereo vision images processing for real-time object distance and size measurements", 2012 International Conference on Computer and Communication Engineering (ICCCE), 2012.
- [23] Y. Mustafah, A. Azman and M. Ani, "Object Distance and Size Measurement Using Stereo Vision System", Advanced Materials Research, vol. 622-623, pp. 1373-1377, 2012.

A Machine Learning Approach for the Early Detection of Dementia

Sven Broman
 Department of Computer Science & Physics
 Rider University
 NJ 08648, USA
 bromans@rider.edu

Elizabeth O'Hara
 Department of Computer Science & Physics
 Rider University
 NJ 08648, USA
 oharae@rider.edu

Md Liakat Ali
 Department of Computer Science & Physics
 Rider University
 NJ 08648, USA
 mdali@rider.edu

Abstract—Longer life spans in today's society have contributed to the growth of degenerative disease prevalence, especially dementia. Dementia causes a deterioration in thought process and a decline in cognitive function, specifically thinking, reasoning, and remembering. While dementia cannot be completely prevented, its early detection can delay the onset of the disease. With the help of a machine learning algorithm, relevant attributes to detect the disease in its early stages can be refined and successful predictions can be made. To conduct this analysis, the Alzheimer Features and Exploratory Data Analysis for Predicting Dementia datasets were utilized. The following machine learning models were applied to the dataset: Naïve Bayes, Decision Trees, K-Nearest Neighbors, and Fully Connected Neural Networks. After evaluation of accuracy scores, confusion matrices for both Naïve Bayes and Decision Trees were determined to provide the best results among the models. While further investigation with a larger dataset is necessary, such models suggest that machine learning algorithms are a promising tool to detect and mitigate the growth of dementia in older populations.

Keywords— *machine learning, dementia, Naïve Bayes, Decision Trees, K-Nearest Neighbors, Fully Connected Neural Network, K-fold Cross Validation*

I. INTRODUCTION

The development of advanced medical technology has shown to be successful through an increase in the quality of life and average lifespan, especially in Western countries. Longer life expectancy, consequently, contributes to the growth in senior populations and presents changing patterns in degenerative disease prevalence. In recent years, the aging society has seen a growth of older individuals with dementia, a disease that results in the loss of cognitive function when thinking, reasoning, and remembering. In 2020, the number of people worldwide living with dementia reached over 55 million. This number is expected to double every 20 years, much of the increase occurring in developing, or low and middle income, countries [1]. The effects of dementia are not experienced individually. Economic, psychological, and physical strains affect patients' families and even society [2]. An overall lack of awareness, understanding, and early prevention cause barriers to diagnosis and proper care.

Dementia leads to deterioration in thought processing beyond the predicted amount from biological aging, affecting comprehension, judgment, and orientation. There are three stages to dementia-related cognitive decline: preclinical, mild cognitive impairment (MCI), and mild through severe dementia [3]. Most studies have focused on the treatment process post diagnosis or disease onset, while current research aims to prevent development of mild through severe dementia in those with MCI. Research shows that three quarters of people with dementia haven't been diagnosed, meaning that they do not have regimented treatment and care that could be provided if a formal diagnosis was received [1]. While dementia cannot be completely prevented, studies point to several factors that can delay the onset of the disease, encouraging discovery of technology focused on early detection and diagnosis.

This research aims to contribute to existing research in risk reduction, early intervention, and timely diagnosis of dementia in older adults. The goal is to develop and refine a machine learning algorithm which identifies relevant attributes that can detect the disease in its MCI stage. To conduct this analysis, the Alzheimer Features and Exploratory Data Analysis for Predicting Dementia datasets were utilized. These datasets target variables such as age, gender, socioeconomic status, years of education, and mini-mental state examination (MMSE) scores. These features are classified into a clinical dementia rating (CDR), determining whether an individual has dementia. [4, 5]

In initial stages of analysis, data preprocessing was formed to transform raw structured data into a complete and useful format. From there, each feature was ranked based on its relevance to the class variable (whether the person has dementia) using a correlation matrix. After feature selection was completed, the dataset was split to train the machine learning algorithms and classify whether the individual had dementia. The prediction was compared to the CDR column in the dataset to determine accuracy of the methodology. A comparative analysis of supervised learning algorithms such as Naïve Bayes, Decision Trees, K-Nearest Neighbors, Fully Connected Neural Networks, and K-fold Cross Validation were conducted to determine which are the best predictors of dementia. The

influence of each feature was determined using the correlation coefficient functionality in Python, showing that the MMSE score, gender and age had the highest correlation to the CDR. Each machine learning algorithm used on the dataset was evaluated using their accuracy rate.

II. RELATED WORK

Along with the methodology proposed in this paper, various studies have been focused on the detection of dementia using machine learning algorithms. While current processes exist in clinical settings for the detection and diagnosis of dementia, the underlying goal for the proposed models are to serve as an additional decision-making aid and a way to identify incipient stages of the disease. While the same supervised machine learning algorithms are utilized, researchers have selected different features to conduct their analysis, altering the accuracy and efficiency of the models. The methodology Bansal and colleagues discuss in their article “Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia” makes use of two brain MRI data sets provided by Randy Buckner [6], from the Open Access Series of Imaging Studies (OASIS) data set [6]. Regarding preprocessing the data sets they are using, Bansal and colleagues reference other works which have dealt with incompleteness by filling in values with the averages of the entire column. Bansal’s team does likewise. Using WEKA, they make use of J48, Naïve Bayes, Random Forest, and Multilayer Perceptron to visualize data and to make their predictions. After attribute selection, only four of the nine attributes remain in their set. However, they do not mention which four remained. The results for most of the algorithms on the first data set are significantly accurate, though for Random Forest it is only 76% accurate [6]. For the second set, a Cross-Sectional OASIS, their use of Multilayer Perceptron results in the weakest accuracy. The exclusion of the attributes they use in making their predictions is an obvious drawback to Bansal and his team’s work, as it makes repetition of their results difficult.

The basis for understanding the Clinical Dementia Rating (CDR) of each of the datasets comes from Shankle and colleagues’ [7] article on detecting dementia from normal aging patterns. According to Shankle and colleagues, the main range of the CDR column is from 0-5, with 0 and 0.5 indicating normal and mildly demented, respectively. Every value above these means increasing levels of impairment. Like Bansal’s team, Shankle made use of Decision Tree learners and Naïve Bayesian classifiers. Also, like Bansal, Shankle filled in incomplete data with the mean of the column, specifically the mean of the classification accuracy. Shankle included the features age, sex, education, as well as the Functional Activities Questionnaire (FAQ). They also made use of the MMSE. According to Alagiakrishnan and colleagues, this exam has been used for several decades and features specific tasks which must be completed in a certain range and measures somewhere around six cognitive domains [8]. Although the effectiveness of this specific test is considered dubious and faced some controversy [8], it has proven to be at least somewhat efficient to the task of early dementia prediction, and therefore earns its place in these studies. They list the MMSE scores as going from 0-30, with 0 meaning severe impairment and 30 meaning none [7]. Another approach used a two-layer method, mimicking the clinical practices of early diagnosis of dementia using machine learning

techniques. So and colleagues utilized various machine learning algorithms to classify data into normal, MCI, and dementia with a two-step process [2]. The first step is composed of a screening test which classifies the data into a normal group and a cognitive decline group. The second step relies on a closer examination where cases are allocated into MCI or dementia. Feature selection was performed using chi square and information gain. Supervised machine learning algorithms used included Support Vector Machine (SVM), Naïve Bayes, Multilayer Perceptron (MLP), Bayesian Network, Bagging, Logistic Regression, and Random Forest. Multilayer Perceptron proved to have the highest accuracy for phase 1 (97.2%), followed closely by Random Forest (96.3%), and Naïve Bayes had the lowest classification accuracy (94.4%) [2]. For phase 2, SVM was the most accurate in the diagnosis of dementia (74.03%), with a close second being Logistic Regression (73.71%), with MLP being the least accurate (68.12%) [2].

While these findings were shown to be consistent with several other researchers, this method aligned closely with the three stages of dementia diagnosis: screening tests for cognitive ability, neuropsychological assessment, and diagnosis for dementia or MCI by doctor consultation. The first stage mimicked a precise screening using MMSE data. In the second stage, CERAD data is added to classify MCI versus dementia. This provides a fast way to interpret test results through classification criteria of the patient. The limitation of this model is the lack of lifestyle or disease information relating to individual patients, which could improve the accuracy of the algorithms [2]. Such data is to be utilized in this research paper.

New research has emerged which uses a combination of signal processing, machine learning, and natural language processing to diagnose dementia based on the patient’s speech. Linguistic processes such as syntax and phonetics are processes uniquely affected with disease progression. Luz and colleagues capitalized on language impairment as a clinical indicator of cognitive status, exploring patterns in dialogue of conversational data [9]. Machine learning models were trained on the data to differentiate between dementia and non-dementia speech. They hypothesize that dementia patients will show identifiable patterns in dialogue interactions via disrupted turn taking and speech rate differences. Also included in the data are demographic and clinical variables. Conversational fluidity data was gathered with speech-silence patterns and basic prosodic information.

This study used classification algorithms such as Logistic Regression, Real Adaboost, Decision Trees, Support Vector Machine, and Random Forest. Real Adaboost and Decision Trees clocked in with the highest accuracy rate at 86.5%, with a close competitor being Support Vector Machine at 83.7% [9]. While this research represents several promising linguistic parameters for the assessment of dementia, the main limitation stemmed from its small sample size of conversations and a degree of bias in the data collection [9]. While the analysis in our paper does not focus on data collected using spontaneous dialogue, future considerations can be made in the data collection process to include these predictors and yield a mostly accurate detection of dementia at the time of onset. This research paper makes use of the same two OASIS sets as Bansal, though the issue of incomplete/inconsistent data presented itself. It has

been found that somewhere around 200 values are missing on average for the EDU, SES, MMSE, and CDR columns each. This makes up around half of the total number of rows, so the usefulness of the OASIS set is dubious at best, especially given the extreme importance of the CDR column. As a result, this team will be working mostly with a different dataset titled Alzheimer Features for Analysis [4]. This set features nearly identical attributes but is more complete. However, the OASIS sets will be scrapped for complete data, and will be combined into a larger set. The remaining data (missing the CDR column) will be used as test data. As a result of surveying this other literature, this experiment will be making use of Naïve Bayes, J48, and Decision Trees, as these appear to provide the most consistently accurate results. The exclusion of the attributes Bansal and his team use is an obvious drawback to their paper, one this team will avoid by listing the features that are utilized. Preprocessing the data is the next key issue. This team will follow in Bansal and Shankle’s footsteps and fill in incomplete data with average values of columns. For the Alzheimer’s set, this was only necessary for the SES column, as other fields were filled. Two rows containing null values in the MMSE set had to be dropped, but this exclusion will not have a significant impact given the scope of the dataset.

III. DATASET AND FEATURES

This research analysis found that the CDR column is rated on scale from 0-3, and in most cases only reached 2 as a maximum. To combat this and create categorical attributes, 0 and 0.5 were defined as 0 (no dementia), and 1 and 2 were defined as 1 (dementia). Like Shankle, the MMSE score will play an important role in our predictions. A correlation coefficient test was run, revealing that the MMSE score provided the most important data in making predictions. The main dataset this paper will focus on is Alzheimer Features for Analysis [4]. Fig. 1 is an example of the dataset when initially loaded into a Pandas data frame. As can be seen visually, the data required some processing. First off, the Group column has proven to not be as helpful as it would seem. This column is directly related to the CDR column, and at first glance appears to function as the class column however, there are two reasons why this is not the case.

	Group	M/F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	Nondemented	M	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	Nondemented	M	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	Demented	M	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	Demented	M	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	Demented	M	80	12	NaN	22.0	0.5	1698	0.701	1.034
5	Nondemented	F	88	18	3.0	28.0	0.0	1215	0.710	1.444
6	Nondemented	F	90	18	3.0	27.0	0.0	1200	0.718	1.462
7	Nondemented	M	80	12	4.0	28.0	0.0	1689	0.712	1.039
8	Nondemented	M	83	12	4.0	29.0	0.5	1701	0.711	1.032
9	Nondemented	M	85	12	4.0	30.0	0.0	1699	0.705	1.033
10	Demented	M	71	16	NaN	28.0	0.5	1357	0.748	1.293
11	Demented	M	73	16	NaN	27.0	1.0	1365	0.727	1.286
12	Demented	M	75	16	NaN	27.0	1.0	1372	0.710	1.279

Fig. 1. The initial Alzheimer’s dataset

Firstly, this set has classified anyone with a CDR of 0.5 or higher as “demented.” However, according to Bansal, 0.5 only indicates possible dementia, not confirmed. Secondly, this paper will be making use of two data sets provided from OASIS, neither of which define a Group column, but feature the CDR column. It should be therefore obvious that the Group column was dropped in favor of using the CDR column as our class variable. The gender column, “M/F,” poses the obvious obstacle of string types, which interfere with the ability of Python to use correlation metrics to select attributes. A simple solution was to map the number 0 to “M” and 1 to “F.” Another issue is with the NaN values in the Socioeconomic Status (SES) column. To combat this, this team followed in suit with Bansal and Shankle and filled in the NaN values with the mean of the entire column. Although this approach could be up to debate, it was taken for two reasons. First off, this approach is recommended and often used by researchers. Second, this team does not see SES as having a profound impact on the study. As seen in Fig. 2, running a Correlation test revealed that the correlation between SES and the CDR column is minimal, and somewhere around 0.01. As a result, this team will most likely drop this column in the future, though this has not yet been decided.

The MMSE column is shown in Fig. 2 to have the most correlation with the CDR column, around 0.6. The rest show minimal correlation, which is believed to be a result of much of the data being continuous rather than discrete. Most of the work conducted thus far has been focused on the age column and turning this into discrete data. The variety of the age data is shown in Fig. 3.

```
Correlation M/F and CDR class -0.08394097143923113
Correlation Age and CDR class -0.08094166210892433
Correlation EDUC and CDR class -0.04942876281554439
Correlation SES and CDR class 0.011557674982375744
Correlation MMSE and CDR class -0.6160703837903448
```

Fig. 2. The results of the correlation coefficient test.

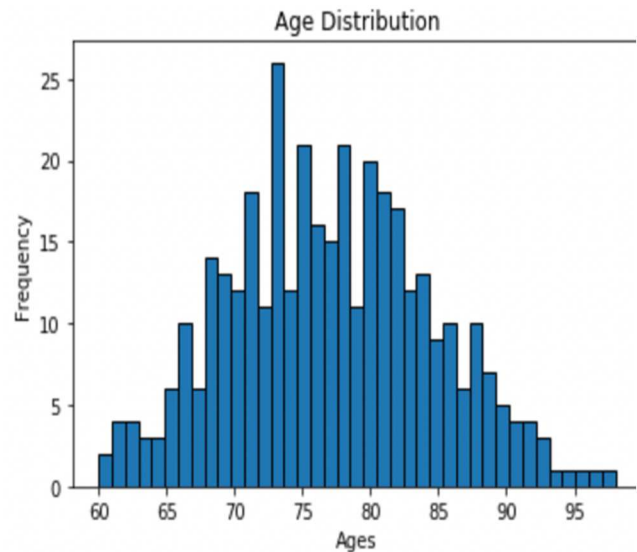


Fig. 3. The distribution of values in the Age column before grouping.

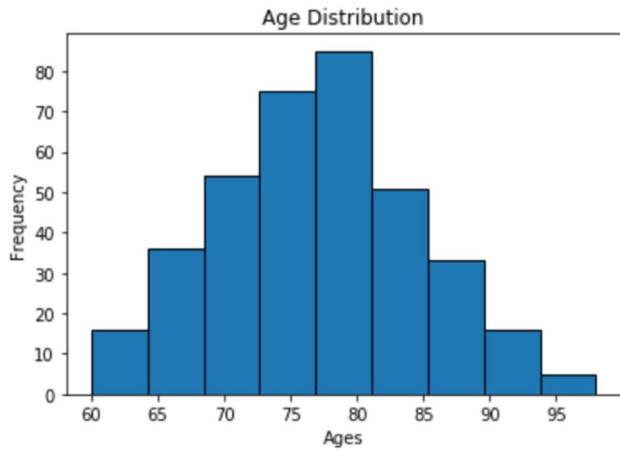


Fig. 4. The distribution of values in the Age column after grouping.

	M/F	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	0	6.0	14	2.000000	27.0	0.0	1987	0.696	0.883
1	0	6.0	14	2.000000	30.0	0.0	2004	0.681	0.876
2	0	3.0	12	2.460452	23.0	0.0	1678	0.736	1.046
3	0	4.0	12	2.460452	28.0	0.0	1738	0.713	1.010
4	0	4.0	12	2.460452	22.0	0.0	1698	0.701	1.034
5	1	6.0	18	3.000000	28.0	0.0	1215	0.710	1.444
6	1	6.0	18	3.000000	27.0	0.0	1200	0.718	1.462
7	0	4.0	12	4.000000	28.0	0.0	1689	0.712	1.039
8	0	5.0	12	4.000000	29.0	0.0	1701	0.711	1.032
9	0	5.0	12	4.000000	30.0	0.0	1699	0.705	1.033
10	0	3.0	16	2.460452	28.0	0.0	1357	0.748	1.293
11	0	3.0	16	2.460452	27.0	1.0	1365	0.727	1.286
12	0	3.0	16	2.460452	27.0	1.0	1372	0.710	1.279

Fig. 5. The dataset after preprocessing.

To make the data more discrete, and therefore make it easier to find correlation, the data was split into 9 separate bins, ages 60-65 in bin 1, 66-70 in bin 2, and so on. The resulting data distribution is shown in Fig. 4. However, when running the correlation metrics again, the correlation of the age column and the CDR column was listed as NaN. The reasons for this and the solution will be the next step of the team, as right now there is no obvious one. For some reason the data is appearing as NaN. This may be related to the methods used to split the data, in which case the solution may simply be using a different method of mapping. The data set processed up to this point is shown in Fig. 5.

To expand the scope of the algorithms, it was necessary to increase the dataset. To do this, the team combined the Alzheimer dataset with the OASIS dataset. Initially the OASIS set had around 200 rows with multiple null values (including within the CDR column), so preprocessing the data was

necessary. Dropping the null values left the OASIS set with about 200 rows. The issue with the Age column was also resolved here. The reason for the issue was that when separating the ages into bins, a few ages fell out of the assigned intervals, and therefore defaulted to null values. To mitigate this and to help with the combination of the datasets (since there is a greater range of ages in the second set), they were sorted into a greater number of bins: approximately 12. Any null values in the age column were replaced with the average of the column. The CDR column was split into the same intervals as the previous set. The 200 rows were then concatenated to the end of the Alzheimer set. The resulting set contained 600 rows, and 6 columns: Gender, Age, Education level, Socioeconomic Status, Mini-Mental State Examination, and the CDR column (the class variable). The correlation metrics are shown in Fig. 6.

IV. METHODOLOGY

The following classification techniques have been applied to the dataset: Naïve Bayes, Decision Trees, K-Nearest Neighbors, Fully Connected Neural Networks and K-fold Cross Validation.

A. Naïve Bayes

Naïve Bayes relies on the assumption of independence between attributes in the class to provide classification accuracy [10]. Put simply, the occurrence of one feature does not signify the occurrence of another, hence why it is called naive. It utilizes Bayes theorem and sample data to evaluate the probability of a class A, given B. In this dataset, the goal is to classify whether dementia is found, given the eight attributes to describe the patient’s condition. This can be shown in (1), where y is the class variable, representing if the patient has dementia given the variables X that represent the features [10].

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{1}$$

Here, X will be given as $X = (x_1, x_2, \dots, x_n)$. Expanding our equation with multiple features that are conditionally independent given the class, (2) becomes the result [10].

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)} \tag{2}$$

Due to the static nature of the denominator across entries in the dataset, the denominator can be removed. In the case of the class variable for this research, the outcome is either yes (dementia) or no (no dementia). Equation (3) shows the equation being used for this experimentation, where x^i is the value of the i^{th} attribute in x, and n is the total number of attributes [10].

$$P(y|x_i) = \sum_{i=1}^n P(x_i|y)P(y) \tag{3}$$

```
Correlation M/F and CDR class 0.0022972940774799726
Correlation Age and CDR class 0.18036141506043923
Correlation EDUC and CDR class -0.18865156350827453
Correlation SES and CDR class 0.1337286056706456
Correlation MMSE and CDR class -0.6046179611211675
```

Fig. 6. The combined dataset’s correlation.

B. Decision Trees

Decision trees consist of internal nodes that split into sub-spaces, dependent upon the values of the inputted attributes [11]. Each iteration of the decision tree considers a single attribute which is then partitioned according to its value [11]. Each leaf node of a decision tree represents the predicted class outcome variable. Within this research, navigation will happen from the root of the tree to the leaf, with the path being dependent on the values read for a person’s age, gender, and MMSE. Note that both numeric and nominal values may be used. The algorithm will predict dementia by sorting the data down the tree with corresponding values [11]

C. K-Nearest Neighbor

K-nearest neighbor classification is a simple methodology, especially used when little prior knowledge about the data’s distribution is known [12]. It looks through all examples in the data and finds the distance between one query and all the others. It performs both classification and regression. In the case of classification, which is the technique used in this research, it finds “K” number of examples like the query and votes for their most frequent label to correctly classify the outcome variable [12].

D. Fully Connected Neural Networks

Contrary to the assumptions in Naïve Bayes, Fully Connected Neural Networks do not utilize any assumptions on input, meaning that they are structure agnostic [13]. As a result, they are broadly applicable to learning any function, but can lack efficiency compared to special-purpose networks [13]. All neurons are connected to those in the previous layer, meaning the output is partly dependent on the input. While this model may be impractical for computationally demanding tasks, the classification used within the framework of this paper is acceptable.

E. K-fold Cross Validation

K-fold Cross Validation is an approach to data partitioning where the dataset is split into k folds/subsets [14-17]. The dataset is split into k equal parts, with each subset being utilized iteratively to either learn or evaluate the performance of the model [14]. One fold is used at a time to test the model, but each fold is incorporated into the testing at some point. Each iteration of cross-validation has a model trained independently than the iteration before and then the test set is validated [14]. By the end of the algorithm, each fold will have been validated and the results will be averaged.

V. EXPERIMENTAL RESULT

A. K-Nearest Neighbors, Naïve Bayes, and Decision Tree

The K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Decision Tree (DT) were used initially to test, each with encouraging results. The model was trained with 80% of the set and tested with 20%. After combining the dataset, the experiment using the same algorithms gave slightly improved results in KNN. Table 1 shows the accuracy comparison of these three models on two dataset.

B. Fully Connected Neural Network

The next step the team took in the model was to create a neural network which could be used to label the data. The architecture used to accomplish this goal was essentially four layers, including an input layer, two hidden layers, and an output layer. This was done using the Keras module in Python. Following the previous experiments, the 6 columns were passed as the input shape. Both hidden layers were defined as “Dense” to create a fully connected layer. The data was split into training and testing sets in similar ways to the earlier experiments, and then run through the model for 100 epochs. The accuracy of the model improved by nearly eleven percent by the end.

The accuracy of the model fluctuated quite a bit during the epochs, specifically with the validation data. The accuracy of the training data remained constant. This is shown in Fig. 7. The overall accuracy of the model was initially reported as around 89%. However, after only running the model several more times, the accuracy increased to and remained at 1.

TABLE I. ACCURACY COMPARISON OF THREE MODELS ON TWO DATASETS

Dataset	Train data	Test data	KNN	NB	DT
Alzheimer’s dataset	296	75	85%	100%	100%
Combined dataset	424	182	91%	100%	100%

C. K-fold Cross Validation

A final test that the team ran was performed using K-fold Cross Validation. It was decided that 10 would be the number of folds used for testing purposes. The team used K-fold Cross Validation alongside logistic regression to make their predictions. According to the results given by the Python program, the accuracy when using 10 folds was about 0.998, with a standard deviation of 0.005. Adding more folds increased the accuracy to 100%, but again, the dataset’s size makes these results tentative at best.

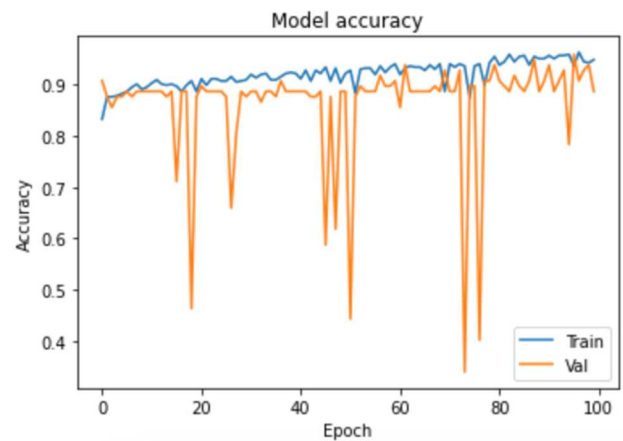


Fig. 7. Models performance in terms of accuracy

VI. CONCLUSION

This research set out to expand on existing research regarding the early detection of dementia through machine learning algorithms. The goal was to conduct a comparative analysis on various methods that best identified relevant attributes and provided the highest classification accuracy in its predictions. Both the Alzheimer Features and Exploratory Data Analysis for Predicting Dementia datasets were utilized and preprocessed to clean up all the missing or inconsistent data. The dataset was used to train the following algorithms: Naïve Bayes, Decision Trees, K-Nearest Neighbors, Fully Connected Neural Networks, and K-fold Cross Validation. Evaluation measurements included accuracy, confusion matrix, and the ROC curve. For the utilized dataset, Naïve Bayes and Decision Trees recorded the best results.

There were two major limitations of this study. Firstly, there was a lack of available data. After combining two datasets and conducting data preprocessing, the dataset only featured approximately 600 rows. Consequently, the models had less data to work with to learn and improve their accuracy scores. As with all machine learning research, the more data provided leads to improved accuracy due to a greater amount of training and testing data. The second limitation was the performance of the Fully Connected Neural Network. At times, the accuracy was extremely low, while also occasionally reaching 100%. The high accuracy scores indicate room for growth, so the focus of the team’s future research will be on increasing the power of the neural network and experimenting with deeper levels of the network. Once a larger dataset is procured, the team will look to explore the optimal number of folds with K-fold Cross Validation. Along with utilizing a larger dataset, additional machine learning models, such as Logistic Regression, Support Vector Machines, and Multilayer Perceptron, will be used to improve this research.

REFERENCES

[1] *Dementia Statistics*. Available: <https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/>.

[2] A. So, D. Hooshyar, K. W. Park, and H. S. Lim, 2017. Early diagnosis of dementia from clinical data by machine learning techniques. *Applied Sciences*, 7(7), p.651.

[3] R. S Turner, 2021. Improving Outcomes in Alzheimer’s Disease and Dementia: Emerging Treatment Advances and Recommendations. *Editorial Review Board*, 267, p.53.

[4] B. Dincer, *Alzheimer Features for Analysis*. United States: Kaggle, 2021. Accessed on: March 23, 2022. [Online]. Available: <https://www.kaggle.com/datasets/brsdincer/alzheimer-features>.

[5] M. Siddhartha, *Exploratory Data Analysis for Predicting Dementia*. United States: Kaggle, 2020. Accessed on: March 23, 2022. [Online]. Available: <https://www.kaggle.com/code/sid321axn/eda-for-predicting-dementia/data>

[6] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, 2018. Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia computer science*, 132, pp.1497-1502.

[7] W. R. Shankle, S. Mani, M. J. Pazzani, and P. Smyth, 1997, March. Detecting very early stages of dementia from normal aging with machine learning methods. In *Conference on artificial intelligence in medicine in Europe* (pp. 71-85). Springer, Berlin, Heidelberg.

[8] K. Alagiakrishnan, N. Zhao, L. Mereu, P. Senior, and A. Senthilselvan, 2013. Montreal Cognitive Assessment is superior to Standardized Mini-Mental Status Exam in detecting mild cognitive impairment in the middle-aged and elderly patients with type 2 diabetes mellitus. *BioMed Research International*, 2013.

[9] S. Luz, S. de la Fuente, and P. Albert, 2018. A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*.

[10] C. Sammut and G. I. Webb, 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

[11] O. Maimon and L. Rokach, 2005. *Data mining and knowledge discovery handbook*.

[12] L.E. Peterson, 2009. K-nearest neighbor. *Scholarpedia*, 4(2), p.1883.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, 2015, April. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). IEEE.

[14] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, 2012. The ‘K’ in K-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)* (pp. 441-446). i6doc. com publ.

[15] M.L Ali and C. C. Tappert, 2018, August. Pohmm/svm: A hybrid approach for keystroke biometric user authentication. In *2018 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 612-617). IEEE.

[16] M. L. Ali, K. Thakur, C. C. Tappert, and M. Qiu, 2016, June. Keystroke biometric user verification using Hidden Markov Model. In *2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud)* (pp. 204-209). IEEE.

[17] M. L. Ali, J. V. Monaco, C. C. Tappert, and M. Qiu, 2017. Keystroke biometric systems for user authentication. *Journal of Signal Processing Systems*, 86(2), pp.175-190

Semantic Segmentation using Modified U-Net for Autonomous Driving

T Sugirtha

Department of Computer Science and Engineering
National Institute of Technology, Tiruchirappalli
Tamilnadu, India
sugi.smile56@gmail.com

*M Sridevi

Department of Computer Science and Engineering
National Institute of Technology, Tiruchirappalli
Tamilnadu, India
msridevi@nitt.edu

Abstract—Scene understanding of urban streets is a crucial component in perception task of autonomous driving application. Semantic segmentation has been extensively used in scene understanding which further provides assistance in subsequent autonomous driving tasks like object detection, path planning and motion control. But, accurate semantic segmentation is a challenging task in computer vision. U-Net is a popular semantic segmentation network used for segmentation task. In this paper, we improve the accuracy of U-Net model by replacing its encoder part with Convolution Neural Network (CNN) architecture. We compared the performance of VGG-16 and ResNet-50 CNN architectures. Extensive analysis was performed on Cityscapes dataset and the results demonstrated U-Net with VGG-16 encoder shows better performance than ResNet-50 encoder. The model is compared with semantic segmentation CNN architectures like Fully Convolutional Network (FCN) and SegNet with mean Intersection over Union (mIoU) improved by 2%.

Index Terms—Semantic segmentation, U-Net, Autonomous Driving

I. INTRODUCTION

In recent years, autonomous driving have gained popularity with the ease of deep neural networks. An autonomous car need to sense its surrounding environment and plan its navigation without human intervention. Hence, accurate perception and decision making of the environment becomes crucial task. Scene understanding becomes the basic building block for the important modules in autonomous driving cars which comprises SLAM, path planning and motion control. Semantic segmentation assists the autonomous driving cars to perceive better understanding about the surrounding.

Scene understanding refers to identification of

lanes, maneuver, cars, pedestrians, cyclists and traffic signs. The key choice to achieve this goal is to apply semantic segmentation based on deep learning which helps to increase the detection accuracy. Semantic segmentation is a pixel level classification that assigns class label to each pixel in an image with different categories like cars, pedestrians, buildings, lanes, traffic lights, sky, trees etc.,. Though many traditional computer vision algorithms [1]–[3] were proposed for semantic segmentation, the recent progress in deep learning based approaches witnessed a large margin with CNN based semantic segmentation methods [4]–[6].

Camera sensors are most widely deployed in urban driving scenarios and traffic surveillance systems. Most of the deep learning based semantic segmentation networks proposed to process RGB images generated by monocular images mounted in autonomous driving car. Various CNN based networks are proposed for semantic segmentation, for example, U-Net, FCN, SegNet, ENet etc. In this paper, we modified the most widely used real time semantic segmentation network U-Net [5] by changing the encoder part. We applied 2 different structures to the encoder such as VGG-16 [7] and ResNet-50 [8]. Extensive analysis on Cityscapes dataset [9] demonstrates that U-Net+VGG-16 shows better performance with mIoU increased by 2%.

The major contributions of this paper is given as follows :

- Modified the encoder part of the U-Net with CNN such as architectures VGG-16 and ResNet-50.

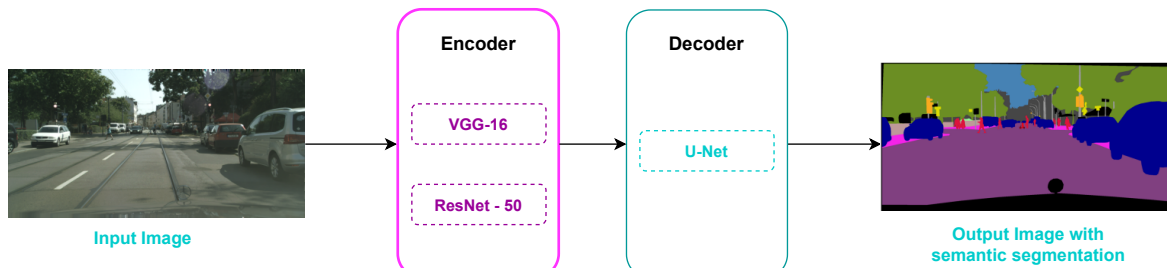


Fig. 1: Block diagram of proposed model

- Analysed the results based on the mIoU and found the accuracy increased with VGG-16 encoder.
- Extensive experiments were carried out on Cityscapes Semantic Segmentation dataset and achieved 2% improvement in mIoU.

The remaining sections of the paper is organized as follows : Section II describes the detailed explanation of various existing techniques used for semantic segmentation. Section III explains the proposed modified U-Net. The experimental results and performance analysis of the proposed method is provided in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

Computer vision achieved remarkable progress with rapid improvement in deep learning. CNN based object classification and detection methods created a huge impact in improving models accuracy and speed. Many CNN based models were proposed for object classification tasks. Example includes LeNet [10], AlexNet [11], VGGNet [7], GoogleNet [12], These methods aimed at improving accuracy by increasing the hidden layers. But, it resulted in vanishing gradient problem. Then few models like ResNet and DenseNet introduced skip connections to overcome the problem of vanishing gradient. But, it increased the number of trainable parameters which restricted the models from being deployed in real time applications. Later, InceptionNet [13], ShuffleNet [14], MobileNetV1 [15], MobileNetV2 [16] were developed with depthwise separable convolutions which made them suitable for real time deployment.

Evolution of semantic segmentation models increased by the first end-to-end FCN proposed in [4]. It is the first encoder-decoder based architecture which replaced fully connected layers with

fully convolutional layers. ESNet [17] is a symmetric encoder-decoder structure with convolution done in 3 levels : parallel convolution, dilated convolution and pointwise convolution. This resulted in reduced inference time and network complexity. ShelfNet18 [18] adopted multiple encoder-decoder structure. It used residual block which has convolution layers and imposed weight sharing among them. This decrease network parameters which in turn decreases the inference time. ERFNet [19] used residual connections while LEDNet [20] is an asymmetric encoder-decoder structure. It deployed receptive field enlargement with channel shuffling and attention pyramid network. ENet [21] applies early downsampling to decrease inference time and computation cost.

Of all the semantic segmentation models discussed above, we chose U-Net architecture and modified its encoder with CNN architectures such as VGG-16 and ResNet-50.

III. PROPOSED WORK

[5] proposed a U-shaped network with encoder decoder architecture for semantic segmentation. Figure 1 shows the block diagram of the proposed model and the algorithmic steps are given in Algorithm 1. Figure 2 shows the architecture of U-Net. The encoder part also called as contracting path to extract the features and the decoder path otherwise called as expanding path to infer class labels. The encoder and decoder are linked by means of skip connections and deploys end-to-end segmentation. The contracting path recognizes what the object is and the expanding path recognizes the pixel positions. The encoder follows typical architecture of convolution network with two convolutions and one pooling at each step. It increase the number of feature maps per layer by downsampling. The decoder comprises of one transposed convolution

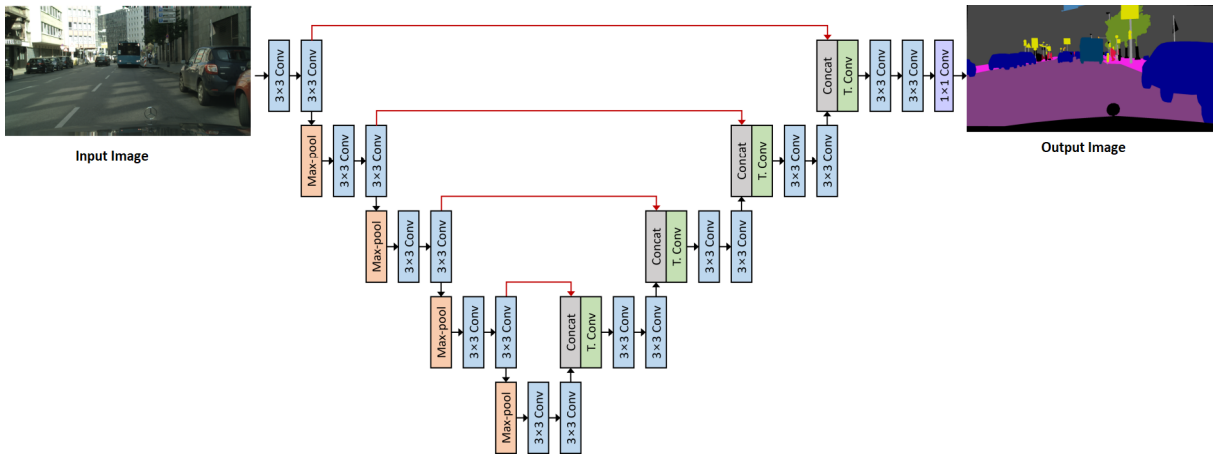


Fig. 2: U-Net architecture for semantic segmentation [5]

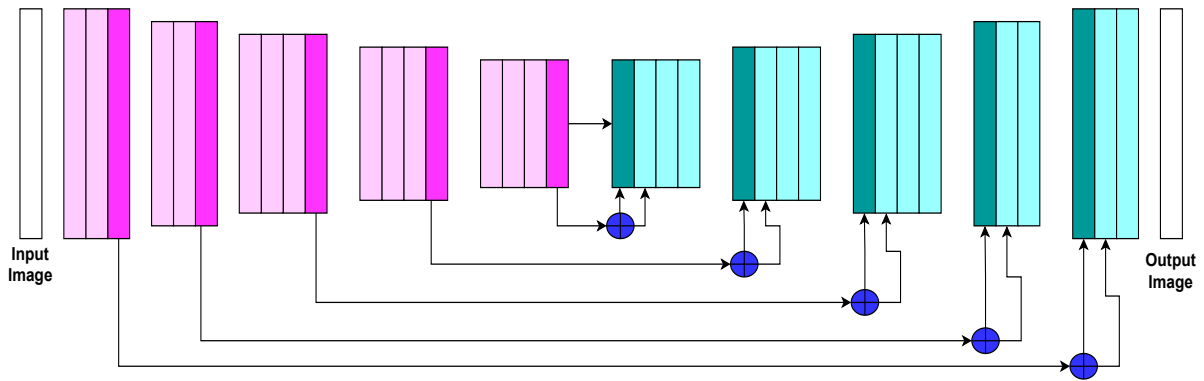


Fig. 3: VGG16 + U-Net architecture

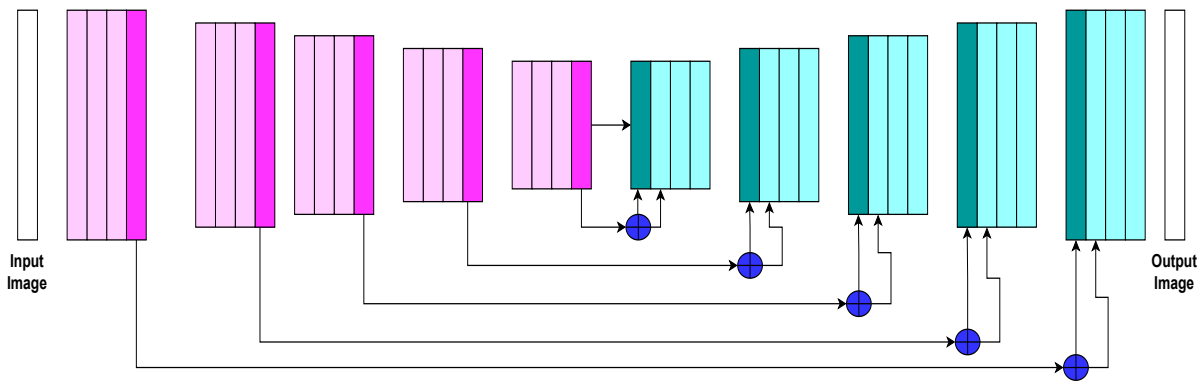


Fig. 4: ResNet-50 + U-Net architecture

and 2 convolutions which performs upsampling of feature map followed by convolution. Hence, the decoder part is meant to improve the resolution.

To infer the class labels, the upsampled features obtained in the decoder through transposed convolution concatenated with the high-resolution

TABLE I: Comparison of two proposed models with existing semantic segmentation models

	Precision	Recall	F1-Score	mIoU
U-Net [5]	0.86	0.85	0.85	78.8
Seg-Net [6]	0.84	0.81	0.82	78.6
FCN [4]	0.83	0.81	0.82	74.3
U-Net+ResNet-50 (proposed)	0.87	0.86	0.86	79.4
U-Net+VGG-16 (proposed)	0.89	0.86	0.87	80.5

features obtained in encoder part via skip connections. The architecture of U-Net was small and elegant with significantly less convolutions which makes it suitable for real time semantic segmentation. This motivated us to choose U-Net for the semantic segmentation in autonomous driving. In this work, we modify the encoder part of U-Net with CNN architectures : VGG16 and ResNet-50. VGG-16 consists of 7 sequential layers, 2 convolution layers followed by ReLU activation function and 5 max pooling layers. ResNet-50 is a CNN architecture which introduced identity mapping the idea of skipping some layers. ResNet-50 has convolution layers followed by ReLU activation function and maxpooling layers. The proposed model uses original layers from VGG-16 and ResNet-50. Upsampling is done to restore image dimensions and concatenation is performed to deploy skip connections which integrates feature maps with their class labels. Figure 3 and 4 shows the modified U-Net with VGG-16 and ResNet-50 in its encoder part.

Algorithm 1 : Modified U-Net for semantic segmentation

Input: Input image 'I'

Output: Semantic segmented image

Step 1: Extract feature maps with encoders VGG-16 and ResNet-50 for each pixel (x,y) in I

Step 2: Send extracted feature maps to decoder of U-Net

Step 3: Infer class labels

Step 4: Perform concatenation of feature maps with class labels using skip connections

Step 5: Generate output image with semantic segmentation of various categories

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Environment Setup : The model was implemented in PyTorch.

Dataset : Cityscapes dataset is most widely used large scale autonomous driving dataset which includes images from urban street scenes. The images were captured in the sheets of 50 different cities in Germany. It has 5000 and 20000 fine labelled and rough labelled images respectively. We utilized 2975, 500 and 1525 images for training, validation and testing. It comprises of 19 different classes important for autonomous driving like car, bus , train, sidewalk, cyclist, etc.

Training details : The model was trained on an NVIDIA 1080 Ti GPU for 50 epochs. We set the learning rate to 0.001 and used Adam optimizer and ReLU activation function.

Evaluation Metrics : The metrics we used to evaluate our model are :

(i) **mIoU** measures the accuracy of semantic segmentation and it denotes the average of the ratio of IoU of all classes with the ground truth and the union of ground truth with the segmentation results.

(ii) **F1-score** is the harmonic mean of precision and recall. Precision, Recall, F1 Score, mIoU are calculated using equations 1, 2, 3, 4 respectively. TP, FP and FN denotes True Positive, False Positive and False Negative.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

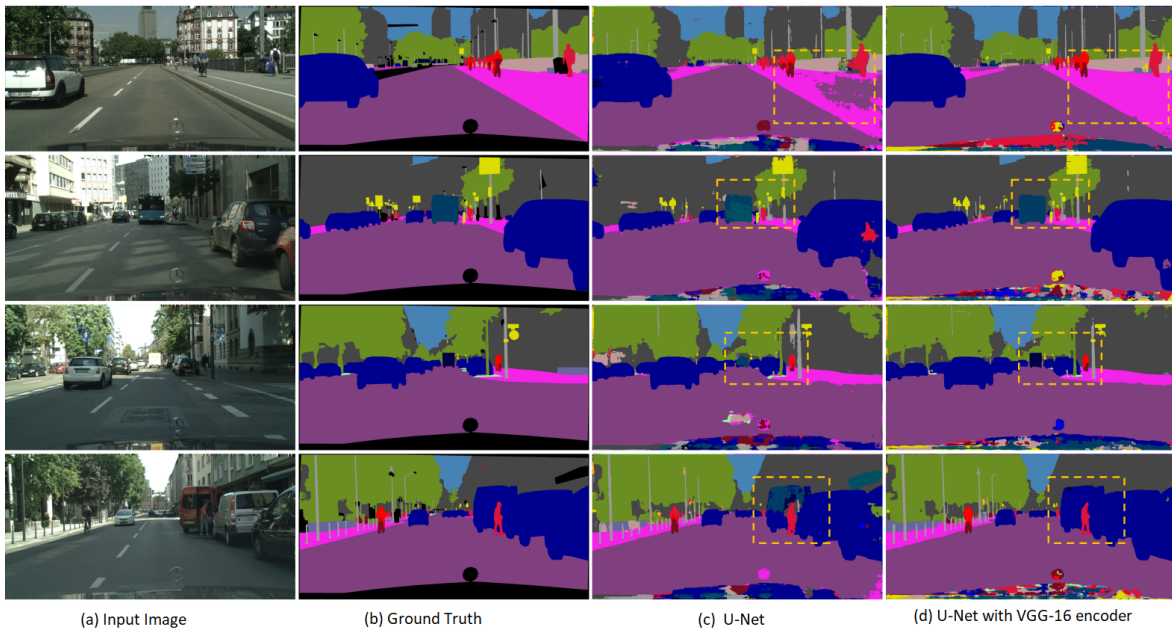


Fig. 5: Semantic Segmentation with VGG16+U-Net architecture

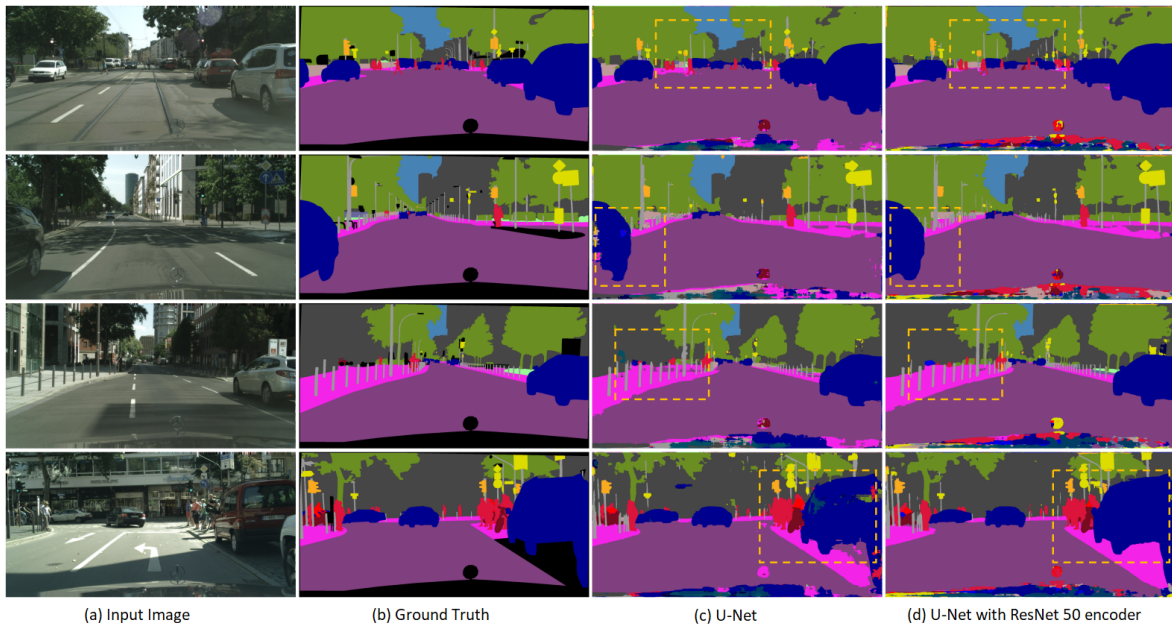


Fig. 6: Semantic Segmentation with ResNet50+U-Net architecture

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$MIoU = \frac{1}{k} \sum_{i=0}^k \frac{TP}{TP + FP + FN} \quad (4)$$

Results & Analysis : Figure 5 and Figure 6 shows the performance of modified U-Net with VGG-16 and ResNet-50 encoders. Figures (a), (b), (c), (d) shows the input image, corresponding ground truth, results of U-Net, results of U-Net with VGG-16 and ResNet-50 encoders. The dotted rectangles in Figures 5 and 6 shows the portions where VGG16+U-Net and ResNet-50+U-Net performed better than original U-Net. Table I shows the comparison among Modified U-Net with CNN architectures VGG-16 and ResNet-50 and other semantic segmentation networks such as FCN and Seg-Net. From the results, we observed that U-Net with VGG-16 encoder gives better performance than ResNet-50. It is because VGG-16 has less number of convolution layers than ResNet-50.

V. CONCLUSION

Semantic segmentation plays a major role in scene understanding which helps in the urban street perception. It is very important for an autonomous car to have complete understanding about their surrounding in order to plan their maneuver. With the improvement in deep learning models in the last decade, many CNN models were proposed for semantic segmentation. In this paper, we improved the accuracy of U-Net model by replacing the encoder with CNN architectures. The model was trained and tested on Cityscapes dataset and inferred that U-Net with VGG-16 showed better performance. In future, we would like to test with other variations of VGG-Net models like VGG-11 and VGG-19.

VI. ACKNOWLEDGEMENT

This publication is an outcome of the R&D work undertaken in the project under TiHAN Faculty Fellowship of NMICPS Technology Innovation Hub on Autonomous Navigation Foundation being implemented by Department of Science & Technology National Mission on Interdisciplinary Cyber-Physical Systems (DST NMICPS) at IIT Hyderabad.

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
 [2] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[3] S. Lakshmi and D. Sankaranarayanan, "A study of edge detection techniques for segmentation computing approaches," *International Journal of Computer Applications*, pp. 7–10, 2010.
 [4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: <http://arxiv.org/abs/1411.4038>
 [5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
 [6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
 [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.
 [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
 [9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.
 [10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
 [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
 [12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: <http://arxiv.org/abs/1409.4842>
 [13] N. S. Punn and S. Agarwal, "Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 1, feb 2020. [Online]. Available: <https://doi.org/10.1145/3376922>
 [14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *CoRR*, vol. abs/1707.01083, 2017. [Online]. Available: <http://arxiv.org/abs/1707.01083>
 [15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
 [16] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018. [Online]. Available: <http://arxiv.org/abs/1801.04381>
 [17] Y. Wang, Q. Zhou, and X. Wu, "Esnet: An efficient symmetric network for real-time semantic segmentation," in *PRCV*, 2019.
 [18] "Shelfnet for real-time semantic segmentation," *CoRR*,

- vol. abs/1811.11254, 2018, withdrawn. [Online]. Available: <http://arxiv.org/abs/1811.11254>
- [19] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.
- [20] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1860–1864, 2019.
- [21] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *ArXiv*, vol. abs/1606.02147, 2016.

Performance Analysis of a New Non-contact, Potentiometric Angle Sensor

¹Mithun Sakthivel and ¹Utpol Tarafdar

Department of Electrical Engineering, National Institute of Technology Calicut, India

Abstract – Resistive, potentiometric type of transducers find lots of applications in angle sensing due to their simple operation and linear outputs. However, they have a major disadvantage that their wiper undergoes wear and tear in course of time. So, a floating wiper is the solution to this problem. Hence in one of the previous works, a new non-contact potentiometric transducer was developed by the authors by modifying a conventional, contact-type resistive transducer. As the wiper is now floating with respect to its fixed terminals, an air gap capacitance got included in the equivalent circuit of this transducer and therefore, to get linear results from it, a novel signal conditioning circuit was developed by the authors. The principle of operation of this circuit has been discussed in a follow-on paper by the authors. This paper is a continuation of these previous works and it reports the performance analysis conducted on this sensor for a few features.

Keywords: Angle sensor, Non-contact transducer, Potentiometer, Precision, Repeatability, Sensing range, Sensitivity, Speed of operation.

I. INTRODUCTION

Resistive potentiometers or pots can be used as angle or sensors that display angle information by measuring the voltage across a movable contact that is sliding over a coiled resistive wire which is supplied with a fixed dc voltage [1]. Their characteristics, such as sensitivity, accuracy, resolution, and dependability, as well as their cost, may vary depending on the application. While the first four features should be very high for medical applications [2], angle sensors for automobiles (for steering rotation detection, throttle position detection and so on [3 - 5]) must be inexpensive and long-lasting. The key advantage of resistive potentiometric sensors is that they produce an output that varies linearly with its shaft position using simple potentiometric principle [5]. Furthermore, its straightforward operation reduces the complexity and expense of the signal conditioning circuit. Their primary downside, however, is that their moving contact or wiper is prone to wear and tear, resulting in poor performance [3]. However, by employing a wiper which does not make any physical contact with the resistive coil as shown in Fig. 1, wear and tear can be greatly reduced. So, the authors developed this non-contact type potentiometric transducer in [6] by slightly modifying a conventional contact-type pot available in the lab and studied the advantages of using it for developing an auxiliary Steering Angle Sensor (SAS) that can be utilized for sensing the front-wheel orientation of automobiles. In Fig. 1, the air gap between the resistive coil

and the wiper constitutes a coupling capacitance C whose value is in pF range, potentially affecting the transducer's sensitivity. Hence, a high frequency alternating current (ac) was utilized to stimulate it. Furthermore, to obtain linear outputs for the measured angles, a novel signal conditioning circuit was developed by the authors in [7]. Its simulation results were found to be quite good and when developed on hardware, it displayed linear outputs as expected. As this is a new sensor in development, it needs to be evaluated for some important features such as sensitivity, sensing range, repeatability, precision and speed of operation with regard to its intended application. The results obtained during the analysis of these parameters are presented in this paper. The results indicate that this new sensor has adequate sensitivity, adequate sensing range, good repeatability and good speed of operation, making it suitable for the developing the auxiliary SAS. However, its precision needs improvement.

II. PRINCIPLE OF OPERATION

Fig. 2 shows the signal conditioning circuit used to obtain angle information from the new non-contact potentiometer. The authors had explained the operation of this circuit in detail in [7]. It essentially uses a differential sensing technique to overcome the effect of C and then gather angle information in a linear fashion from its outputs. The circuit can be split into two branches: the sensing branch on the upper half and the reference branch on the lower half. Each half has two identical potentiometers with equal resistance R and inductance L . While pot-1 & pot-3 in Fig. 2 are conventional pots, pot-2 & pot-4 are non-contact pots. In each branch, the conventional pot acts as the input impedance, whereas the non-contact pot acts as the feedback impedance of an integrator. Here, pot-2 of the Integrator I-1 depicts the angle sensor with its R and L varying with respect to angle in the form of $(1 - k)R$ and $(1 - k)L$, where k denotes the proportion of the coil between the moving contact and the fixed contact 2. So I1's output will be always varying. I-2, on the other hand, provides a output that is always fixed since it always includes the total R and L of pot-4 during operation. The outputs of the two integrators are fed to an Instrumentation Amplifier (IA) which amplifies the difference between the two signals. In Fig. 2, the integrators contain a feedback resistor R_f which prevents their saturation due to opamp offset voltage. But for easy explanation of the

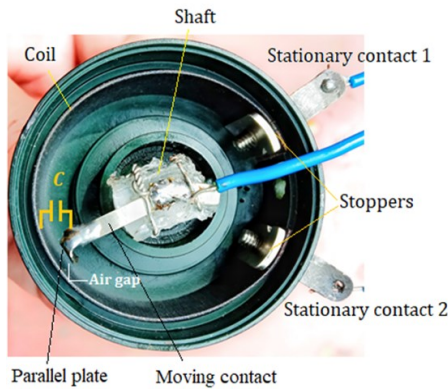


Fig. 1. The new non-contact potentiometric transducer. The air-gap capacitance is indicated by C .

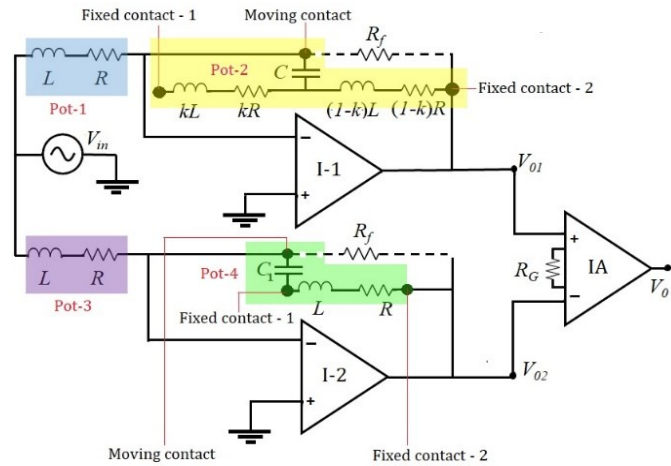


Fig. 2. The novel signal conditioning circuit developed for the non-contact transducer.

theory of operation of this circuit, the equations below have been presented without accounting these R_f .

The output of the integrator I-1

$$V_{01} = -\frac{\frac{1}{jC\omega} + (1-k)[R + jL\omega]}{R + jL\omega} \times V_{in} \quad (1)$$

This is derived from the general equation for an integrator's output. Here, V_{in} is the input voltage, and ω (or f) is the excitation frequency. Rearranging,

$$V_{01} = -\left[\frac{1}{jC\omega(R + jL\omega)} + (1-k) \right] \times V_{in} \quad (2)$$

Similarly the output from I-2 is obtained as in eq.(3). Here, C_1 is the air-gap capacitance of pot-4.

$$V_{02} = -\left[\frac{1}{jC_1\omega(R + jL\omega)} + 1 \right] \times V_{in} \quad (3)$$

The final output of this circuit which is the output of IA is obtained as eq. (4).

$$V_0 = G(V_{01} - V_{02}) \quad (4)$$

Substituting (2) and (3) in (4), the final expression of V_0 is obtained in eq. (5) when $C_1 = C$. Here G is the gain of the IA, determined by the resistor R_G .

$$V_0 = GkV_{in} \quad (5)$$

From the above eq. (5), it is clear that the new signal conditioning circuit produces linear outputs with respect to angle (indicated by the parameter k). However, when $C_1 \neq C$, the equation of V_0 is as follows:

$$V_0 = GkV_{in} + j \frac{1}{(R + jL\omega)\omega} \left[\frac{1}{C} - \frac{1}{C_1} \right] GV_{in} \quad (6)$$

In the above eq. (6), the second term is the offset voltage of the sensor. This can be reduced by the following methods:

- Replace pot-4 with a normal contact-type pot and substitute its air-gap capacitance with a physical variable capacitor whose value closely matches C .
- Raising the excitation frequency and
- Using pots with large value of R (and hence L)

Incorporating these corrective measures, the R value of the non-contact transducer finally selected for performance analysis was $50 \text{ k}\Omega$ and f was chosen as 150 kHz for most of the measurements. The results of the performance analysis are presented in the following section.

III. PERFORMANCE ANALYSIS

As it is required to evaluate the performance of any new sensor under development to establish its efficacy, the new non-contact potentiometric transducer along with its linear signal conditioning circuit was tested for the five parameters mentioned in this section. Since this angle sensor is being developed for making an auxiliary SAS for automobiles, the

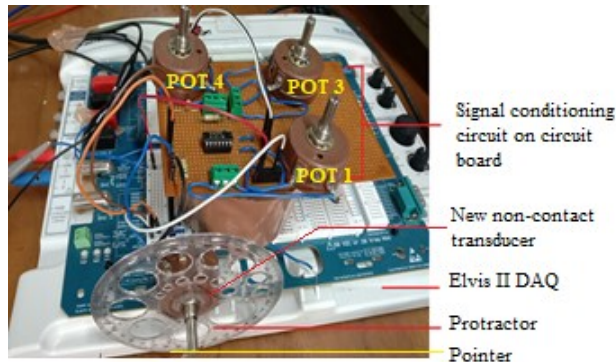


Fig. 3. Experimental setup of the transducer and the signal conditioning circuit on ELVIS II

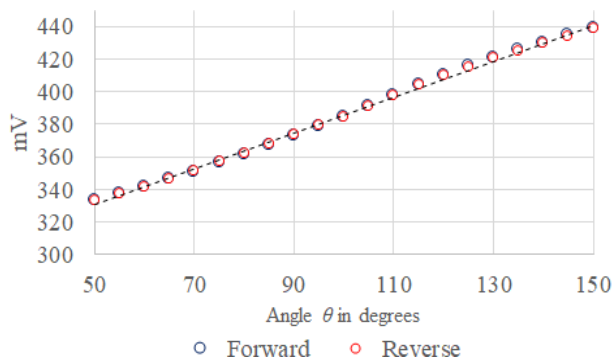


Fig. 5. Results of repeatability analysis, V_{p-p} versus θ for $G = 6.03$. Here, the forward and reverse output at each angle is the average of 100 sets of readings.

interpretation of the results obtained for each parameter is with respect to this application, that is, it will be explained if the performance of the new sensor is adequate for this application or not. The setup used for the performance analysis is shown in Fig. 3. The transducer and its circuit were soldered on a dot board and interfaced with NI ELVIS II [8] data acquisition (DAQ) board that provided the input V_{in} as well as the power supply for the various ICs in the circuit. The opamp IC and the IA used in the circuit were TL084 from Texas Instruments and AD620 from Analog Devices respectively. The voltage outputs from the IA for different angular displacements of the shaft of the angle sensor were measured either using the same ELVIS board or a separate DSO (Digital Storage Oscilloscope) depending on the frequency of excitation used. When ELVIS was used to take the reading, a LabVIEW program was used to display the IA's output on a computer screen. To identify the angular position of the shaft, a stationary protractor was placed around it as seen in Fig. 3. When the pointer connected to the shaft of the angle sensor was placed over a particular marking

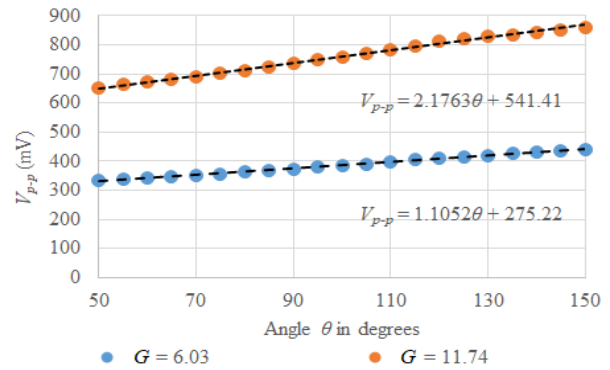


Fig. 4. Results of sensitivity analysis

TABLE I
RESULTS OF SENSITIVITY ANALYSIS

k	Angle in degrees	$G = 6.03$	$G = 11.74$	k	Angle in degrees	$G = 6.03$	$G = 11.74$
		V_{p-p} (mV)	V_{p-p} (mV)			V_{p-p} (mV)	V_{p-p} (mV)
0.83	50	334	651	0.65	105	391	771
0.82	55	338	663	0.63	110	397	784
0.8	60	342	671	0.62	115	404	797
0.78	65	346	681	0.6	120	410	812
0.77	70	351	691	0.58	125	415	819
0.75	75	357	702	0.57	130	420	829
0.73	80	362	713	0.55	135	425	835
0.72	85	368	724	0.53	140	430	842
0.7	90	373	736	0.52	145	434	852
0.68	95	379	751	0.5	150	439	859
0.67	100	385	756				

on this protractor, the corresponding angle was noted down along with the magnitude of the output waveform seen on the DSO or the computer screen. This activity was repeated one after the other for different angles within the sensing range of the sensor. The parameters that were evaluated are as follows:

a. Sensing range

Sensing range of a sensor is the range of magnitude of the measurand within which it produces effective outputs. For this angle sensor, it was taken as the angles within which it produced linear outputs as per the theory. It was determined while estimating the sensitivity of the sensor. Although, the total span of the commercially available pot that was tweaked to develop this non-contact transducer was 300 degrees, its sensing range was found to be 100 degrees within the limits of 50 degrees and 150 degrees (refer Fig. 4). The reason for this small range is the variation in the value of C along the perimeter of the transducer from fixed contact 1 to fixed contact 2. Between 0 & 50 degrees and 150 & 300 degrees, this air gap capacitance was observed

to be varying and therefore, the expected linear trend in the output with respect to angle was not observed. So, further work will be done to eliminate this issue. The reason for this variation of C is the variations in the radius of the resistive coil with respect to the axial shaft. However, for the proposed auxiliary SAS that was intended to identify the angular position of the front wheels of the vehicle during motion, this sensing range is sufficient as the front wheels themselves displace by ± 30 degrees only from the central position [9]. Using suitable conversion gears, this 100-degree operational range can be easily made to read the total of 60 degrees movement of the front wheels from left to right.

values at $G = 11.74$ are almost double to $G = 6.03$ as the selected gain values were not too high that the output voltage V_O was within the slew rate of the IA ($= 1.2 \text{ V}/\mu\text{s}$) [10] at 150 kHz. The sensitivity of the circuit for the two gain values can be identified from the trendline equations presented in Fig. 4 that has been plotted using the values in Table I. For $G = 11.74$, it is equal to $2.2 \text{ mV}/\text{degree}$ while for $G = 6.03$, it is $1.1 \text{ mV}/\text{degree}$ approximately. Although these are not too high values, an ADC like ELVIS whose resolution is in μV can easily identify the corresponding angle from the output values. So, ideally the obtained sensitivity of this new sensor is adequate for the proposed auxiliary SAS. Moreover, the sensitivity can be further improved to $3.2 \text{ mV}/\text{degree}$ by increasing G to 17 which

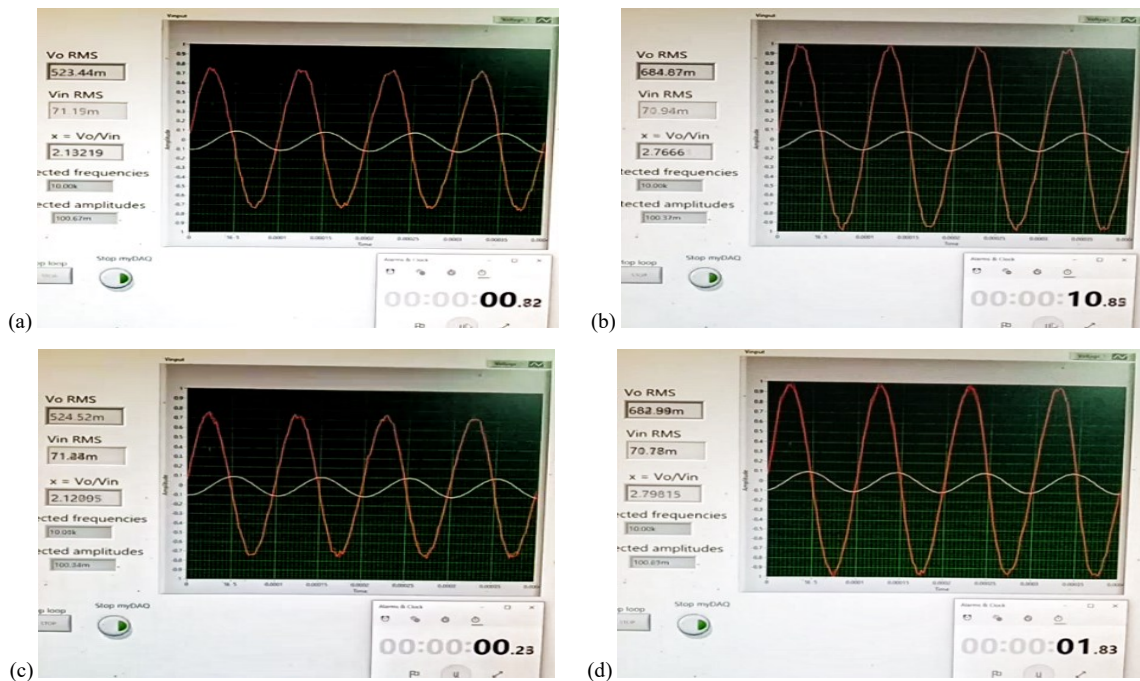


Fig. 6. Evaluating the speed of operation of the new non-contact transducer - (a) & (b) are the initial and final time stamps of the 10 s operation and (c) & (d) are the initial and final time stamps for the 1.5 s operation. In these figures, the red signal is V_O while the white is V_{in} .

b. Sensitivity

The sensitivity of a sensor is the voltage output it produces for unit change in the measurand's value, which is angle in this case. Sensitivity of this sensor (transducer + its signal conditioning circuit) basically depends on the gain of the IA used. The values of the gain resistor R_G used to obtain gains of 6.03 and 11.74 were 9.63 k Ω and 4.60 k Ω respectively. The peak-to-peak (V_{p-p}) values of V_O obtained from the DSO for $V_{in} = 75 \text{ mV}$ amplitude and 150 kHz, for various values of k are shown in Table I. In this table, it can be seen that the output

is the maximum gain possible without affecting the shape of V_O due to the slew rate of the IA when $V_{in} = 75 \text{ mV}$ at 150 kHz.

A point worth noticing in Fig. 4 is that the offset voltage of the sensor denoted by the y-intercepts of the linear trend-line equations is seemed to have reduced when compared to [7]. In [7], the V_{p-p} values of offset voltage obtained during the experimental evaluation should have been 352 mV and 685 mV respectively for $G = 6.03$ and $G = 11.74$ respectively, but these are now 275.2 mV and 541.4 mV respectively. This shows that

the remedial measures suggested earlier have been effective in reducing this nonideality.

c. Repeatability

This is one of the most important attributes of a sensor as it is supposed to produce same outputs for the same magnitude of the measurand irrespective of the direction in which the measurand is varying. The first step in determining repeatability was to set the wiper to 50 degrees and record the obtained output for this position (the features of V_{in} were kept the same as earlier). Then, for every 5 degrees, the wiper was displaced and outputs were collected up to 150 degrees. This can be considered as the forward trend of the measurand. Following this, outputs were collected in the reverse trend from 150 degrees to 50 degrees with the same spacing of 5 degrees. A total of 100 sets of such forward and reverse measurements were taken and their respective averages for each angle were calculated. These averaged values are presented graphically in Fig. 5. As the output values of both trends for each angle are exactly superimposing in this figure, the repeatability of the sensor can be claimed to be very good and therefore, it is very suitable for its intended application. However, an aspect to be noted is the slight deviation in the output values from the dotted trendline provided in this figure. This is due to variation in the value of C along the perimeter of the transducer that was mentioned earlier. Although not easily seen, this aspect is present in Fig. 4 also. In fact, Fig. 4 and Fig. 5 are the same for $G = 6.03$, but in Fig. 4, an additional set of readings has been taken for $G = 11.74$ as the objective there was to identify the best change in output for unit angular displacement in degrees.

d. Precision

Precision is the consistency with which a sensor produces its output for a given magnitude of the measurand. The precision of the new non-contact sensor was also studied while performing the repeatability analysis. The maximum, average and minimum V_{p-p} values obtained for each angle taken in Table I in a set of 200 measurements while $G = 6.03$ are shown in Table II. In this table, the equation for the mean absolute deviation \bar{u} =

$$\bar{u} = \frac{\sum_{i=1}^{200} |V_{p-p-i} - \overline{V_{p-p}}|}{200} \quad (7)$$

Here, V_{p-p-i} is the V_{p-p} value obtained for the i^{th} V_O measurement corresponding to an angle and $\overline{V_{p-p}}$ is the mean V_{p-p} value of these 200 measurements. It can be seen that, although the deviations of the extreme values from $\overline{V_{p-p}}$ are only a few mV, these cannot be considered negligible as the mean absolute deviation \bar{u} is comparable to the difference in

$\overline{V_{p-p}}$ for every 5-degree angle evaluated. This will affect the ability of the sensor to distinguish or resolve angle changes less than 5 degrees which means, the proposed SAS will not be able to sense small changes in front-wheel orientations. Hence, the precision of the sensor needs to be improved for the proposed SAS. The possible reasons for these V_{p-p} variations are ambient noise, quantization noise of the ADC of the DSO (The DSO, Tektronix Model No. 2014C will have a quantization error of 1.56 mV as its 8-bit ADC [11] was operating on a scale of 100 mV/division) and the human error in placing the pointer correctly over a particular angle on the protractor. So, work will be done to improve the precision of the sensor. Although, both repeatability and precision analysis used the same set of (two hundred) V_{p-p} readings, the results of repeatability analysis were presented using $\overline{V_{p-p}}$ values alone. The effects of all the above reasons were minimalized in it due to averaging. Hence $\overline{V_{p-p}}$ values are the most accurate outputs corresponding to the angles evaluated and this is also the reason for the high repeatability obtained earlier. In fact, averaging a good technique to improve the accuracy of the sensor. But for this, the DAQ system has to operate at very high speed as steering angle will change continuously in real-time.

e. Speed of operation

As the intended application of the proposed non-contact transducer was steering angle sensing, the expected speed of operation of the sensor will be in the order of seconds as steering wheels are not abruptly rotated by the driver under normal conditions. Therefore, to investigate if the sensor's output would vary proportionally with angle at these ranges of speed, it was interfaced with a LabVIEW program through NI ELVIS II. The graphical interface of this LabVIEW program would indicate the RMS value of V_O along with a time stamp. Due to technical limitations of ELVIS, the excitation frequency was reduced to 10 kHz and the amplitude of V_{in} was increased to 125 mV. The idea of this test is as follows: If the sensor is able to produce the same outputs that it produced at a lower speed when displaced through its effective operating range at a faster pace later, then it is deemed suitable for the intended application. So initially, its shaft was displaced from 50 degrees to 150 degrees in about 10 s and the outputs (RMS values) corresponding to these limiting values of angles were recorded. These were found to be 524.5 mV and 683.0 mV respectively. Later, the same angular displacement was made in about 1.5 s and the corresponding RMS values were noted. In this attempt, the obtained values were 523.5 mV and 684.7 mV respectively. Similar results were obtained in another 20 more following attempts. As the output values obtained for 10 s and 1.5 s evaluations were very close to each other, it can be said that the

new non-contact sensor is consistent at higher speeds of operation as well. The screenshots of the time stamps of 10 s operation and 1.5 s operation are presented in Fig. 6. In this figure, 6a & 6b show the time stamps corresponding to the initial angular position (50 degrees) & final angular position (150 degrees) of 10 s operation respectively, while 6c & 6d show the corresponding values of 1.5 s operation respectively. As 1.5 s is the quickest time amongst the two, it can be said that the best angular velocity displayed by this transducer is about 67 degrees/s. An angular displacement of 67 degrees in 1 s is a pretty high value for the front-wheels of an automobile and will occur rarely only during an accident. Hence, it can be confirmed that the new sensor has good speed of operation, sufficient enough for its intended application.

Summarizing, the performance analysis of the new non-contact potentiometric angle sensor revealed the following – while it has good repeatability & speed of operation, its sensing range & sensitivity are adequate for steering angle sensing; its precision however needs improvement. So, further work has to be done for improving it. Moreover, if this sensor needs to be used for other applications that require sensing a wider range of angles, then additional work needs to be done for improving this feature too.

V. CONCLUSION

This paper presents the results of performance analysis of a new non-contact potentiometric angle transducer and its signal conditioning circuit designed for the purpose of developing an auxiliary SAS for automobiles. The features that were analyzed were sensing range, sensitivity, repeatability, precision and speed of operation. The sensing range of this angle sensor was observed to be 100 degrees which is adequate for the intended application. But, it can be improved further by minimalizing the effect of its air-gap capacitance if this sensor needs to be used for other applications that require sensing a wider range of angles. Its sensitivity with an amplifier gain of 12 was found to be around 2.2 mV/degree which is also adequate for SAS. From the mean V_{p-p} values obtained from sets of 200 measurements corresponding to different angles, it was identified that this sensor has very good repeatability for both ascending and descending trends of angle. Its speed of operation was found to be about 67 degrees/s, which is also good enough for the intended application. However, due to various reasons, the precision of this sensor was found inadequate. So, further work needs to be done to improve this. Thereafter, the improved sensor will be mounted on the steering column of a test vehicle in the lab and its suitability will be checked in real-time.

TABLE II
RESULTS OF PRECISION ANALYSIS

Angle (degrees)	Max V_{p-p} (mV)	$\overline{V_{p-p}}$ (mV)	Min V_{p-p} (mV)	\bar{u} (mV)	Difference in V_{p-p} for every 5 degrees, \bar{v} (mV)
50	336	333.6	328	2.1	- -
55	344	337.5	332	2.7	3.9
60	348	341.6	334	2.7	4.1
65	352	346.3	340	2.8	4.7
70	356	351.3	344	2.8	5.0
75	364	356.5	348	2.7	5.2
80	370	362.0	352	2.8	5.5
85	376	367.5	358	3.0	5.5
90	382	373.4	364	3.2	5.9
95	388	379.0	368	3.8	5.6
100	396	384.7	374	4.0	5.6
105	400	390.8	382	4.2	6.1
110	404	397.4	388	3.3	6.7
115	410	404.2	396	3.0	6.8
120	416	410.2	404	2.4	6.0
125	420	415.4	410	2.3	5.2
130	426	420.4	416	2.4	5.0
135	430	425.2	420	2.9	4.9
140	434	430.0	424	2.4	4.8
145	438	434.3	428	2.3	4.4
150	440	439.0	436	1.5	4.6

REFERENCES

- [1] (2020, Jul 10). Ixthus instrumentation, “Position sensor types” [Online]. Available: <https://www.ixthus.co.uk/blog/detail.php?aid=68&did=Position-Sensor-Types>
- [2] (2012, Jun 1). Piher Sensors and Controls S.A., Tudela, Spain. Rotary Sensing Technologies for Medical and Robotic Shaft Angle Sensing Applications [Online]. Available: <https://www.medicaldesignbriefs.com/component/content/article/mdb/tech-briefs/13928>
- [3] J. Svensson, “Design of a portable steering wheel angle measurement system,” M.S. thesis, Vehicle Dynamics, Aeronautical and Vehicle Engineering, Royal Institute of Technology, Sweden, 2015. Accessed on: Nov. 25, 2021. [Online]. Available: <https://www.semanticscholar.org/paper/Design-of-a-portable-steering-wheel-angle-system-Svensson/e64995a436c6644571cea06abe89d076-873686bb>
- [4] W. J. Fleming, “Overview of automotive sensors,” *IEEE Sensors J.*, vol. 1, no. 4, pp. 296–308, Dec. 2001
- [5] (2021, Dec. 12). Variohm Eurosensor. VTP11 Throttle Position Sensor [Online] Available:

<https://www.variohm.com/products/motorsports-sensors/rotary-position-sensors-for-motorsport/vtp11-throttle-position-sensor>

- [6] Mithun Sakthivel, Utpol Tarafdar and Hemanth Sankar P., "A New Auxiliary Steering Angle Sensor for Power-steering in Four-Wheelers," *IEEE CATCON*, Dec, 2021.
- [7] Mithun Sakthivel, Hemanth Sankar P., Utpol Tarafdar and Anoop C. S., "A Simple Linear Circuit for Angle Measurements Using a Non-Contact Potentiometer," *IEEE DELCON*, Jan., 2022.
- [8] National Instruments. *NI ELVIS II datasheet*, [Online]. Available:
www.ni.com/pdf/products/us/cat_nielvisii_plus.pdf
- [9] (2021, Nov. 25). How does a steering angle sensor work [Online]. Available: <https://findanyanswer.com/how-does-a-steering-angle-sensor-work>
- [10] Analog Devices. *AD620, low power instrumentation amplifier datasheet*, [Online]. Available:
<https://datasheetspdf.com/pdf-file/819392/AnalogDevices/AD620/1>
- [11] Tektronix. *DSO Model No. 2014C datasheet*, [Online]. Available:
<https://download.tek.com/manual/077044600web.pdf>

Identifying Functional and Non-functional Software Requirements From User App Reviews

Dev Dave¹, Vaibhav Anu²
 Department of Computer Science
 Montclair State University
 Montclair, NJ, USA
 {daved2¹ | anuv²}@montclair.edu

Abstract—Mobile app developers are always looking for ways to use the reviews (provided by their app’s users) to improve their application (e.g., adding a new functionality in the app that a user mentioned in their review). Usually, there are thousands of user reviews that are available for each mobile app and isolating software requirements manually from such a big dataset can be difficult and time-consuming. The primary objective of the current research is to automate the process of extracting functional requirements and filtering out non-requirements from user app reviews to help app developers better meet the wants and needs of their users. This paper proposes and evaluates machine learning based models to identify and classify software requirements from both, formal Software Requirements Specifications (SRS) documents and Mobile App Reviews (written by users) using machine learning (ML) algorithms combined with natural language processing (NLP) techniques. Initial evaluation of our ML-based models show that they can help classify user app reviews and software requirements as Functional Requirements (FR), Non-Functional Requirements (NFR), or Non-Requirements (NR).

Keywords— requirements, mining, classification, machine learning, natural language processing

I. INTRODUCTION

The requirements phase in the software development life cycle (SDLC) is one of the most important stages of software development. Incorrect or missing requirements can lead to an incomplete product that does not satisfy customer demand. Requirements are collected from various stakeholders (end-users, client, etc.) and recorded in a formal document called the Software Requirements Specification (SRS). The quality of the SRS document has an immense impact on the quality of the final software product. The SRS document outlines the functional and non-functional capabilities of the software-being-built. Thus, the development team, the end-users, and the client must share the same understanding [1]. The industry is still seeking to establish and apply good practices to identify and classify key software requirements that are often missed by software engineers during the initial releases of the software. Capturing such missing requirements is a key requirement for successful future releases of the software product.

In recent years, the growth of mobile devices has led to an increase in mobile software. App distribution platforms such as the Google Play Store and Apple’s App Store have had 4 million apps as of June 2016 with the number of monthly app downloads hovering around 1 billion per month [2]. Users that download these applications can rate the application and provide textual

feedback relating to the application. User reviews contain important information that can assist a developer in better understanding their users’ needs. Making use of user reviews for app upgrades can lead to an increase in new users and retain existing users. Studies indicated that one-third of users have modified their app ratings after a developer’s response. Mobile applications can receive more than 20 reviews per day and popular mobile applications such as Facebook can receive more than 4000 application user reviews a day [2]. Therefore, user reviews are an important source of feedback for developers to elicit (i.e., identify) requirements so as to provide fixes or updates in their apps. However, it is difficult to sift through vast amounts of user reviews and filter out reviews that do not contain a software requirement. Automatic extraction of software requirements through various approaches and frameworks can enable developers to respond quickly to customer needs and reduce the time spent on eliciting requirements. This paper introduces a *novel hybrid dataset* consisting of Functional Requirements (FR), Non-Functional Requirements (NFR), and Non-Requirements (NR) from Software Requirements Specifications (SRS) document and user app reviews. The hybrid dataset is used to create data models that use machine learning (ML) algorithms such as Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Random Forest (RF) combined with natural language processing (NLP) techniques such as term frequency-inverse document frequency (TF-IDF). The data models are then evaluated by using 10 k-fold cross-validation and calculating accuracy metrics such as recall, precision, and the F1 score. Our research addresses the following research questions (RQs):

RQ1: *What is the type and size of data that is required to automate the identification of requirements from user app reviews?*

To answer this research question, we created a novel hybrid dataset. The hybrid dataset consists of formal software requirements from SRS documents from the PROMISE Software Engineering Repository and user app reviews from the dataset prepared by Maleej et. al [3]. The hybrid dataset consists of a total of 1,249 requirements and user app reviews. 507 requirements labeled as functional requirements, 371 labeled as non-functional requirements, and 371 labeled non-requirements.

RQ2: *To what extent can the Machine Learning Algorithms combined with NLP techniques accurately identify and classify Functional and Non-Functional Requirements?*

The results of all three data models using SVM, SGD, and RF combined with TF-IDF are evaluated by using 10 k-fold cross-validations and the performance of each model is evaluated by metrics such as accuracy, precision, recall, and F1 score for Functional and Non-Functional Requirements.

RQ3: *How effective are the data models in isolating Non-requirements from Functional or Non-Functional Requirements?*

To evaluate the effectiveness of the data models in identifying Non-Requirements, the precision, recall and F1 scores are calculated for user app reviews classified as Non-Requirements.

The primary contributions of this paper are:

1. Introducing a novel hybrid dataset with formal software requirements from SRS documents and user app reviews to identify functional and non-functional requirements as well as non-requirements that are unimportant.
2. Introducing and evaluating three data models using various ML algorithms with TF-IDF for the automatic identification and classification of software requirements and user app reviews.

II. BACKGROUND

Identifying and classifying functional requirements from SRS documents and User App reviews is an open research problem. Data related to requirements elicitation may not be well documented by the stakeholders and it is up to the developer to meet the requirements as per their understanding. Ambiguous requirements are a key factor in the failure of software projects. A lot of requirements are not initially captured but are captured during testing by the end-user or client to accept the software project before it is moved on to production [1].

In the rest of this section, we first describe FRs and NFRs, followed by a discussion on existing research on classification of software requirements.

A. Functional Requirements (FRs) and Non-functional Requirements (NFRs)

Software requirements are of two types: Functional requirements (FR) and Non-Functional requirements (NFR). For the purpose of our research, we propose a third type of requirement: a *Non-Requirement (NR)*. A non-requirement (NR) is a statement that contains no information whatsoever about software's features or qualities and thus has no impact on software development.

A functional requirement (FR) describes the required behavior of the software-being-built in terms of required activities, such as reactions to inputs, and the state of each entity before and after an activity occurs. FRs should describe the following [3]: what will the software system do; are there several modes of operation; What kinds of computations or data transformations must be performed; what are the appropriate reactions to possible stimuli?

A Non-Functional requirement describes some quality characteristics that the software solution must possess. It

describes the following [3]: performance; security; reliability and availability; maintainability; usability and human factors.

Thus, it is vital to identify and classify software requirements during the requirements engineering (RE) phase to ensure that the software development project will meet client requirements, is within the budget, and is completed on time.

B. Existing Research on Automatic Identification and Classification of Requirements from Formal Requirements Documents

This section explores the various approaches and tools that have been used to identify and classify software requirements from formal requirements documents (usually referred to as SRS documents).

Machine Learning Based Method: Binkhonain and Zhao [4] give an overview of various machine learning algorithms implemented and their performance to identify and classify non-functional requirements. The performance of 16 machine learning algorithms is evaluated. The 16 algorithms are divided into the following types of algorithms: 4 unsupervised, 5 supervised, and 5 semi-supervised machine learning algorithms [4]. The 4 unsupervised algorithms used are Latent Dirichlet Allocation (LDA), K-means, Hierarchical Agglomerative, and Biterm Topic Modelling (BTM). The five supervised machine learning algorithms used are Support Vector Machines (SVMs), Naïve Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (K-NN), and Multinomial Naïve Bayes (MNB). Lastly, the Semi-Supervised algorithms used are Expectation-Maximization (EM), Self-training, Active learning, Random Subspace Method for Co-training (RAS-CO), and Relevant Random Subspace Method for Co-training (Rel-RASCO). The machine learning approaches had the same data preprocessing steps which included selecting appropriate features and preprocessing the text to be classified. The text was preprocessed by using methods such as stop word removal, stemming, and tokenization. To select the appropriate features, the text was converted into a numeric matrix using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). The important features are evaluated by using information gain and the Chi-squared test. After preprocessing the text, the machine learning algorithms are implemented and evaluated by metrics such as accuracy, precision, recall, and F1 scores. Overall, ML algorithms achieve an accuracy score of more than 70% when classifying NFRs [4]. Supervised ML algorithms perform better than Unsupervised and Semi-supervised algorithms. SVM and NB algorithms achieved the best performance.

Data Visualization based Method: InfoVis is a tool proposed to identify requirements based on how ambiguous or incomplete a requirement is [5]. The tool combines data visualization and NLP methods to identify requirements. A novel algorithm, Semantic Folding Theory (SFT), takes in user input and then calculates the similarity between two words [5]. The ambiguity score is calculated based on the term and context similarity. Based on the scores, InfoVis generates Venn diagrams to easily visualize and explore requirements.

Template Based Method: An approach to identify security requirements involves the use of templates. Riaz et al. propose

the use of templates that automatically suggest security requirements to aid in the process of requirements gathering [6]. The template serves as a list of important security requirements to be included for the developers as they are eliciting the requirements. The template receives requirements as the input and based on the requirements, it creates a list of security requirements [6]. The template can serve as an assistive tool to be used with other tools to primarily focus on eliciting security requirements and addressing privacy concerns.

Furthermore, there are tools like the PROMIRAR tool that utilizes previous software requirements to create new requirements based on analogy reasoning [7]. The tool uses NLP techniques to identify hidden requirements.

C. Existing Research on Automatic Identification and Classification of Requirements from User App reviews

Lu and Liang et al.'s Approach: Lu and Liang propose four classifications techniques to classify functional requirements and non-functional requirements (NFRs). NFRs are classified into 4 categories which are reliability, usability portability, and performance. Four classification techniques Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Chi-Squared (Chi²) and Augmented User Reviews Bag-of-Words (AUR-BoW) are used in conjunction with three machine learning algorithms Naïve Bayes, J48, and Bagging to classify user reviews [8]. The results of each classification technique are evaluated by the F-measure for each experiment. The classification techniques are defined as follows:

- **Bag-of-Words (BoW):** BoW is a model used to represent textual documents in information retrieval systems. It uses a dictionary that consists of all unique terms of user reviews and uses term frequency (TF) and the number of times a phrase appears in a user review as the weight of the textual features. This is used to train the classifiers.
- **Term Frequency-Inverse Document (TF-IDF):** TF-IDF is like BoW but also uses inverse document frequency (IDF) as the weight of the textual features.
- **CHI2:** A statistical test that is used to select important textual features from user reviews that are acquired from BoW.
- **Augmented User Reviews Bag-of-Words (AUR-BoW):** User reviews are often short and split into sentences that can make classification difficult. AUR-BoW augments short user reviews by the most similar words and uses them as input for BoW.

To evaluate the trained classifiers, the following metrics are used: precision, recall, and F-measure. 10-fold cross-validation was performed on the dataset with each fold consisting of 400 user reviews [8]. The AUR-BoW classifier resulted in the highest F-measure score of 71.7%, a precision score of 71.4%, and a recall score of 72.3%. Augmenting user reviews leads to better results when classifying user reviews.

Chen et al.'s Requirements Mining Framework: Chen et al. proposed a requirements mining framework for mobile app upgrades. A new ranking model is developed to classify customer requirements and rank the importance of each

requirement. The effectiveness of the framework is then evaluated by product quality improvements [9]. The proposed mining framework seeks to transform user reviews into product upgrade requirements. The mining framework has four main components which are context-aware segmentation, opinion target extraction, opinion target grouping, and requirements summarization. To evaluate the effectiveness of the framework, an empirical analysis is conducted to evaluate the effect of requirements mined from customer reviews on app upgrades.

Yang and Liang's Information Retrieval and NLP approach: Yang and Liang propose an approach that utilizes information retrieval and NLP techniques to automatically identify and classify software requirements into functional and non-functional requirements. The approach consists of two components which are User Reviews Extractor and Requirements Identifier and Classifier [10]. The User Reviews Extractor uses an API to collect user reviews of iBooks on the app store. The Requirements Identifier and Classifier are used to automatically identify and classify software requirements into functional and non-functional requirements.

Williams and Mahmoud's Text Classifiers Method: Williams and Mahmoud propose using text classifiers such as Support Vector Machines (SVM) and Naïve Bayes (NB) to mine Twitter feeds to automatically elicit software user requirements. The approach also seeks to summarize common software concerns by making use of algorithms such as TF and hybrid TF-IDF. 4,000 tweets regarding 10 software systems are collected and manually classified. SVM and NB are used to automatically identify and classify the software requirements [11]. The dataset of 4000 tweets is collected by using Twitter's API which utilizes hashtags to search for specific words [11].

III. RESEARCH METHODOLOGY

Previous studies have focused on identifying and classifying functional and non-functional requirements from either requirements artifact such as SRS documents or from user app reviews [9] [11] [12] [13] [4] [2] [10] [5] [8] [6] [1] [7]. This study proposes a hybrid approach to identify and classify requirements from both SRS documents and user app reviews by using machine learning and NLP techniques. Fig. 1 provides an overview of the primary steps. The approach consists of four primary stages as follows:

1. **Data Collection:** A novel hybrid dataset consisting of formal software requirements and user app reviews that have been labeled and will be used as input to the data models.
2. **Data Preprocessing:** NLP techniques such as tokenizing and stop word removal are applied to preprocess the data. The text is then vectorized using TF-IDF.
3. **Model Training:** The preprocessed dataset is used to train the SVM, SGD, and RF ML algorithms.
4. **Model Evaluation:** The performance of the data models is evaluated on the test set by using 10 k-fold cross-validation to calculate the precision, recall, and F1 scores for each category defined as FR, NFRs, and NRs.

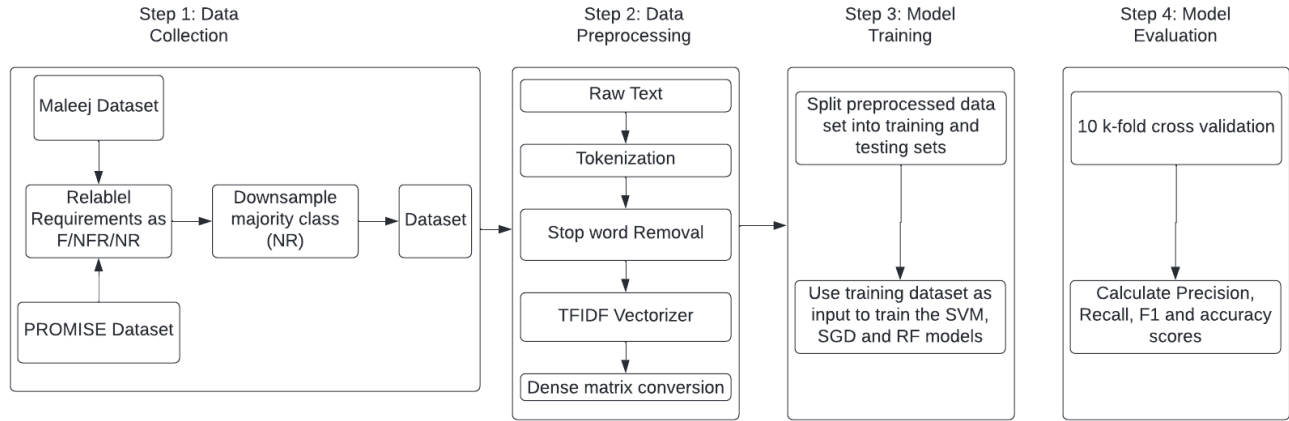


Fig. 1. Classification Process

A. Dataset

A novel hybrid dataset is proposed which consists of formal software requirements from the PROMISE Software Engineering Repository [14] and user app reviews from the dataset prepared by Maleej et al. [15]. The PROMISE dataset consists of 626 software requirements that are classified into FRs and NFRs. The NFRs are further classified into sub-categories: Availability, Legal, Look & Feel, Maintainability, Operational, Performance, Scalability, Security, and Usability. Fault Tolerance and Portability requirements are removed due to having a very small sample size. The user app reviews dataset consists of 3691 reviews from different Google’s apps store and Apple’s app store. The dataset is classified into the following categories: Feature Request, Bug Report, Problem Discovery, User Experience, and Rating. To prepare the dataset for preprocessing and training, the requirements from both datasets had to be labeled as FR, NFR, and NR to form a common set of classes between them. All NFR categories from the Maleej dataset were labeled as NFR and the FR label was unchanged. In the user app reviews dataset, Feature Requests were FRs and were labeled as such. The rest of the requirements were labeled as NRs. Due to the class imbalance of the requirements, the majority class, NRs, had to be downsampled to the count of the minority class, NFR, with 371 observations. The final dataset distribution with relabeling and downsampling is shown in Table I.

TABLE I. DISTRIBUTION OF FUNCTIONAL, NON-FUNCTIONAL AND NON-REQUIREMENTS IN THE FINAL DATASET

Category	Count
Functional Requirement (F)	507
Non-Functional Requirement (NFR)	371
Non-Requirement (NR)	371

B. Data Preprocessing

To have the data ready to be trained by the machine learning models, several NLP techniques were applied to the textual software requirements and user app reviews. The textual dataset is converted into a matrix of TF-IDF features using the TfidfVectorizer function. The TfidfVectorizer function

extracts unigrams and bigrams, splits the words into tokens, ignores terms that appear in less than 1 document, and removes stop words. After the raw text is vectorized with preprocessing, it is converted into a dense matrix.

The terms are defined as follows:

- N-gram: Contiguous sequence of n items from a given sequence of text
- Tokens: An instance of a sequence of characters in some document that are grouped as a useful semantic unit for processing
- Stop words: A set of commonly used words such as “at”, “which”, “the” that do not add a lot of meaning to the text.

C. Model Training and Evaluation

After the data is preprocessed, it is ready to be used as input for the ML algorithms. The dataset is split into train and test sets. The training dataset consists of 80% of the data with 999 observations and the test dataset consists of 20% of the data with 250 observations. The training dataset is used as input for the SVM, SGD, and RF ML algorithms. The algorithms are defined below:

- Stochastic Gradient Descent (SGD) [16]: Method to find the optimal parameter configuration for a machine learning algorithm. It iteratively makes small adjustments to a machine learning network configuration to decrease the error of the network
- Support Vector Machine (SVM) [17]: A classifier that creates a hyperplane in multi-dimensional space to classify observations
- Random Forest (RF): is an ensemble of individual tree predictors [18]

After the models are trained, their performance is validated using the testing dataset. The following metrics are used to evaluate the performance of each model.

- Recall: Calculates the true positives in a class out of all the observations in the class. It is defined as True Positive (TP) / TP + False Positive (FP)
- Precision: Calculates the number of true positives out of all the input classes. It is defined as TP / TP + False Negative (FN)
- F1: Calculated based on the precision and recall scores. It is defined as $2 * \text{Precision (P)} * \text{Recall (R)} / \text{P+R}$
- Accuracy: Calculates the number of true positives out of all the data points. It is defined as TP + True Negative (TN) / TP + TN + FP + FN

In addition to calculating the accuracy metrics, 10 k-fold cross-validation is used to ensure proper model validation by reducing any bias in the training and testing sets that may arise due to the random splitting of the data. In k-fold cross-validation, the data is split into equal-size groups. The number of groups is defined as k. k iterations of model training and testing are performed so that each iteration will have a different train and test set. The average scores of all the 10 groups are calculated to compare the performance of each model.

IV. EXPERIMENT RESULTS AND DISCUSSION

This section provides the results of the experiments to evaluate the performance of each data model. The precision, recall, and F1 and accuracy scores are discussed to answer the research questions.

RQ1: What is the type and size of data is required to automate the classification of requirements from user app reviews?

To answer this question, we proposed a novel approach to create a hybrid dataset consisting of formal software requirements from SRS documents from the PROMISE Software Engineering Repository and user app reviews from the

dataset prepared by Maleej et al. To create a well-balanced hybrid dataset, the requirements were re-labeled to only consist of NFR, FR, and NR. After re-labeling, the issue of class imbalance was resolved by downsampling the majority class, NR, to the minority class count of NFR. The relabeling of requirements and downsampling ensured consistent labels in the final dataset and reduced bias towards the majority class. The final dataset consisted of 507 FRs, 371 NFRs, and 371 NRs. Therefore, the final dataset consisted of a relatively equal count of each type of requirement.

RQ2: To what extent can the Machine Learning Algorithms combined with NLP techniques accurately identify and classify Functional and Non-Functional Requirements?

The precision, recall, and F1 scores of each data model were evaluated to determine their performance. The scores of each model is shown in Fig. 2.

The scores of all 3 machine learning algorithms were compared for each requirement type. In identifying FRs, the RF algorithm had the highest precision score of 0.812 whereas the SGD algorithm had the highest recall score of 0.795, and SGD had the highest F1 score of 0.788. For identifying NFRs, SVM had the highest precision score, and SGD had the highest recall and F1 scores of 0.916 and 0.913 respectively.

RQ3: How effective are the data models in isolating Non-requirements from Functional or Non-Functional Requirements?

The precision, recall, and F1 scores of NRs were evaluated to measure the performance of the algorithms in identifying NRs from FRs and NFRs. NRs are general feedback for an application and do not specify any requirement. For identifying NRs, the SGD and SVM algorithms had the best precision score of 0.783 and the RF algorithm had the highest recall score of 0.839. The SGD algorithm also had the highest F1 score.

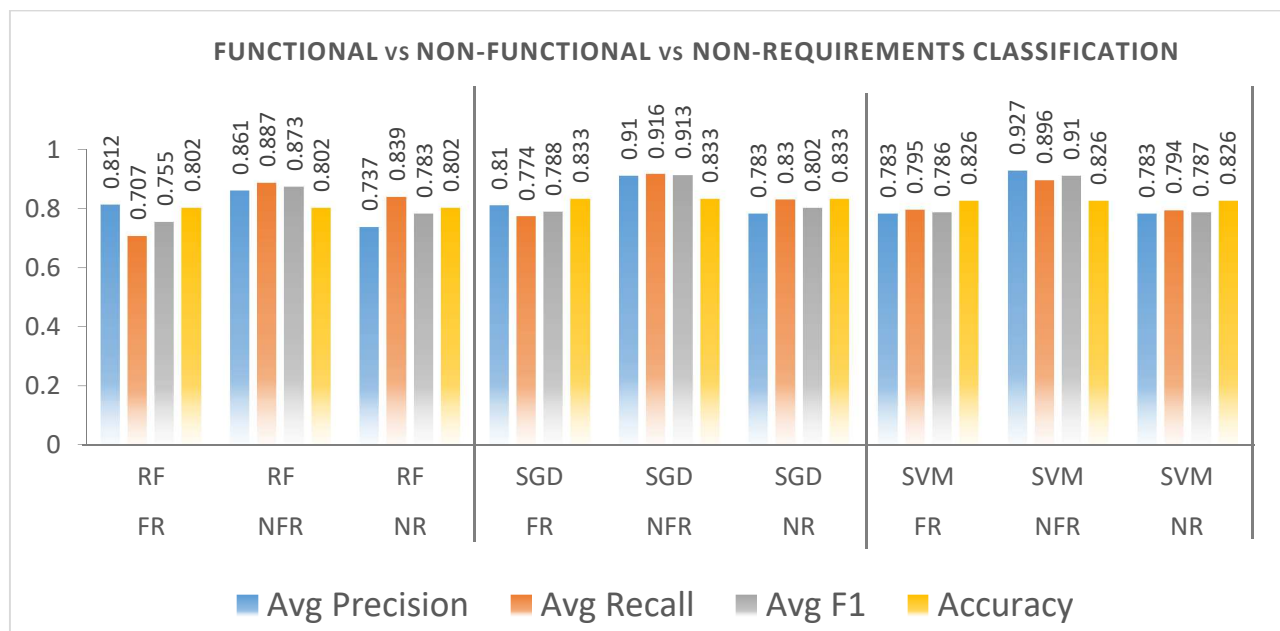


Fig. 2. Comparison of the scores for each machine learning algorithm

V. CONCLUSION AND FUTURE WORK

This paper presented the creation of a novel hybrid dataset that addressed the issue of extracting requirements from both, software requirements and user app reviews. Additionally, we proposed the use of NLP techniques such as TF-IDF to preprocess the data combined with using 3 different machine learning algorithms (SVM, SGD, and RF) to identify and classify requirements. Furthermore, we evaluated the performance of each algorithm for each requirement category. The automatic identification and classification of requirements from SRS documents and user app reviews can enable project managers and software engineers to catch any requirements early during the requirements engineering (RE) phase as well as improve the next iteration of their application by being able to quickly sift through vast amounts of user reviews to highlight user wants and needs that may be missed otherwise. In our future work, we intend to incorporate additional datasets consisting of formal software requirements and user app reviews from the various domain. We also plan on using more NLP techniques such as Bag-of-Words (BoW) and Word Embedding to preprocess the raw text data and implement additional machine learning algorithms such as XGBoost and deep learning methods such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

REFERENCES

[1] O. Daramola, T. Moser, G. Sindre, and S. Bi. Managing implicit requirements using semantic case-based reasoning. In REFSQ, Springer LNCS, pages 7915:172-178, 03 2012.

[2] D. M. Fernandez, S. Wagner, M. Kalinowski, M. Felderer, P. Mafra, A. Vetro, T. Conte, M.-T. Christiansson, D. Greer, C. Lassenius, et al. Naming the pain in requirements engineering. *Empirical software engineering*, 22(5):2298-2338, 2017.

[3] C. Page, *Software engineering: Theory and practice*. Willford Press, 2019.

[4] Binkhonain, Manal, and Liping Zhao. "A Review of Machine Learning Algorithms for Identification and Classification of Non-Functional Requirements." *Expert Systems with Applications: X*, vol. 1, 12 Mar. 2019, p. 100001., doi:10.1016/j.eswax.2019.10000.

[5] L. Dalpiaz. Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and NLP. In REFSQ, pages 119-135, 2018.

[6] M. Riaz, J. Slankas, J. T. King, and L. A. Williams. Using templates to elicit implied security requirements from functional requirements - a controlled experiment. In ACM-IEEE Intl. Symp. on Empirical Software Engineering & Measurement, ESEM 2014.

[7] O. Emebo, O. Daramola, and C. K. Ayo. Promirar: Tool for identifying and managing implicit requirements in SRS documents. In WCECS Conference, 2018.

[8] M. Lu and P. Liang, "Automatic Classification of Non-Functional Requirements from Augmented App User Reviews," *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, 2017.

[9] R. Chen, Q. Wang, and W. Xu, "Mining user requirements to facilitate mobile app quality upgrades with big data," *Electronic Commerce Research and Applications*, vol. 38, p. 100889, 2019.

[10] H. Yang and P. Liang, "Identification and Classification of Requirements from App User Reviews," *Proceedings of the 27th International Conference on Software Engineering and Knowledge Engineering*, 2015.

[11] G. Williams and A. Mahmoud, "Mining Twitter Feeds for Software User Requirements," *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017.

[12] D. Dave, V. Anu and A. S. Varde, "Automating the Classification of Requirements Data," *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5878-5880, doi: 10.1109/BigData52589.2021.9671548.

[13] A. A. Alshazly, A. M. Elfatraty, and M. S. Abougabal, "Detecting defects in software requirements specification," *Alexandria Engineering Journal*, vol. 53, no. 3, pp. 513–527, 2014.

[14] S. Shirabad, J. and T.J. Menzies (2005) *The PROMISE Repository of Software Engineering Databases*. School of Information Technology and Engineering, University of Ottawa, Canada.

[15] H. Assem (2019), "A dataset of Mobile application reviews for classifying reviews into software Engineering's maintenance tasks using data mining techniques", *Mendeley Data*, V2, doi: 10.17632/5fk732vkw.2.

[16] DeepAI, "Stochastic gradient descent," DeepAI, 17-May-2019. [Online]. Available: <https://deepai.org/machine-learning-glossary-and-terms/stochastic-gradient-descent>. [Accessed: 02-Apr-2022].

[17] X. Wu, V. Kumar, J.R. Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. 2008. Top 10 algorithms in data mining. *Knowledge and information systems* 14, 1 (2008), 1–37.

[18] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

Studying-Alive: A Holistic Wellness Application for College Students

Natasia Fernandez¹, Vaibhav Anu²
 Department of Computer Science
 Montclair State University
 Montclair, NJ, USA
 {fernandezn9¹ | anuv²}@montclair.edu

Abstract—Over the past few years, there has been an increased emphasis on mental health awareness and wellbeing. It is a topic that many do not feel comfortable talking about or acknowledging. Taking care of one’s mental health is important for the individual’s overall wellbeing. College students are a group of individuals who face several pressures that may affect mental health. It can become difficult to juggle their class schedules, assignments, work schedules, and social lives. Struggling to maintain these responsibilities can cause stress and anxiety for the students. Students may be uncomfortable to express their thoughts, as it may be difficult to discuss mental health. With the COVID-19 pandemic there has been an additional stress placed upon college students by taking extra precautions when going out and maintaining their social lives. To alleviate some of the stress and pressures faced by college students, a mobile app called Studying-Alive for holistic wellbeing management is being developed. This wellness app focuses on helping college students reduce stress and organize their schedules while tracking their responsibilities and overall health.

Keywords—mental health, college, students, motivation, anxiety, stress, mobile computing, wellness

I. INTRODUCTION

Mental wellbeing has become more prominent as individuals become more vocal about their mental health. As a college student, there are several responsibilities that one must face. College students have the weight of figuring themselves out while also finding a career path. This can be a great amount of pressure as it may determine the type of job they will be doing for the foreseeable future. In addition, college students have the pressure of maintaining their course load, social lives, internships, and dealing with personal matters. During the COVID-19 pandemic, the stress and anxiety of students increased [1]. They had the pressure of being in a lockdown which may have impacted their social lives. It also decreased the confidence of many college students as they were less active physically and stuck indoors. They also struggled with online classes, as the learning environment completely shifted. In Wang’s study on college students, “among the 2031 participants, 48.14% (n=960) showed a moderate-to-severe level of depression, 38.48% (n=775) showed a moderate-to-severe level of anxiety, and 18.04% (n=366) had suicidal thoughts” [1]. Overall, these are a significant number of students struggling with mental health. The majority also indicated that their stress levels had increased overall during the pandemic.

Several factors need to be addressed to help college students’ declining mental health. There may be several causes for a decline of mental health. Since the pandemic began, the major factors that impact the mental health of college students are fears about one’s physical health and of those around us, difficulty concentration, sleeping problems, decreased sociability, and concerns on academic standings [2]. These factors deteriorate the mental health of college students. Without any help, the students may have difficulty succeeding in school and lack ambitions to keep moving forward in life.

To alleviate some of the issues that students face, a mobile application called Studying-Alive is under development to allow college students to organize their busy schedules, maintain a balance for their social lives, and find ways to de-stress, while remaining safe from COVID-19. Although the app focuses mainly on mental health, it also considers the physical health of the students, as these two are interconnected. If one is not in a good mental state, it makes it difficult to have the motivation to take care of one’s physical health. Similarly, if one is having issues with their physical health they may be stressed out, which will affect their mental health. Therefore, this app helps organize one’s mental and physical and academic record. It also helps the user identify certain routines they may have developed to give them a chance to improve these routines, such as sleep routine. To test the usefulness of the app and gain feedback to improve development, a usability survey was conducted on college students (shown in Section IV). Overall, the app is aimed at improving the mental wellness of college students by providing more time and health management, and methods to de-stress.

II. RELATED WORK

There are several wellness apps that currently exist, but they are not specific to college students. The leading wellness app currently is Calm, which focuses on meditation, improving sleep, relaxing music, stretching, audio classes, and nature sounds. According to Conferroni posthoc tests conducted in Huberty’s study, the app Calm decreased stress, mindfulness, sleep disturbances, and self-compassion significantly with a p-value of less than .001 [3]. This shows that wellness apps are effective, but the app Calm is not specifically helpful to college students during a pandemic. It is more beneficial if students could have an app that is specifically design for their busy schedules and targets their specific stressors, especially during the time of a pandemic.

There are several other mental health applications (MHA), but they have struggled to get interest of college students. When 741 students were surveyed about MHAs, “26.1% of respondents were open to using an MHA yet only 7.3% had used an MHA” [4]. The study shows that there is potential in MHAs to help improve the mental health of college students. However, the students need to be more exposed to the applications as many have not used them. Therefore, a wellness app that is specifically for college students may help students feel more welcomed to using it.

Recently, there was a mobile app that was created for college students, but it faced a low uptake due to the pandemic. This application is called IntelliCare for College Students, which provides “evidence-based cognitive and behavioral skill-building exercises” to decrease depression and anxiety. Some of the features included in the app are mood recording feature, symptom assessment, available resources, and lessons on mental health. It is an app that provided some form of counseling, while also referring the users to more resources. Overall, the app proved to be efficient in decreasing anxiety and stress, but it had a low number of students using it. This may have been due to the pandemic reducing the forms of advertising the app. As this app was created before the pandemic, it may not address certain fears or anxieties that were developed from the pandemic. Studying-Alive will help manage stress, anxiety, and one’s responsibilities rather than help diagnose symptoms of a student. The Studying-Alive app also considers that there is a pandemic occurring. For example, it includes activities that the user may participate in that will help them maintain their social lives while practicing precautions to avoid contracting COVID-19.

III. APPLICATION DESIGN OUTLINE

The Studying-Alive App is being developed to include several features that would help reduce stress and anxiety. The features are Goal Tracker, Schedule, Activities, Routines, Calming Techniques, Recipes, Social Forum, Motivation, and Achievement. Before starting the development of the app, a prototype was created to outline the features that the app would include and the overall design. This prototype was created using a website Proto.io, which facilitates the design process with a drag and drop interface [5]. In this website, each page of the app was outlined, and each feature was created. This created a blueprint of how the app is going to be designed. From this prototype design, information can be gathered on the response of students so that the feedback can be used to enhance the development of the application.

A. Database

The database that it is currently being developed with is Google’s Firebase [6]. This database facilitates authentication of the login page using an email and password. It helps store the information efficiently and easily. For the Studying-Alive app, the Cloud Firestore option for the database is being used as it is a flexible server that syncs the data across apps and allows for support even when there are issues with the Internet connection [9].

B. Login/Starting Survey

The Studying-Alive app will begin with an option to take a personalized survey or sign in (see Fig. 1). The personalized

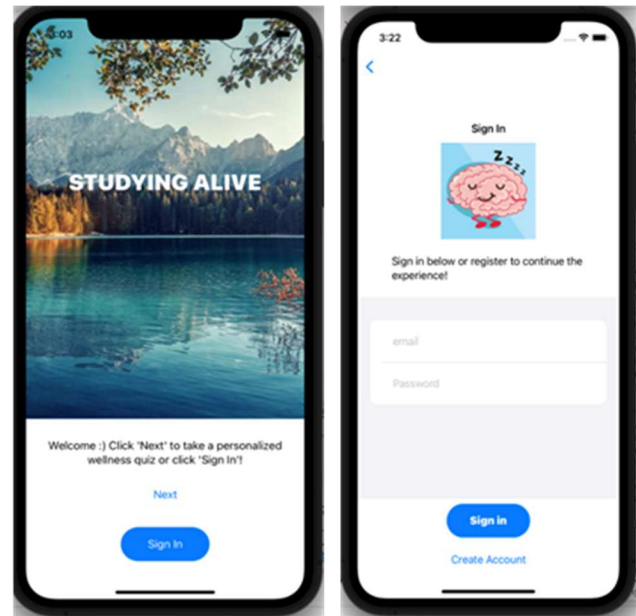


Fig. 1. Studying-Alive App: Initial screen and Sign-in page.

survey helps identify what the user may be struggling with. This may help gain the interest of the user as the app will be specifically focusing on the areas in which they are having difficulties. To sign in, the Studying-Alive app will require the user to enter an email and password. This will then bring them to the home page, which will include the nine features that the Studying-Alive app will provide.

C. Goal Tracker

The Goal Tracker feature will allow users to keep track of the daily, monthly, yearly, and long-term goals while also viewing their progress. When the user adds a new goal, it will be stored in the database, which will add the goal into the lists with an unchecked box. When the goal is completed, the user will mark it as completed, which will increase the percentage of completed goals in the progress ring. In Fig. 2, the current Goal Tracker page is displayed while the app is under development. Overall, this feature will help keep the student motivated, while also helping them to remember that the small daily tasks should be celebrated as well. Students should be proud of their daily accomplishments in addition to keeping their end goals in sight.

D. Schedule

The Schedule features allows the users, which are mainly college students, to organize their classes, assignments, and upcoming events. One of the most stressful parts of being a college student is time management. It is also one of the most essential skills to acquire that is linked to succeeding in one’s academic career [7]. This skill can help decrease procrastination in students as they will learn how to organize their schedules, which can lead to a decrease in stress and anxiety. If students are waiting to complete their assignments when the due dates are closer, they will be more stressed out as they feel like they need to crunch in as much as they can in a short amount of time. This can also help them organize their social lives.

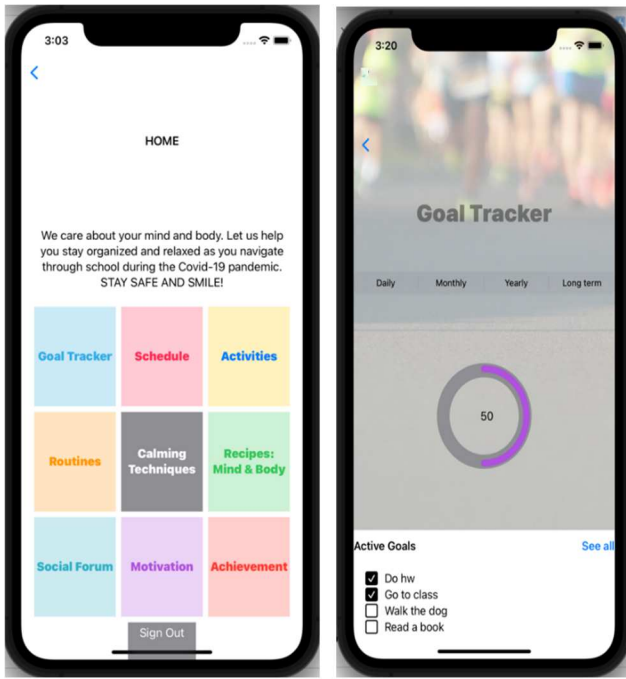


Fig. 2. Current Home screen and the Goal Tracker feature in the app

E. Activities

The Activities feature of the Studying-Alive app offers a variety of activities that the user can do while still taking precaution against COVID-19. It offers activities for solo participants, groups, and outdoor activities. This allows college students to remember that in addition to focusing on their schoolwork, it is also important to take time for themselves to enjoy different activities and hobbies. It is also crucial that “Social support may have indirect effects on health through enhanced mental health, by reducing the impact of stress, or by fostering a sense of meaning and purpose in life. Supportive social ties may trigger physiological sequelae (e.g., reduced blood pressure, heart rate, and stress hormones) [8]. Maintaining a balanced and healthy social life can help improve both mental and physical health. These activities include COVID-friendly places and activities, such as bike riding and rollerblading with friends.

F. Calming Techniques

A crucial part to helping mental health is finding useful techniques to reduce anxiety and try to maintain calm. The Calming Techniques feature offers different techniques that can help reduce stress and anxiety and allows the user to favorite the techniques that work for them. The users can keep these favorited techniques saved so that they can go back to these techniques when needed. Progressive muscle relaxation (PMR) is one of the techniques that help trigger relaxation. It involves the contraction of muscles and slowly relaxing them, which has proven helpful to reduce stress in nursing students [10]. Deep breathing is another technique that helps increase blood oxygen levels and decrease anxiety. In a survey of 4,793 pre-surgical patients, about 40% of the participants showed a decrease in anxiety from using the deep breathing technique [10]. These techniques in the Studying-Alive App will offer students a

healthy, safe, and effective method to decrease anxiety easily and relatively quickly.

G. Recipes: Mind & Body

Physical health and mental health are intertwined. Therefore, eating healthy foods and maintaining a balanced diet can improve mental health. As a college student, one’s diet may be limited to the food on campus, which may include several unhealthy choices. As students navigate more independence on top of academic stress, they may find it difficult to make healthy food choices. In a study of 1956 college students, there was a correlation between anxiety and added sugars intake with a p-value of 0.005, which indicates statistical significance [11]. This shows how what is consumed affects both the body and the mind. Therefore, this Recipes feature in the Studying-Alive app will provide college students with quick and easy recipes to make them feel healthy and provide them with ample energy.

H. Social Forum

The Social Forum allows the students to have a platform in which they may discuss their current mental state and any stresses they may be facing. This can help them acquire feedback from other students that may be experiencing the same events or feelings as them and listen to how others have handled this situation. It can help users that they are not alone and do not need to go through their academic stresses and personal struggles by themselves. This can also be a way to expand the social lives of the users. The users can meet other friends through this application to complete the activities with. It can help the users of the app form a community where they may help each other. This feature can make advice from others and the ability to talk about one’s feelings and thoughts more accessible. Three benefits that social platforms can provide for others are that it “facilitates social interaction”, provides “access to peer support network”, and it “promotes engagement and retention in services” [13]. The accessibility of this social forum can offer these three benefits, while also giving the users an option to not use the feature if they do not feel comfortable.

I. Motivation

College students often need motivation to help them keep persevering and have the determination to complete their degree. The Motivation feature in the Studying-Alive app will have motivational stories and quotes that can resonate with the users and get them to keep moving forward. It will also have quotes that can help push them beyond their comfort zone and set higher goals for themselves.

J. Achievement

The Achievements feature is a reward system that helps the user acquire badges as they use the app. These badges may be given for using different features or for accomplishing a goal. In Platt’s “Evaluating User’s Needs in Wellness Apps,” a survey of 519 participants indicated that one of the features that wellness apps should include to encourage users to interact with the app more frequently is an achievements feature [12]. The users would be getting rewarded for their consistency while also helping them improve their mental health. It will make the app more enjoyable to use as users will want to unlock more badges and it makes the user feel more accomplished. It also makes the

app more interactive as the students are rewarded for using the app.

IV. USABILITY SURVEY

To test the app’s usability, an informal survey was conducted with college students at Montclair State University. A Google form was created with a video of the prototype of the app and nineteen questions to acquire feedback from the students. This helped provide insight into which features, and modifications students may or may not want in a wellness app. The survey received responses from 36 students overall with 2.8% being sophomores, 19.4% being juniors, 63.9% being seniors, and 13.9% being graduate students. Of these students, 72.2% have never used a wellness app, while 27.8% have. Therefore, most of these students have not been exposed to a wellness application before.

A. Interface Design Feedback

An important aspect that needs to be considered during app development is the user interface design. This is the overall style of the app, which the user will interact with. Ideally, an app should be easy to use, interesting to use, and have a style that

attracts the user. It should be pleasing to the eye and motivate the user to keep using it. As shown in Fig.3, according to the survey, about 55.5% of the participants gave the user interface design a four or five rating, which means more than half of the students liked the design. This means that students find the overall look of the application to be pleasing. According to Liew et al.’s survey, the three most important attributes that contribute to the usability of an application are “Satisfaction, Learnability, and Efficiency.” The feedback on the user interface shows that satisfaction can be completed by the Studying-Alive app as the interface feels comfortable and attractive to the user. In other questions, students were allowed to comment more upon the design. Overall, the feedback was positive, but there were several students that believes that the app needed more rounder edges and less colors. When the students were asked for how easy it seems to use the Studying-Alive app, 80.6% of the students gave it a score of four or five, indicating that overall, the Studying-Alive app will be easy to use. This satisfies the second most important attribute, which is Learnability. The students found the app to be easy to use and learn. The app needs to be easy to use to attract users and to accomplish its purpose of alleviating stress.

B. Features Feedback

The Studying-Alive app includes nine features, which are Goal Tracker, Schedule, Activities, Routines, Calming Techniques, Recipes, Social Forum, Motivation, and Achievement. The students who completed the survey were asked for their opinions on these features. The feedback of the students is crucial to ensure that the app will provide features that are useful and are what students feel that they need for their academic success and to improve their mental state. The app offers a great variety of feature, which helps cover different contributors to improving one’s mental health. The students were asked to organize the nine features from 1 to 9, with 1 meaning the best and 9 meaning the worst feature. The stacked bar chart in Fig.4 shows how many students placed each feature under each rating. For example, students voted ‘Schedule’ and ‘Goal Tracker’ the most for the number 1 spot, indicating that these two features are the students’ favorite features, with 13 and 12 votes, respectively. These two features may be the ones that students find the most useful as they are both features that can help in school. Students can use the ‘Schedule’ feature to insert their class schedules and keep track of other events, such as extracurricular and personal events. The feature that came in third place for the number 1 rating was ‘Achievements’. Wellness apps should promote a goal and provide nurturing to lead to “habit formation, self-awareness, and goal attainment” [14]. The ‘Achievements’ feature helps provide small goals within the app by giving badges for using the app and completing certain goals in the app, such as favoriting a calming technique. This will help develop a habit of using techniques to calm oneself and using features for organizations and other habits. The badges are similar to accomplishing small goals within the app, which will further motivate the users. In the second best (2) rating, the most voted were both ‘Goal Tracker’ and ‘Routines’, which nine votes for both. For the third favorite rating (3), the feature ‘Activities’ received the most votes, with nine votes. According to this question, the top five features of the app starting with the highest rated are Schedule, Goal Tracker, Achievements, Routines, and Activities. These are the

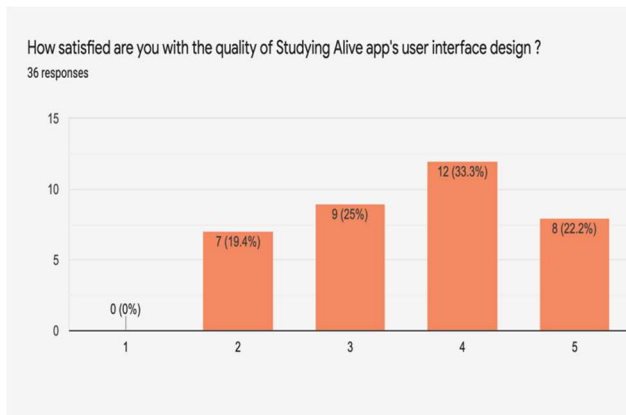


Fig. 3. User satisfaction with the interface design of the Studying Alive app prototype

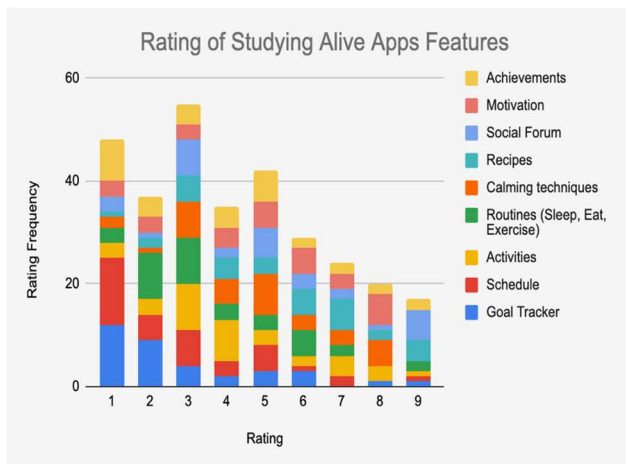


Fig. 4. User ratings of the features with 1 being the best and 9 being the worst.

features that students would be most likely to use and to find the most necessary as a college student. The feature that received the most votes for the worst rating (9) is the Social Forum feature, with six votes. The user may find this feature to be not as useful or may be concerned with how effective it may be.

Some other features that the students recommended are a journal section to keep track of mental state, yoga positions, meditation techniques, an educational feature, brain teasers, music, and daily jokes. These are all features those students believe could be a good addition to the Studying-Alive applications. The students may find that these additions could help them be more engaged with a wellness application or this may be the features that they feel may be missing from other wellness applications. These ideas can be possibly incorporated into the app while development. For example, meditation techniques and yoga positions can be added onto calming techniques and can possibly be created as categories within this feature. “Most common evidence-based strategy was mindfulness-meditation, followed by positive psychology and goal setting” [15]. Therefore, the addition of meditation techniques and positive psychology, such as the addition of jokes in the motivation feature, may help the effectiveness of the application.

C. Organization

The organization of an application can also be a factor in the success of an application. The way an application is organized will determine whether users want to continue to use it. If an application lacks organization, it may frustrate the user. According to the informal survey (shown in Fig. 5), the students found the prototype of the application to be organized overall with 30.6% of the students giving it a five in organization and 50% giving it a four. A rating of five indicated that the app is very organized, while a four indicates it is organized. In the additional comments, students expressed that the colors and text need to be more cohesive throughout the application. The original prototype of the app incorporated a wide variety of colors and different size fonts for each feature. The students found that this made the app slightly less appealing. Consistency with colors and text are more appealing to the eyes. They prefer a simpler theme across the entire application as it helps keep more focus to the features itself rather than to the design. They also found that in some features, such as the stories in Motivation, the font was too small and difficult to read. Therefore, in the implementation of the Studying-Alive application, modifications will be made to make the colors and organization more appealing. The fonts need to be legible in order for the users to be captivated and understand the contents within the features. The colors between text and pictures need to contrast well against each other to make it easy to see. The edges will also be rounded as the users find round edges to portray a cleaner appearance to the application. Overall, the organization is adequate for the application, but it still needs some modifications to make the Studying-Alive application more user-friendly.

D. Effectiveness/Usefulness of the Studying-Alive app

The goal of the Studying-Alive application is to provide students with tools that can help alleviate stress and anxiety while helping them succeed as they navigate through college.

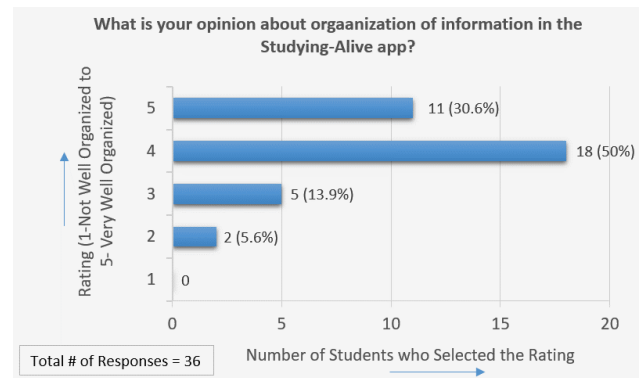


Fig. 5. User responses towards the organization of the Studying Alive application.

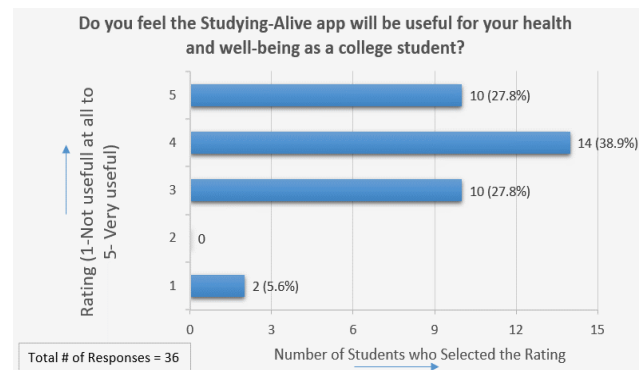


Fig. 6. User response for the effectiveness/usefulness of the Studying Alive

Since the Studying-Alive application is still under development, students cannot be asked to use the application to test its effectivity. Instead, students were asked to rate how useful they believe the application would be in improving their mental and physical health. The students were asked to rate the applications usefulness according to a scale from one to five with one being not useful and five indicating that the application is very useful.

According to Fig. 6, about 66.7% of the students rated the effectivity of the Studying-Alive application a four or five indicated that they believe the application would be useful in alleviating their stress and anxiety. Additionally, 27.8% of the students gave a 3-rating indicating that they were not sure how useful it would be. Only 5.6% of the students indicated that they did not think the application would be useful. Therefore, the majoring of the students has found that the application contains at least one feature that can help them in decrease the stress in their personal lives and academic career. An additional study after the application is fully developed could indicate if any of the responses of the students would change once they are able to use the application and apply it to their lives.

E. Frequency of usage

As previously mentioned, the application IntelliCare was a wellness application developed recently for college students, but it failed to grasp the attention of students as it has a low rate of usage. A mobile application needs to be able to attract users so that it can accomplish its purpose. The application’s usefulness

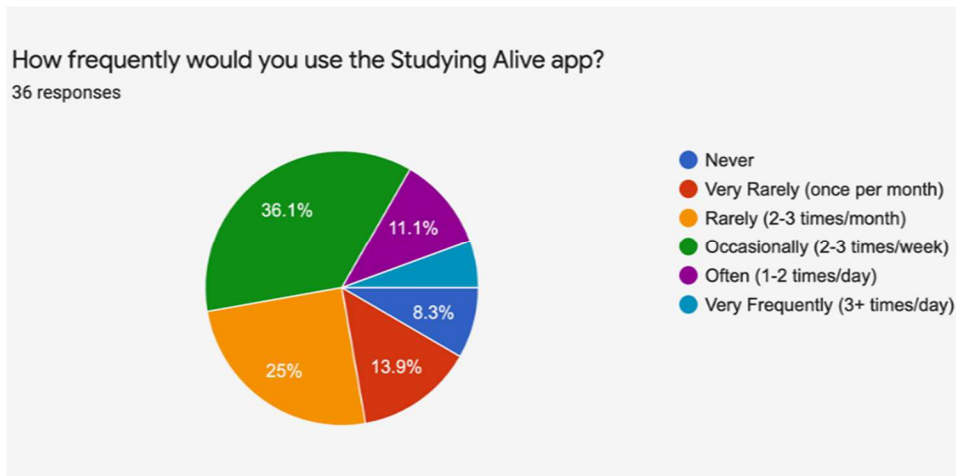


Fig. 7. Frequency-of-Usage Responses

and worthiness are also correlated to how often a user may use the application. In the informal survey, the college students were asked to give an honest opinion on how often they would use the Studying-Alive application. About 52.8% of the students said they would at least use the application occasionally (see Fig. 7). This shows that a majority of college students are willing to use a wellness application, specifically the Studying-Alive application. Only 5.6% of the students, however, said that they would use it very frequently, with the majority claiming they would use it occasionally. Approximately 8.3% of the students said they would never use the application. Therefore, about 91.7% of students would at least use the application once, which is a large number of students. Students may have different ways of dealing with stress that an application may not be able to provide. Therefore, it is expected that some students may not be willing to test out the application or some may not use it as often. However, the Studying-Alive application, has several different

features which means it can appeal to various students with different methods of de-stressing and various students in need of different areas for help. Overall, the initial responses from students indicate that the usage of the application can be relatively high, especially once the students can test the implementation of the Studying-Alive application.

F. Additional Suggestions From Students

At the end of the survey, the students were given open ended questions to allow for more detailed explanations and further suggestions. Several students gave helpful advice that will be considered while completing the development of the application. In Fig. 8, there are a few suggestions that students have included, which may be added into the development or further versions of the application. One that stands out is the fear of the social forum platform because social media can negatively impact students and young adults. However, a great enhancement could be to restrict and filter the comments people post to only allow encouraging comments and motivational comments. This is a valid concern and will need additional modifications.

V. CONCLUSION

This paper has presented a wellness application called Studying-Alive aimed at college students. Overall, the application received positive feedback during the informal survey (responders were college students). In Fig.9, some of the positive responses and comments received have been included.

- “Add a Personal journal feature (deleted daily or saved), fitness and diet/calorie tracker”
- “Better design. Be wise when picking the colors displayed, and make it match. For example, for calming activity, match it with a color that is calming and so on.”
- “a feature I think that should be added should be a habits section. you can track specific habits that a person needs to track, such as medication, skincare, or drinking enough water.”
- “Research visual vs auditorial alarms and consider if it is possible to have the app connect with a google home or Alexa to connect with smart lights to achieve this idea.”
- “College students like apps that are supported by some sort of incentive.”
- “To be honest, the social forum is my least favorite feature. It can be a good addition, but social forums and media can also have a negative affect instead of a positive one. It would have to be moderated.”

Fig. 8. Students’ suggestions for improving the Studying-Alive application

- “I think this app could go a long way. it reminds me of Calm, except it's something a person can personalize and track themselves. the many features available are a great way to help people also start habits, for example, if a person thinks they should start doing physical activities.”
- “App will be helpful to improve mental health and overall health”
- “Can’t wait to try”
- “No dislikes at this time”

Fig. 9. Students’ positive reactions to the Studying-Alive application

The interface design had relatively high ratings, but needs some improvements with the shaping, color, and text sizing. The top 3 features were Schedule, Goal Tracker, and Achievements in the ratings (see Fig. 4). These features seem to be what students mostly want in a wellness application. The organization of the Studying-Alive application was highly rated but may need a few minor adjustments. The students also positively rated the usefulness of the application with most students indicating that the application will be very useful (see Fig. 6). The frequency-of-usage feedback from the college students was also positive as the majority of students claimed they would at least use the Studying-Alive application 2-3 times a week (see Fig. 7). The application may require some marketing by universities so that more students can know about the existence of the application so that more students can benefit from its features.

As the world continues to struggle with the COVID-19 pandemic, college students have endured a large hit to their mental health as they increasingly are struggling with anxiety and depression [1][2]. They are expected to become accustomed to the new methods of online teaching while also maintaining their social lives, dealing with their personal situation, and succeeding academically. The Studying-Alive application offers various features that can help minimize the stress that is accumulating on these college students while allowing them to still take precautions against COVID-19.

As for future works, the application is currently in development as a few of its feature are already being made. The feedback from the survey is being considered in the development of the application. Conducting a larger survey once the development of the application is completed can help give greater insight on the effectivity and usage of the application. The additional features that were suggested by the students will also be considered by adding them into the nine features or keeping them as possible updates for future versions of the application.

Furthermore, after the first iteration of the application development is completed, we intend to identify the features that are found ineffective (i.e., not useful/desirable) by our users. Such features will be replaced by features suggested by students during our survey. In the recent future, the Studying-Alive application will be continuously developed (and refined) and the authors anticipate that this app will bring significant positive change to the lives of college students by improving their mental and physical health.

REFERENCES

- [1] X. Wang, S. Hegde, C. Son, B. Keller, A. Smith, and F. Sasangohar, "Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional Survey Study," *Journal of Medical Internet Research*, vol. 22, no. 9, 2020.
- [2] C. Son, S. Hegde, A. Smith, X. Wang, and F. Sasangohar, "Effects of COVID-19 on college students' mental health in the United States: Interview survey study," *Journal of Medical Internet Research*, vol. 22, no. 9, 2020.
- [3] J. Huberty, J. Green, C. Glissmann, L. Larkey, M. Puzia, and C. Lee, "Efficacy of the mindfulness meditation mobile app 'calm' to reduce stress among college students: Randomized Controlled Trial," *JMIR mHealth and uHealth*, vol. 7, no. 6, 2019.
- [4] A. Kern, V. Hong, J. Song, S. K. Lipson, and D. Eisenberg, "Mental health apps in a college setting: Openness, usage, and attitudes," *mHealth*, vol. 4, pp. 20–20, 2018.
- [5] D. S. D. Lead, J. Y. I. Designer, L. C. D. of M. Design, and C. G. A. C. Director, "Prototyping for all," *proto.io*. [Online]. Available: <https://proto.io/>. [Accessed: 24-Mar-2022].
- [6] "Firebase documentation," *Google*. [Online]. Available: <https://firebase.google.com/docs>. [Accessed: 01-Apr-2022].
- [7] R. V. Adams and E. Blair, "Impact of time management behaviors on undergraduate engineering students' performance," *SAGE Open*, vol. 9, no. 1, p. 215824401882450, 2019.
- [8] D. Umberson and J. Karas Montez, "Social Relationships and Health: A flashpoint for health policy," *Journal of Health and Social Behavior*, vol. 51, no. 1_suppl, 2010.
- [9] "Cloud firestore | firebase documentation," *Google*. [Online]. Available: <https://firebase.google.com/docs/firestore>. [Accessed: 04-Apr-2022].
- [10] L. Toussaint, Q. A. Nguyen, C. Roettger, K. Dixon, M. Offenbacher, N. Kohls, J. Hirsch, and F. Sirois, "Effectiveness of progressive muscle relaxation, deep breathing, and guided imagery in promoting psychological and physiological states of relaxation," *Evidence-Based Complementary and Alternative Medicine*, vol. 2021, pp. 1–8, 2021.
- [11] R. Wattick, R. Hagedorn, and M. Olfert, "Relationship between Diet and mental health in a young adult Appalachian College population," *Nutrients*, vol. 10, no. 8, p. 957, 2018.
- [12] A. Platt, C. Outlay, P. Sarkar, and S. Karnes, "Evaluating user needs in wellness apps," *International Journal of Human-Computer Interaction*, vol. 32, no. 2, pp. 119–131, 2015.
- [13] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, "Social Media and Mental Health: Benefits, risks, and opportunities for research and Practice," *Journal of Technology in Behavioral Science*, vol. 5, no. 3, pp. 245–257, 2020.
- [14] M. Alshawmar, H. Mombini, B. Tulu, and I. Vaghefi, "Investigating the affordances of wellness mhealth apps," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021.

A Review of Cognitive Dynamic Systems and Its Overarching Functions

Waleed Hilal
*Department of Mechanical
 Engineering*
 McMaster University
 Hamilton, ON, Canada
 hilalw@mcmaster.ca

Alessandro Giuliano
*Department of Mechanical
 Engineering*
 McMaster University
 Hamilton, ON, Canada
 giuliana@mcmaster.ca

Stephen A. Gadsden
*Department of Mechanical
 Engineering*
 McMaster University
 Hamilton, ON, Canada
 gadsden@mcmaster.ca

John Yawney
Adastra Corporation
 Toronto, ON, Canada
 john.yawney@adastragr.com

Abstract—Cognitive dynamic systems are a new field of physical systems inspired by several areas of study such as neuroscience, cognitive science, computer science, mathematics, physics and engineering. Building on Fuster’s paradigm, a system is considered cognitive when it is capable of five fundamental processes to human cognition: the perception-action cycle, memory, attention, intelligence and language. With these capabilities, a cognitive dynamic system can sense its environment, interact with it, and learn from it through continued interactions. The goal of this paper is to provide a thorough review of the cognitive dynamic system framework, along with its theory, applications, and its two special functions: cognitive control and cognitive risk control.

Keywords—cognitive dynamic systems, cognitive control, cognitive risk control, cognitive radio, cognitive radar, cognitive internet of things, smart systems

I. INTRODUCTION

Cognitive dynamics systems (CDS) are a new class of systems which combine knowledge from several disciplines such as neuroscience, cognitive science, computer science, mathematics, physics and engineering [1]. The goal of CDS is to provide a framework for dynamic systems to augment them with cognitive capabilities, allowing them to sense, interact and learn from their environments. A dynamic system is considered to be cognitive when it can carry out five fundamental processes which are critical to human cognition. As defined by Fuster’s paradigm, these processes are the perception-action cycle (PAC), memory, attention, intelligence, and language [2].

Interest in the field of CDS has been growing at an increasing rate in recent years thanks to the seminal work carried out by Haykin on cognitive radio and cognitive radar [3] [4]. These two applications are among the earliest examples of CDS, stoking research into the theory and design of the CDS framework [1]. With cognitive radio, the primary goal is to solve the spectrum scarcity issue by providing the means for radio systems to access underutilized bandwidths. Cognitive radar, on the other hand, has been proposed as a means of providing improved accuracy and reliability in remote-sensing applications [2]- [4].

In essence, the CDS framework can be broken down into two special functions: cognitive control (CC) [5] and cognitive risk control (CRC) [6]. In the former, the limitation of current adaptive controllers and neurocontrollers when faced with

unmodeled dynamics or unstructured environments are addressed. Specifically, CC is additive in nature, meaning that it is augmented to existing system designs by introducing a new state known as the entropic state. The entropic state is based on the notion of an information gap that must be controlled alongside a system model. The second special function, CRC, expands on the CC architecture to account for the risks associated with the uncertainties that are faced by a system, and bring them under control. Such risks may include security threats frequently encountered by physical systems, like cyberattacks on smart grids on jammers acting on radar systems.

The first goal of this paper is to provide a detailed background on the theory and the architecture behind the CDS framework and its two special functions CC and CRC. The second goal is to present a short, structured overview of the current state of the field by reviewing recent literature on applications within the CDS framework. We aim to present a thorough discussion on the methodologies, key findings, experimental results, and limitations of the surveyed literature. Finally, we offer insight into the most promising areas for future research efforts in this field. This paper is organized as follows: Section II of this paper presents a background the CDS framework, theory and architecture. In Section III, cognitive control is introduced as the first special function of the CDS framework, with a review of relevant applications of this architecture. Similarly, Section IV introduces CRC as the second special function of the CDS framework, with a short review of applications and recent advancements. Finally, concluding remarks and suggestions for future researchers are discussed in Section V.

II. COGNITIVE DYNAMIC SYSTEMS

A. Perception-action Cycle

There are two parts to any CDS: the perceptual and the executive. On the right-hand side of Fig. 1, the perceptual component or perceptor is located, whereas the executive or cognitive controller is located on the left-hand side [7]. Depending on the CDS application, the perceptor is in charge of directly observing the system and the environment using appropriate sensors. During perception, for example, a Bayesian estimator might be utilised, which computes the posterior of a system’s state in each PAC and extracts relevant

information from what is experienced. A feedback connection transmits the perceptor's extracted relevant environmental information to the executive, which is then tasked with conducting cognitive or physical actions on the environment or system based on this knowledge [7].

The executive's cognitive efforts are designed to continuously improve the information extracted by the perceptor in following cycles. As a result, the executive indirectly observes the environment through the perceptor and acts on the information obtained, completing the PAC with a global feedback loop. In order to indirectly impact the system's perception, cognitive actions are frequently applied to the surroundings, such as increasing the lights in a dark room [5]. The physical state in this scenario contains the positions of objects that are not changed by light. Other forms of cognitive operations, such as altering the system's own sensors or actuators, can be performed alone on the system. The adaptation of a transmitted waveform in a cognitive radar system is an example of this. Furthermore, by adding a new component to the state controller's cost function called the entropic state, cognitive actions can be used to impact state-control actions [5].

The cognitive controller, as shown in Fig. 1, is in charge of making decisions about the aforementioned cognitive operations in the executive, based on the entropic state established by the perceptor [8]. However, all of the distinct types of cognitive actions are not necessarily present in a given problem. A cognitive radar system, for example, can assess the target's states without being able to physically manipulate them because it only executes cognitive actions on its own actuators and the environment [5]. The application of reinforcement learning (RL) for the cognitive control agent is the mechanism underpinning executive decision-making.

B. Memory

The cognitive process of memory occupies its own physical space in a CDS, as shown in Fig. 1, in three forms: perceptual, executive, and working memory. Both sorts of memories have slightly different functions and responsibilities, but the primary purpose of equipping a CDS with memory is to allow for the acquisition and storage of long and short-term information [2]. A CDS can learn from its past experiences in terms of action

and perception with access to this data, resulting in enhanced performance and resilience.

According to the CDS concept, perceptual memory should have a hierarchical structure with multiple layers [7]. The goal of this setup is to perform perceptual abstraction of incoming inputs or measurements in order to represent the essence of an object, event, or experience, similar to how the human memory system works. Relevant information is kept whereas irrelevant information is eliminated, allowing for long-term memory in the perceptual component of the CDS [9]. In response to feedback information from the perceptor, executive memory serves a dual function with perceptual memory. The executive memory stores long-term cognitive actions made by the cognitive controller based on input from the perceptor and can be utilised as a reference for future cognitive activities. A new policy that considers both long-term and short-term experiences is produced by adding the output of the executive memory to the cognitive controller and incorporating it into future policies [8]. The executive memory effectively keeps track of the cognitive controller's action space in a probabilistic fashion.

The working memory's function is to reciprocally couple the executive and perceptual memories, acting as a short-term memory interface between the two inside the CDS [4]. The cognitive controller may carry out its actions in each PAC in a synchronised manner as a result of this integration, and in summation, memory's general job is to continuously learn from and model the behaviour of the environment and the CDS's action space [8].

C. Attention

Unlike the PAC and memory, which have their own physical locations in the CDS, the cognitive process of attention reveals itself through computational mechanisms inside the framework. There are two types of attention: perceptual and executive attention, which are both based on localised cycles and feedback linkages in their respective sections of the CDS [5]. Their responsibilities include prioritisation of activities and effective resource allocation, and they work closely with and are driven primarily by the presence of memory. This is accomplished in the perceptor, for example, by a variety of strategies that can be utilised to filter out unnecessary input using previously stored characterizations of the environment. On the executive side, attention uses the well-known explore-

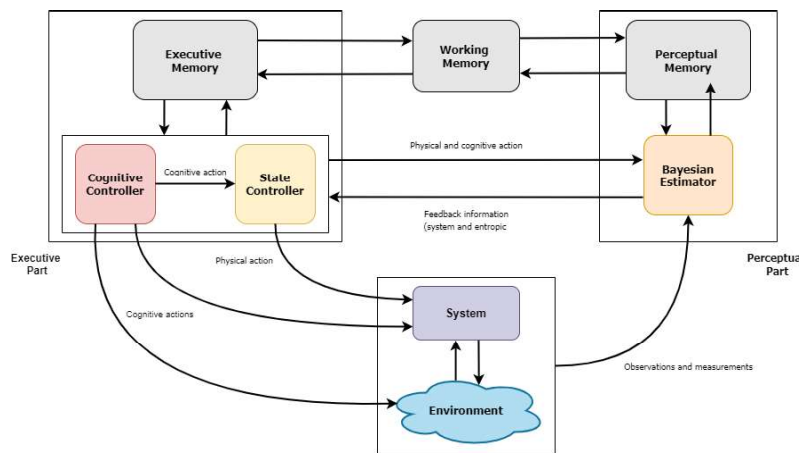


Fig. 1 Block diagram of the basic structure of a cognitive dynamic system and its architecture

exploit tradeoff to help the RL learn from and plan cognitive actions for future cycles, primarily by reducing the action space of the memory for the RL to consider based on the relevance to the perceived information in each cycle [4] [8].

D. Intelligence

Intelligence, like attention, does not have its own physical location in a CDS. It, on the other hand, draws on all prior cognitive processes like memory and attention and integrates them with the PAC to promote computational intelligence through efficient decision-making [2] [10]. Intelligence has an impact over the entire CDS, and its power and effectiveness in information processing are generated from leveraging all of the system's feedback loops, both global and local. As a result, intelligence plays a critical part in the CDS framework in terms of optimal decision-making in terms of the controller's actions on the system or environment of interest [7].

III. COGNITIVE CONTROL

A. Overview of Cognitive Control

Cognitive control is a paradigm that was first introduced in 2012 by Haykin et al. [5] and is additive in nature rather than a replacement system design paradigm. CC can increase the usage of computational resources and lower the correctly defined risk functional for the work at hand by complementing state-control paradigms such as adaptive control and neurocontrol. However, before defining CC, the concept of an information gap must be established, which is related to the risk associated with a policy or action. It is thus possible to define CC with the goal of minimizing the information gap.

The following is a useful summary of the concept of the information gap [5]: From noise-affected data, available information is retrieved and transformed from the measurement space to the information space. The accessible information is then partitioned into useful and redundant information based on the task at hand. In addition, sufficient information must be defined as the information required to complete the task at hand while reducing risk; relevant information is thus the intersection of available information and sufficient information. Furthermore, the information gap can be defined as the difference between sufficient information and relevant information.

Quantifying and reducing the information gap necessitates the development of a task-specific metric – this concept equates to a new state that must be managed. The state of a dynamic system represents the basic information characterising the system's conditions at a given point in time, and the state trajectory, or change in state over time, describes the system's behaviour. The state, on the other hand, can only be obtained through noisy measurements, which necessitates a perception process in order to determine a posterior distribution of the state using a Bayesian generative model. The information gap is the difference between the maximal relevant information in the posterior distribution and the required statistics for a specific job. this quantity is also known as the entropic state, which gets its name from Shannon's entropy [5] [11]. As a result, thinking about a two-state model of a CC system, composed of a state-

space model, which describes the evolution of the system state over time, and an entropic-state model, which quantifies the information gap given the posterior computed by perception, seems intuitive. According to statistical differences in the environment, both models may change from one cycle to the next. It's also worth emphasising that the entropic state is simply the feedback information sent on to the cognitive controller, and that CC is merely a paradigm for minimising the entropic state [5].

The mathematical paradigm of RL is concerned with learning the best possible actions purely through positive and negative reinforcement or rewards. In CC, RL is in charge of ensuring that the entropic state is reduced after each cognitive cycle and establishing a policy in a particular environment that is driven solely by rewards [8]. The entropic reward is defined as the entropic-state decrement after each consecutive cycle. It can be anticipated using a Bayesian filter if the environment is modelled. Thus, learning and planning are two independent ideas in RL for CC, the former using real values of the entropic reward for a particular action and the latter using the predicted entropic reward from the Bayesian filter [8] [12]. It is worth noting that RL can only learn once for each PAC's chosen action; but, RL can prepare for arbitrary number of simulated future cycles and actions [5]. The number of actions that can be performed during planning is limited by considerations such as computing effort and cost, as well as time limits that need planning to be done before a single PAC finishes.

B. Related Works in Cognitive Control

1) Tracking Radar

The authors of [9] create a cognitive radar system and install a cognitive controller within it. The goal of the research is to show how powerful the cognitive controller's information processing capabilities are, as well as the system's potential tracking performance increases as a result. The literature provides details on the parameters of relevance, such as the measurement and system noise covariances, as well as relevant state-space and entropic-state models for the study's applicability. The cubature Kalman filter (CKF) [13] is used to estimate the perceptor's state covariance matrix, which is then utilised to compute the entropic-state.

The system is stated as having 382 distinct cognitive actions (the number of different transmit-waveform library combinations), each of which affects the measurement noise covariance matrix during a cycle of the PAC. Three different scenarios are investigated in the first trials; the first is the absence of CC on the system (fixed radar waveform). In the second situation, the cognitive controller learns but does not plan; the algorithm merely remembers the entropic reward values from the previous step. The third and final scenario introduces planning by using an explore-only method in the planning phase, which means the learning process is repeated for three different cases: the exploration of only one, two, or three random cognitive actions in each cycle of the PAC.

To reduce the impacts of randomness in the author's experimental simulations, 50 cycles were run across 1000 Monte Carlo runs [9]. Because the total number of cycles is substantially fewer than the number of possible cognitive activities (50 vs. 382) [9], performance in entropic-state

reduction in the scenario without planning is not much better than in the absence of CC. However, in the scenario with varying numbers of cognitive actions per cycle, it was discovered that even just one random cognitive action in the planning phase, which is a fraction of the total number of possible cognitive actions, is sufficient to demonstrate a four-order-of-magnitude improvement in entropic-state reduction [9]. In addition, the cases of two or three random cognitive actions showed the same drop in entropy but with faster convergence. In comparison to standard fixed waveform techniques, the planning process in CC was shown to dramatically improve the entropic-state of the model.

Further simulations in the study attempted to see how three distinct CC algorithms, such as dynamic optimization, Q-learning, and the authors' newly suggested algorithm, which combines Q-learning with learning and planning processes, affected the results. The suggested approach, which was configured to plan for three cognitive acts, outperformed both Q-learning and dynamic optimization in terms of minimising the entropic-state while also having a lower computational load [9]. The suggested technique achieved an entropic state value of $10^{0.4}$ in a 250-cycle trial, compared to about $10^{0.7}$ for both Q-learning and dynamic optimization [9]. Finally, the authors point out that while the Q-learning technique is computationally tractable, it can be inefficient in terms of performance. As a result, it may be useful to focus future research on developing algorithms designed specifically for this purpose.

2) Communication-Based Train Control

Communication-based train control (CBTC) systems are automated train control systems that use bidirectional train-ground wireless communications to assure the safe and efficient running of rail vehicles. These solutions aid in the better usage of railway network infrastructure while also improving customer service. However, there are concerns with train-ground communications and train control, which are usually treated as different topics in the literature.

Recent studies have explored combining many concerns into a single problem with the goal of overcoming each challenge using a CC-inspired technique, as in [14]. The authors employ the entropic state to objectively explain the packet delay and drop of information exchanged between the train-ground connection and the train control centre [14]. Wireless local area networks (WLAN) are widely employed in urban rail transit systems around the world as a medium for train-ground communication [15]. The linear-quadratic cost is utilised as a performance metric for train control, and Q-learning is then used to find the best policy based on this metric and the entropic state. In order to characterise high-speed railway and Rayleigh fading, the wireless channels are modelled as finite-state Markov chains with various state transition probability matrices. The CC model is in charge of ensuring that wireless communications and handoffs are reliable and uninterrupted, guaranteeing that the current train receives accurate information about the front train. As a result, the authors believe that adopting CC to increase communication between the train and the control centre will result in a more robust control of CBTC systems in terms of acceleration, deceleration, speed, distance, and emergency brake profiles [14].

With the proposed approach, more reliable velocity management between the system's front and back trains was proved through experimental trials with measurements retrieved from antennas on a train placed within a tunnel and subsequent MATLAB simulations [14]. When compared to alternative control policies like the semi-Markov decision process (SMDP) and greedy policies, which showed tiny disturbances in the difference in front and back train velocities, CC showed completely smooth and significantly safer behaviour. Furthermore, when compared to handoff delays of one second using the SMDP and greedy policies [14], the handoff delay with CC was significantly reduced to 0.2 seconds, half of the train response time parameter. Finally, while looking at the failure rate of the CBTC system under various policies, it is obvious that the CC approach proposed is the most effective due to its 99.78 percent availability. In comparison to the SMDP policy, which has a 10^{-2} unavailability rate, and the greedy policy, which has a 10^{-1} unavailability rate, this figure leads to unavailability rates of the order of 10^{-3} using CC [14]. Overall, the results demonstrate the usefulness of the proposed approach; however, the authors advise that more research is needed to investigate more advanced train-ground communication technologies, such as relaying, in order to increase the performance of CBTC systems.

3) Smart Grid Control

Oozeer and Haykin [16] suggest a CDS as a supervisor for smart grid networks utilising a CC method, as shown in Fig. 2. The authors present a new method for calculating the entropic state that is customised to the smart grid application, and they use it to create a control-sensing mechanism that can recognise and detect incorrect data from sensor measurements in the grid network. Bad measurements caused by erroneous readings, broken hardware components, or power system disruptions can result in a cascade of domino effects that obstruct the state estimation process and can degrade the performance of ordinary control systems [16].

The direct current (DC) state estimator is regarded as the environment in which the CDS acts in the author's suggested framework because it is the recipient of measurements in the network. To classify the observables from the environment, a generative model based on the cumulative sum (CUSUM) is used, followed by a Kalman filter (KF) filter to produce updated estimates for future cycles. The cognitive controller is thus in charge of learning and planning (as illustrated in Fig. 2), as well as providing the network with the ability to prioritise and disregard specific measurements for optimal state estimation by customising the weights assigned to each sensor or metre. By functioning in the opposite direction and independently of the PAC, the shunt cycles facilitate planning. With the use of the memory system in the perceptual and executive sections, this cycle engages both the perceptor and the executive to account for all planned prospective actions in each PAC [16]. The Bayesian Upper Confidence Bounds (Bayes-UCB) algorithm is used to optimise the system's newly tailored entropic state and provide a means for the cognitive controller to learn the best policy of actions, in this case, measurements weights, which are stored in working memory and applied to the system [16].

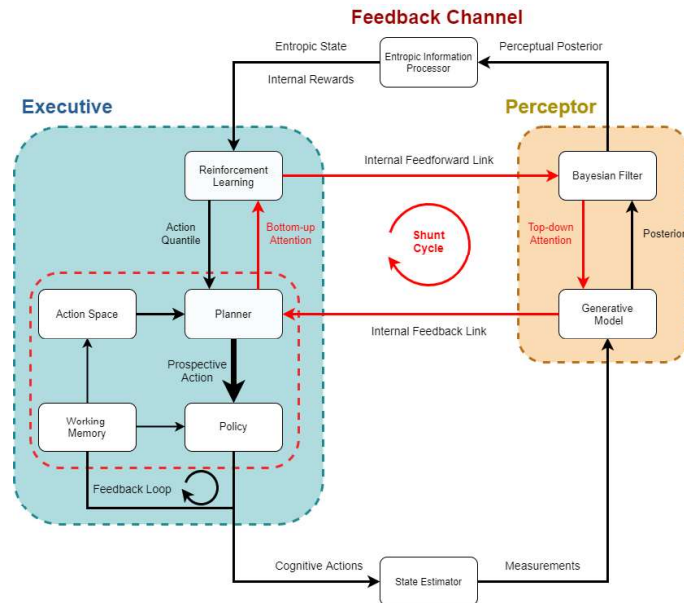


Fig. 2 Structure of a CDS with cognitive control as the supervisor of a smart grid network.

Experimental simulations on a 4-bus network are used to evaluate the suggested CC approach's performance in detecting and correcting erroneous data by changing measurement weights of various metres in the network. With CC, the system was shown to respond dynamically, selecting the optimum collection of metres to obtain readings from at the same time and efficiently assigning the best weight to each measurement for optimal state estimate [16]. Only a few PACs are required when a metre malfunctions in order to learn from the situation and adapt by lowering the weight of the faulty metre.

It is also demonstrated that the cognitive controller's weight assignment to the various measurements is done in such a way that it adapts to the probabilistic characteristics of noisy signals, and that the mean squared error (MSE) of the estimates obtained with the cognitive controller is much lower than without CC [16]. Furthermore, the authors demonstrate that when dealing with cyberattacks such as fake data injection assaults (FDI), the entropic state can be used as a metric to detect such attacks. However, it is noted that the model's structure must be expanded to include CRC in order to effectively deal with and reduce the risk associated with these types of attacks, which the authors address in later works [17] that will be discussed in section IV.

A disadvantage of the suggested framework in dealing with inaccurate measurement data is that it is not scalable to real smart grid networks, which typically have thousands of metres. The rationale provided is that performing an inverse calculation during state estimation is computationally expensive. We believe that machine learning techniques, notably a neural network, might be used to speed up processing and reduce the complexity of the approach. Regardless, the proposed model was more accurate, less prone to false positives, and cost less to compute than existing detection algorithms proposed in [18]. Finally, when scaling up to larger networks, the Bayes-UCB algorithm in the proposed CC model is expected to encounter challenges in terms of response time in determining optimal

configurations in the face of metre malfunctions [16]. In this situation, it may be possible to reduce the algorithm's response time by fine-tuning it and increasing the algorithm's sensitivity.

IV. COGNITIVE RISK CONTROL

A. Overview of Cognitive Risk Control

The CRC model adds a subsystem to the executive side of the CC model to allow for more complex reasoning, which necessitates the coining of a new term, the classifier, as shown in Fig. 3. In this illustration, the subsystem is configured in version II, which entails a disturbed cognitive action to the classifier rather than directly to the physical system. The executive memory, in turn, selects a collection of past acts or experiences for the classifier. There are also two pairs of switches, as indicated in Fig. 3, switches 1 and 2, and switches 3 and 4. Switches 1 and 2 are open in version II, preventing the controller from acting directly on the physical system and providing feedback to the executive memory. Instead, as previously stated, a perturbed cognitive action is delivered to the classifier, which is then in charge of making decisions by selecting a past experience that most closely fits the supplied perturbed cognitive action and then updating executive memory [19]. Switches 1 and 2 are closed, but switches 3 and 4 are opened, when the physical system is working without ambiguity, or under version I. In this case, the controller has direct access to the physical system and can change the executive memory.

The classifier's disrupted cognitive action is of probabilistic origin, and the executive memory's projected past events are similarly probabilistic because they are picked at random from its own action space. The Bayesian paradigm is used as a mechanism of decision-making in recognition of these truths, laying the stage for CRC [19]. For each of the past experiences in the specified set, the probability of the perturbed action's

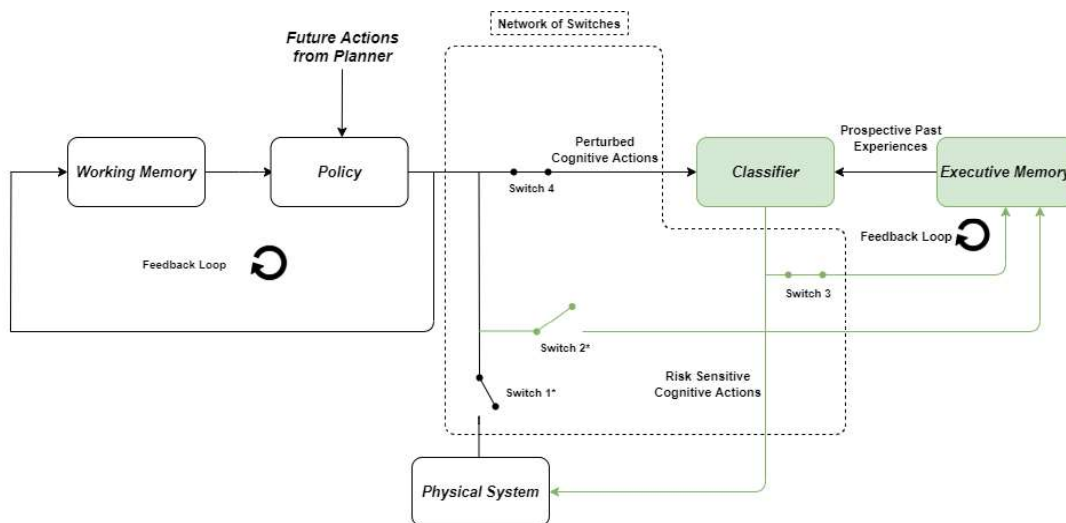


Fig. 3 Structure of the subsystem that is tasked with dealing with risk in the CRC architecture.

posterior given a past experience is determined using Bayes' rule. The risk-sensitive cognitive activity directed to the physical system is then defined as the experience with the highest likelihood [19].

Task-switch control (TSK) is a CRC framework function that exploits the presence of pairs of switches and controls their configuration depending on the presence or absence of uncertainty. The entropic rewards from the feedback channel serve as the foundation for defining TSC methods. The entropic reward in CRC can either be positive or negative, and it can never be zero. These two qualities are crucial in characterising TSC: positive rewards indicate the absence of uncertainty, while negative rewards imply uncertainty [19]. The choice of function to compute the rewards, as well as the tuning of design parameters in the chosen function, are essential considerations when applying CRC to physical systems using this approach. To summarise, when there are no uncertainties, the entropic reward must be positive, so switches 1 and 2 are closed, but switches 3 and 4 are open. When there are uncertainties, the entropic reward must be negative, therefore switches 1 and 2 must be opened and switches 3 and 4 must be closed.

B. Related Works in Cognitive Risk Control

1) Radar and Communications

Feng and Haykin published the first experimental research employing the CRC framework in [20]. The paradigm is studied and used in a cognitive vehicular radar system for self-driving cars by the authors. Recognizing the hazards posed to autonomous vehicles in the presence of uncertainty, the authors attempt to improve the performance of vehicular radar systems in such dangerous situations. The literature discusses the architectural structure of the CRC adapted to the problem of transmit-waveform selection in vehicular radar systems, as well as a simple vehicle-following scenario. A host vehicle is going forward in the stated scenario, and ahead of it is a target vehicle moving in the same direction, both of which are defined by their own velocities and accelerations. Details on state-space dynamics and modelling of the scenario are provided, and we refer the reader directly to the literature in [20] for these specifics. The purpose of the proposed model is to deal with

risky events caused by other physical entities robustly when applied as the supervisor for transmit-waveform selection in the radar system.

The authors remove the Bayesian generative model from the perceptual element of the CDS in their work because, in the case of automotive radars, observables are typically taken in a fashion that can be directly processed by the Bayesian filter [20]. As a result, the Bayesian filter has been relocated to the bottom of the perceptor, and the entropic-information processor has been added to take its place and preserve the feedforward link. Aside from that, the suggested work follows the same structure as Fig. 4. The KF is chosen as the Bayesian filter to represent the vehicle-following scenario, and it is formulated according to the transmit-waveform option, which mixes the linear frequency modulated (LFM) waveform with Gaussian amplitude modulation. Invoking Shannon's information theory [20], the entropic state is determined using the filtered posterior from the KF as input. The entropic state uses a defined function to determine entropic or internal incentives, which it subsequently passes on to the executive. The internal incentives are sent through a defined function in the CRC framework's TSC mechanism, which is then subjected to particular conditions and thresholds formulated in the literature to determine the existence or absence of uncertainty. The rest of the methodology follows the standard CC framework described in earlier sections, which is also depicted in Fig. 4 by the red dashed boundary.

The suggested CRC model with Q-learning for RL was compared to alternative systems, such as a radar with fixed transmit-waveform (FTW), the CC framework, and merely Q-learning on its own for waveform creation, using experimental simulations [20]. The root mean squared error (RMSE) was calculated against each model's five states: the velocity and acceleration of the host vehicle, the longitudinal distance between the host and target vehicle, and finally the velocity and acceleration of the target vehicle. Based on simulation findings, the suggested model has the lowest RMSE for each state in the depicted qualitative graphs, with CC and Q-learning performing similarly [20]. These findings show that regardless of the algorithm used, the executive's learning algorithms will

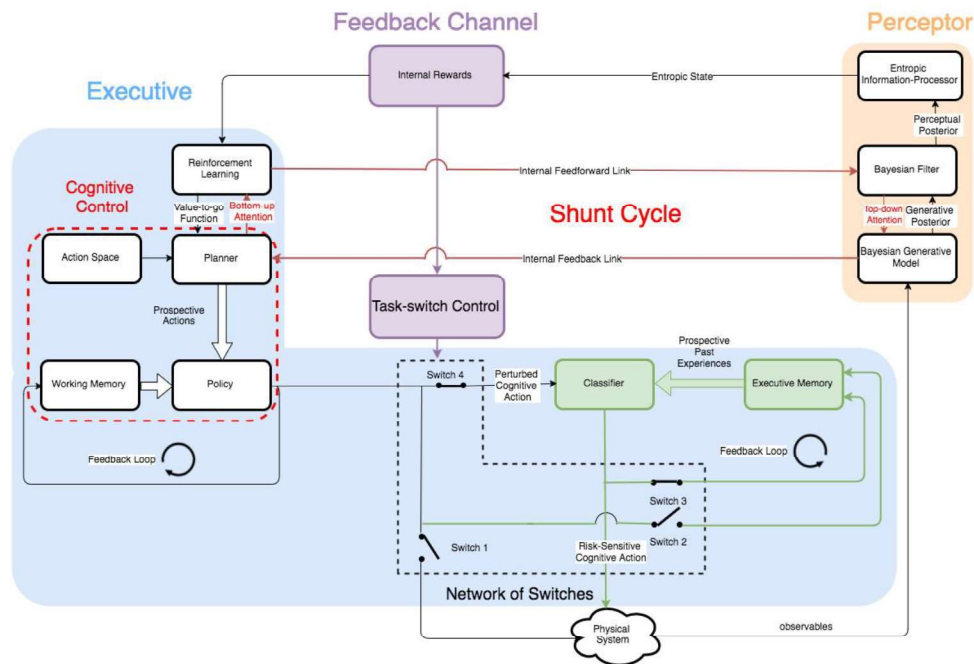


Fig. 4 Structure of the CRC framework. Green coloured elements represent the newly introduced risk-sensitive subsystem.

result in better decision-making and action choices. During the trial, the authors add a structural uncertainty term to the system model to create a dangerous scenario in the experiment, which lasts less than a second. In this circumstance, the Q-learning and CC algorithms were unable to adjust to the uncertainty, resulting in substantial spikes and erratic behaviour in the RMSE, which took upwards of eight seconds to recover from. However, the proposed CRC model relative to the other approaches was only slightly affected in terms of RMSE and recovered within a matter of two or three seconds at most. Overall, the model achieved impressive results and was also deemed a promising alternative to traditional approaches in handling uncertainties and risky events in vehicular radar [20].

These research have culminated in the most current work of Feng and Haykin in [21], in which the authors propose merging cognitive vehicular radar (CVR) and vehicle to vehicle (V2V) communications with a coordinated CRC (C-CRC) model that bridges both systems together. C-CRC investigates the benefits of mutual aid by utilising information emanating from one of the systems that may be useful to the other. In addition, unlike earlier investigations, a nonlinear target-tracking model is used, and the analysis is done with a CKF [21]. The formula for the interference measure in V2V communications in this approach includes tracking findings as well as other practical aspects inferred from those results, such as vehicle motion and channel availability. The CVR also relies on communication system data to determine whether it is operating on a one-vehicle or two-vehicle model, with the latter indicating the presence of a second vehicle engaged in target tracking [21]. Each system in the C-CRC model has its own TSC mechanism and is implemented with a risk-sensitive subsystem. Furthermore, the TSC in each system plays a part in determining what information should be transferred from one system to the next for their dual system.

The authors show that the suggested C-CRC model outperforms previous radar techniques such as FTW and Q-learning in minimising the peak RMSE achieved in tracking longitudinal distance when faced with uncertainty by up to 70% and 67 percent, respectively. Although the typical CRC design in [22] had equivalent performance, it was still 41 percent worse to the C-CRC in terms of RMSE peak reduction. C-performance CRC's is improving across the board, including tracking performance in terms of the utility of power selection in vehicle and jammer communications, total regret from channel selection, and finally, user utility.

However, the ability of V2V communications to keep up with busy networks in specific locations or conditions has been noted in the literature. With less spectrum opportunity, user utility decreases while jammer utility grows, according to further examination of studies relating to the effects of channel availability on power and channel selection [21]. This scenario also leads to a greater regret measure for the host vehicle, a reduced multi-armed bandit (MAB) related reward, and increased channel switching costs. As a result, vehicle networks with several entities sharing available wireless resources in a local area present fascinating and practical V2V performance issues that demand additional research. The authors also highlight that security vulnerabilities in large-scale adversarial CAV networks would be investigated in the future as part of their research efforts.

2) Cybersecurity in Smart Grids

In [17], Oozeer et al. improved on their CC technique for smart grid attack detection from [16] by proposing an upgraded CRC-enabled model capable of also defending against such assaults. The entropic state was utilised as a metric in the initial experiments to detect the existence of FDI attacks, which was signified by the entropic state dropping below a predetermined

threshold, setting the stage for TSC in the extended CRC version of the model. When an FDI assault is detected and TSC is triggered in the extended framework, the cognitive controller is deemed inactive, while CRC is activated to protect against these attacks [17].

The authors note that the action space involved in this scenario involving CRC differs from CC, recognising that FDI attacks try to produce deviation in specific states to trigger a cascade of incorrect control decisions. Unlike the cognitive controller, which has an action space of possible measurement weights, CRC includes picking tuning parameters to be applied to the DC system's configuration matrix [17]. However, if an assault has been discovered, the predictor or classifier must first recognise the states that are at risk. As explained in the literature [17], the affected states are recognised by whether they exceed the maximum deviation allowed by a formulation based on each estimate's mean recorded in the perceptual memory. Following that, once the attacker states have been identified, the planner must carefully pick tuning parameters in the columns of the DC system's configuration matrix corresponding to the impacted states without interrupting the estimation of other states [17]. Each shunt cycle is dedicated to resolving the hazards associated with one of the states at a time during this planning phase of operation, and a new reward connected with a specific action in the cycle is determined. The Bayes-UCB algorithm, as suggested in [16], is then in charge of optimising the policy in such a way that it prioritises actions that will return current attack states to a condition that is closest to past perceptual memories. Similar to CC, the actions that get the highest quantile from the Bayes-UCB algorithm are stored in the working memory and applied once the shunt cycles have expired. Once the impacted states have been restored to acceptable levels, the risk is considered controlled, and no further actions will be taken in those columns of the system configuration matrix. Finally, once the attacks have been determined as having finished, a mechanism is implemented by supplying the TSC with memory and a watchdog timer that restores the system configuration to its original state and marks the end of the current PAC [17].

The experimental simulations used in the author's research are comparable in configuration to those used in their prior investigations [16], which used IEEE 4-bus and 14-bus networks. The literature shows how the cognitive controller and CRC can operate together in the 4-bus network to bring FDI attacks under control once they've been introduced to the system. The network configuration matrix of the system is detailed in the study along with other pertinent parameters, and it is mentioned that the simulations run for a total of 2000 seconds while allowing for 15 shunt cycles in each PAC for learning and planning. The action space for CRC consists of 63 different tuner values, each of which can be tuned with a specific range of values for relevant columns of the network configuration matrix. In the 4-bus network, three states are measured, and an assault is launched on the first two states 1000 seconds into the simulation, lasting 300 seconds. The phase angles of the desired states are shifted by predefined values to imitate FDI attacks [17]. Before the attacks, the CRC only needs 20 cycles to get the estimations or measurements for the impacted states under control and restored to a tolerable

threshold [17]. When the attacks are over, the suggested model continues to run under CRC for another 39 cycles, according to the authors, because the model ensures that conditions for matching current and previous events are met. The altered network configuration matrix is then restored to its original condition before TSC and CRC are triggered.

However, one of the proposed studies' disadvantages is that when other forms of FDI attacks, such as the slowly evolving ramp attack, are used, the detection time can be altered and increased [17]. Furthermore, scaling up the architecture to larger and more realistic networks will necessitate more shunt cycles in each PAC, creating major processing resource and efficiency difficulties. This problem is exacerbated by the fact that other types of FDI attacks, such as developing ramp attacks, necessitate a longer sample time for the DC estimator to overcome them. Another use of this CDS that the author's investigations have not looked into is not just identifying attacked states, but also identifying which sensors or metres have been attacked. Finally, there is a suggestion in the literature that applying a predetermined threshold on the absolute estimated error of the estimated readings might be used to identify the attacked metres in a network [17].

V. CONCLUDING REMARKS

Cognitive dynamic systems, as well as their two particular functions, cognitive control and cognitive risk control, were comprehensively examined in this study. This work is the first attempt to compile data from the voluminous literature published in this young and evolving topic. The goal of this study's methodology was to encourage and facilitate additional research into cognitive dynamic systems. We did so by alerting the reader about advancements in each specialized field, outlining the benefits and limits of the surveyed literature, and providing suggestions and directions for future research. Finally, the contents and outcomes of this survey will serve as a foundation for future research, and will hopefully be useful to other academics working in this fascinating new subject.

VI. REFERENCES

- [1] S. Haykin, *Cognitive Dynamic Systems: Perception-action Cycle, Radar and Radio*, Cambridge, UK: Cambridge University Press, 2012.
- [2] S. Haykin, "Cognitive Dynamic Systems," *International Journal of Cognitive Informatics and Neural Intelligence*, vol. 5, no. 4, pp. 33-43, 2011.
- [3] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201-220, 2005.
- [4] S. Haykin, "Cognitive radar: a way of the future," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30-40, 2006.

- [5] S. Haykin, M. Fatemi, P. Setoodeh and Y. Xue, "Cognitive Control," *Proceedings of the IEEE*, vol. 100, no. 12, pp. 3156-3169, 2012.
- [6] S. Haykin, "The Cognitive Dynamic System for Risk Control [Point of View]," *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1470-1473, 2017.
- [7] S. Haykin and J. M. Fuster, "On Cognitive Dynamic Systems: Cognitive Neuroscience and Engineering Learning From Each Other," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 608-628, 2014.
- [8] S. Haykin, A. Amiri and M. Fatemi, "Cognitive control in cognitive dynamic systems: A new way of thinking inspired by the brain," in *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, Orlando, FL, 2014.
- [9] M. Fatemi and S. Haykin, "Cognitive Control: Theory and Application," *IEEE Access*, vol. 2, pp. 698-710, 2014.
- [10] S. Haykin, "Cognitive Dynamic Systems," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1910-1911, 2006.
- [11] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, 1948.
- [12] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 2018.
- [13] I. Arasaratnam and S. Haykin, "Cubature Kalman Filters," *IEEE Transaction on Automatic Control*, vol. 54, no. 6, pp. 1254-1269, 2009.
- [14] H. Wang, F. R. Yu, L. Zhu, T. Tang and B. Ning, "A Cognitive Control Approach to Communication-Based Train Control Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1676-1689, 2015.
- [15] L. Zhu, F. R. Yu, B. Ning and T. Tang, "Cross-Layer Handoff Design in MIMO-Enabled WLANs for Communication-Based Train Control (CBTC) Systems," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 719-728, 2012.
- [16] M. I. Oozeer and S. Haykin, "Cognitive Dynamic System for Control and Cyber-Attack Detection in Smart Grid," *IEEE Access*, vol. 7, pp. 78320-78335, 2019.
- [17] M. I. Oozeer and S. Haykin, "Cognitive Risk Control for Mitigating Cyber-Attack in Smart Grid," *IEEE Access*, vol. 7, pp. 125806-125826, 2019.
- [18] R. Xu, Z. Guan, L. Wu, J. Wu and X. Du, "Achieving Efficient Detection Against False Data Injection Attacks in Smart Grid," *IEEE Access*, vol. 5, pp. 13787-13798, 2017.
- [19] S. Haykin, J. M. Fuster, D. Findlay and S. Feng, "Cognitive Risk Control for Physical Systems," *IEEE Access*, vol. 5, pp. 14664-14679, 2017.
- [20] S. Feng and S. Haykin, "Cognitive Risk Control for Transmit-Waveform Selection in Vehicular Radar Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9542-9556, 2018.
- [21] S. Feng and S. Haykin, "Coordinated Cognitive Risk Control for Bridging Vehicular Radar and Communication Systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-16, 2020.
- [22] S. Feng and S. Haykin, "Cognitive Risk Control for Anti-Jamming V2V Communications in Autonomous Vehicle Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9920-9934, 2019.
- [23] S. Soatto, "Actionable Information in Vision," in *IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009.
- [24] R. A. Fisher, "On the mathematical foundation of theoretical statistics," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 222, pp. 309-368, 1922.
- [25] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed., New York, NY: Springer-Verlag, 2008.
- [26] J. J. Gibson, "The Myth of Passive Perception: A Reply to Richards," *Philosophy and Phenomenological Research*, vol. 37, no. 2, pp. 234-238, 1976.
- [27] J. J. Gibson, *The Ecological Approach to Visual Perception*, Classic ed., New York, NY: Psychology Press, 2014.
- [28] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139-154, 2009.
- [29] P. Dayan and Y. Niv, "Reinforcement learning: The Good, The Bad and The Ugly," *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 185-196, 2008.
- [30] D. J. Surmeier, J. Plotkin and W. Shen, "Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection," *Current Opinion in Neurobiology*, vol. 19, no. 6, pp. 621-628, 2009.
- [31] R. Poovendran, K. Sampigethaya, S. K. S. Gupta, I. Lee, K. V. Prasad, D. Cormann and J. L. Paunicka, "Special issue on cyber-physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 6-12, 2012.
- [32] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah and C. S. Hong, "Caching in the Sky: Proactive Deployment of Cache-Enabled Unmanned Aerial Vehicles for Optimized Quality-of-Experience," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1046-1061, 2017.
- [33] B. Paden, M. Čáp, S. Z. Yong, D. Yershov and E. Frazzoli, "A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33-55, 2016.

- [34] S. Feng and S. Haykin, "Cognitive Dynamic System for Future RACE Vehicles in Smart Cities: A Risk Control Perspective," *IEEE Internet of Things Magazine*, vol. 2, no. 1, pp. 14-20, 2019.
- [35] S. Feng and S. Haykin, "Anti-Jamming V2V Communication in an Integrated UAV-CAV Network with Hybrid Attackers," in *IEEE International Conference on Communications*, Shanghai, China, 2019.
- [36] S. Feng and S. Haykin, "V2V Communication-Assisted Transmit-Waveform Selection for Cognitive Vehicular Radars," in *IEEE Canadian Conference of Electrical and Computer Engineering*, Edmonton, AB, 2019.
- [37] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235-256, 2002.
- [38] J.-Y. Audibert, R. Munos and S. Csaba, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876-1902, 2009.
- [39] A. Humayed, J. Lin, F. Li and B. Luo, "Cyber-Physical Systems Security—A Survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802-1831, 2017.
- [40] X. Yu and Y. Xue, "Smart Grids: A Cyber-Physical Systems Perspective," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1058-1070, 2016.
- [41] G. Liang, S. R. Weller, J. Zhao, F. Luo and Z. Y. Dong, "The 2015 Ukraine Blackout: Implications for False Data Injection Attacks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317-3318, 2017.
- [42] Y. Liu, P. Ning and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, pp. 1-33, 2011.
- [43] J. Hao, R. J. Piechocki, D. Kaleshi, W. H. Chin and Z. Fan, "Sparse Malicious False Data Injection Attacks and Defense Mechanisms in Smart Grids," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 5, pp. 1-12, 2015.
- [44] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor and A. Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 106-115, 2012.
- [45] W. Sun, F. R. Yu, T. Tang and S. You, "A Cognitive Control Method for Cost-Efficient CBTC Systems With Smart Grids," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 568-582, 2017.
- [46] X. Wang, L. Liu, T. Tang and W. Sun, "Enhancing Communication-Based Train Control Systems Through Train-to-Train Communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1544-1561, 2019.
- [47] L. Lei, J. Lu, Y. Jiang, X. S. Shen, Y. Li, Z. Zhong and C. Lin, "Stochastic Delay Analysis for Train Control Services in Next-Generation High-Speed Railway Communications System," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 48-64, 2016.
- [48] M. L. Tuballa and M. L. Abundo, "A review of the development of Smart Grid technologies," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 710-725, 2016.
- [49] R. Deng, J. Chen, X. Cao, Y. Zhang, S. Maharjan and S. Gjessing, "Sensing-Performance Tradeoff in Cognitive Radio Enabled Smart Grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 302-310, 2013.
- [50] X. Fang, Y. Han, J. Wang and Q. Zhao, "A cognitive control approach for microgrid performance optimization in unstable wireless communication," *Neurocomputing*, vol. 355, pp. 168-182, 2019.

Balancing Sampling Frequencies for Multi-modality IoT Systems: Smart Shoe as an Example

1st Wenwu Deng

Dept. of Computer Science and Technology
University of Science and Technology of China (USTC)
 Hefei, China
 dengwenw@mail.ustc.edu.cn

2nd Kangyu Chen

Department of Cardiology
The First Affiliated Hospital of USTC
 Hefei, China
 ahslyycky@126.com

3rd Qijun Ying

Dept. of Data Science
University of Science and Technology of China
 Hefei, China
 yqj@mail.ustc.edu.cn

4th Jingyuan Cheng

Dept. of Computer Science and Technology
University of Science and Technology of China
 Hefei, China
 jingyuan@ustc.edu.cn

Abstract—With the rapid development of Internet of Things, more and more sensing modalities are being integrated into single small devices. The data throughput, however, is always limited due to factors like transmission bandwidth and storage. How to balance the sampling frequencies among different sensing modalities, is little studied. Given a concrete system and its application, the question becomes: 1) how to qualitatively evaluate the loss of information as the sample frequency drops? 2) how to find the optimal sampling frequency ratio and set the multi-modality system according to it? In this paper, we will provide a general workflow to answer these questions. The workflow is evaluated on a self-developed smart-shoe, equipped with one pressure sensing matrix and two Inertial Measurement Units. This workflow could be interesting to both multi-modality IoT system developers and users, support them in finding the best frequency setting and obtaining datasets of higher qualities.

Index Terms—System Optimization, Multi-Modality Sensing, Data Quality, Embedded Systems

I. INTRODUCTION

With the rapid development of sensors and integrated circuits, more and more sensing modalities are being brought into small devices around us. While more sensing modalities enable sensor fusion, they play also a zero-sum game: the overall data throughput is always limited by for example the transmission bandwidth, while all individual sensors want a higher share in the overall throughput to maintain the high-frequency components in its own data.

The Whittaker-Shannon sampling theorem points out that an analog signal can be perfectly recovered as long as the sampling frequency is at least twice as large as the highest-frequency component of the analog signal to be sampled. For systems containing multiple sensing modalities, it is impossible to push at the same time all the sampling frequencies

to their corresponding maxima. It is a common practice to put the sampling frequencies of all the sensors to the same. However, is the 1:1 ratio really the best option for saving the most information? We will show in section IV-B with a concrete example, viz. a smart-shoe, equipped with one textile pressure sensing matrix (TPM) and two Inertial Measurement Units (IMU), that this common practice might not be the optimal: with the 1:1 setting, the overall information loss for IMU is as high as 15% (measured by Cramer-Von Mises test) and for TPM 2%. After frequency balancing, when the frequency ratio is set to 2:1 (IMU over TPM), the information loss of IMU is enhanced to only 2%, while that of TPM remains 2%.

For people who design or use such multi-modality sensing systems, two questions might then come up:

- 1) Do I also need frequency balancing, or not?
- 2) If yes, what is the optimum frequency ratio?

This paper will provide answers to these questions. The contribution lies in:

- 1) We demonstrate with a concrete example that the commonly practiced 1:1 sampling frequency ratio is not optimal if the goal is to keep as much information as possible within limited data throughput.
- 2) We design a general workflow to support multi-modality IoT system developers in finding the optimal ratio and modifying the embedded system's configuration.
- 3) We demonstrate how the workflow works with a concrete example, namely a smart-shoe, featured with two Inertial Measurement Units and a pressure matrix. This is also a typical hardware setting with both discrete multi-channel sensors and a sensing matrix.

This paper could be interesting to both IoT system developers and users, promote their second thought before carrying out large-scale data acquisition, and support them in finding

This work is supported by "the Fundamental Research Funds for the Central Universities" (Grant No. 2150110020).

out the optimal frequency configuration to obtain datasets of higher qualities.

II. RELATED WORK

A. Multi-modality IoT systems

The amount of multi-modality IoT systems are growing together with the need on information. A lot of efforts are given on the sensor level, done by material scientists and hardware developers, including new materials [1], new sensing modalities [2], [3], physical and electronic modeling [4] and better ADCs [5]. The result are IoT systems with more and more sensing modalities. We list a few in Table I, focusing on wearable systems in healthcare applications. There are far more such systems, but the fact is, little thought is given on how to balance the sensing frequencies. The 1:1 ratio is a common practice, because it ensures that data from all sensors are synchronized, which eases the sensor fusion algorithms. Whether this is the optimal ratio, stays with a question mark.

B. Sampling Frequency Considerations

To the best of our knowledge, there was no research on improving information quality by balancing sample frequencies of multiple sensors and/or sensing matrices. We thus go for the most related research category: finding the best frequency of a single sensor/sensing module, mainly for the purpose of reducing power consumption.

Inertial Measurement Unit is widely used in wearable systems. Its sampling frequency is studied for activity recognition applications. Taleb et al. [11] used an entropy-based optimization algorithm to derive the optimal sampling frequencies of the accelerometer in the cell phone for different activities, the conclusion was sitting (10Hz), walking (70Hz), jogging (90Hz). Allik et al. [12] explored the effect of sampling frequency on classification performance. By downsampling triaxial accelerometer data they concluded that there is no

significant difference in classification performance for most activities with sampling frequencies at 50 Hz, 25 Hz and 13 Hz, only that 13 Hz is not sufficient to capture vibrations from outdoor riding. Khan et al. [13] used two-sample K-S test to derive its optimal frequency for activity recognition based on 5 read-to-use public datasets. The conclusion was 45Hz. Anish et al. [14] presented a sensor-classifier co-optimization technique for wearable devices using human activity recognition as a driver application, they dynamically power down the accelerometer and lower the sampling frequency when the user is performing low-intensity activities. Using these optimizations, the proposed approach achieves up to 49% reduction in total platform energy consumption with less than 1% decrease in the accuracy.

Another type of sensing modality that was studied is the Force Myography (FMG). By downsampling the data and analyzing the RMSE (Root Mean Square Error), Xiao et al. [15] suggested that FMG at the forearm and wrist should sample at minimum 54 Hz and 58 Hz for deciphering isometric actions, and 70 Hz and 84 Hz for deciphering dynamic actions. Lei et al. [16] studied 16-channel FMG, and concluded that for recognizing static postures, the frequency can be further reduced to 5 Hz.

While we can borrow the validation methods from these papers, none of them considered multiple sensing modalities in the same system. In such systems, each sensor occupies a part from the overall system resources (CPU time, bandwidth, and etc.), the question changes from how to save system resource given the optimum recognition goal, to how to balance the frequency setting to lose less information with the limited, predefined resource.

III. THE EXAMPLE: A MULTI-MODALITY SMART SHOE

Before going into the general workflow for balancing multi-sensors' frequencies, we first introduce the example system:

TABLE I
MULTI-MODALITY IOT SYSTEM EXAMPLES

Work	Sensing modalities	Application	Note
[6], 2005	Electrocardiogra(ECG) sensors (100Hz) Piezoresistive sensors (16Hz)	A fabric sensing system for monitoring individuals affected by cardiovascular diseases	Arbitrary sampling frequencies
[7], 2009	ECG sensors (200Hz) Accelerometer (200Hz)	A smart shirt for continuous health monitoring	Frequency ratio set to 1:1
[8], 2015	Accelerometer (100Hz) Gyroscope (100Hz) Temperature (1Hz) Humidity (1Hz) Barometric air pressure (5Hz)	A wrist-worn device for recognizing very fine-grained and complex inhome activities of human users	Arbitrary sampling frequencies
[9], 2020	Audio (46.875kHz) Electrical bioimpedance (0.02Hz) Accelerometer (250Hz&100Hz) Gyroscope (100Hz) Temperature (1Hz)	A wearable, multimodal sensor brace for assessing knee joint health	Accelerometer at different sampling frequencies
[10], 2022	Pressure array (30Hz) Accelerometer (30Hz) Gyroscope (30Hz)	A smart shoe for monitoring gait	Frequency ratio set to 1:1

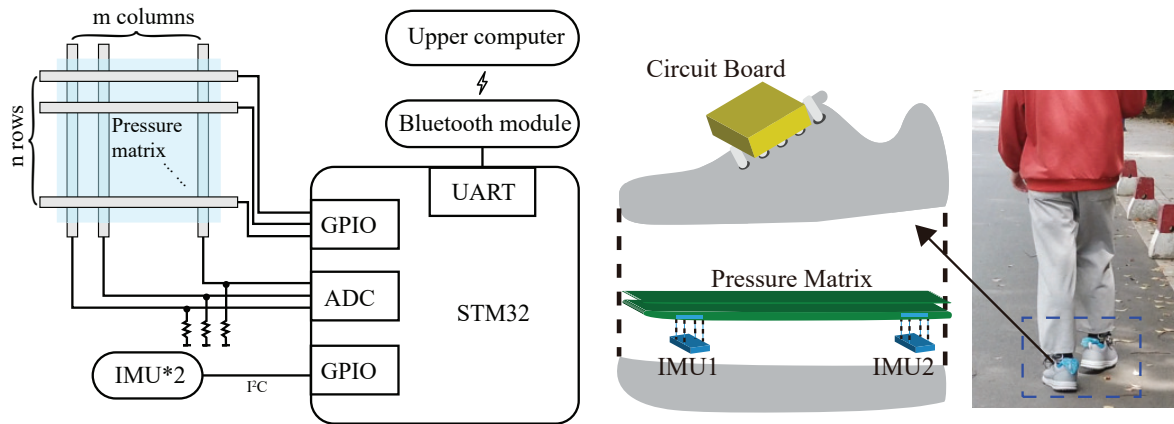


Fig. 1. The smart shoe: driving architecture (left) and sensor configuration (right)

a pair of smart shoes developed by our lab for gait analysis, as shown in Fig. 1. In the later examples, only the right shoe data are used and the system is thus mentioned only as "the shoe". It features two IMU's (we use only accelerometers and gyroscopes in the study, thus 6 channels each, resolution 16 bits) and one pressure sensing matrix (28 × 16 channels, resolution 12bits), all integrated into a single sole and put into the shoe. One IMU is put to the forefoot position and the other to the heel. The pressure sensing matrix covers the whole planta. The data are streamed to a phone or PC via Bluetooth with a maximum bandwidth of 48KB/s, resulting in the sample rate of 68Hz, shall the 1:1 ratio be adopted.

Signals from the shoe are demonstrated in Fig. 2. A health adult wears a pair of smart shoes and walks in a straight line. We first record the data with the TPM turned on and IMUs off, then with the TPM turned off and the IMUs on. This is to push the sampling frequency of each sensing modality to its maxima. It can be seen that the two sensors and the matrix obtain different types of information. While the IMUs are good at capturing the in-air posture, the TPM is good at capture on-ground force distribution. Even the two IMUs demonstrate detailed difference of the forefoot and the heel, as our feet are not as stiff as wood and its shape changes throughout the whole gait cycle. That is to say, we need all sensors for monitoring the gait details. Turning off one IMU means losing quite much information and is not recommended. Also there exist high-frequency components in the signal, demonstrated by the sharp rising and falling edges. That is to say, lower sampling frequency would very likely mean losing more high frequency details.

Here comes the zero-sum game: all sensing modalities want a higher sampling frequency. How to balance the sample frequencies is thus an important question to be answered.

IV. TOWARDS THE BALANCED FREQUENCY SETTING

A. The General Workflow

Given a multi-modality system and its application domain, a pre-experiment on data quality and the corresponding fre-

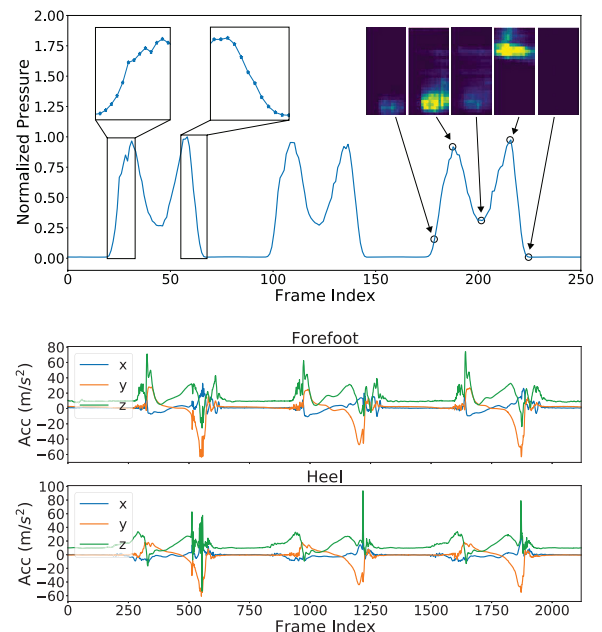


Fig. 2. Smart shoe data from three steps, TPM data (up), two IMU data (down)

quency balancing is recommended before creating the large-scale dataset. Certain amount of information might get lost already during the dataset creation process and the only way to get them back is to repeat the experiment, which normally takes much more efforts, time and human resource than the simple pre-experiment. Below we provide the general workflow and concrete methods, including:

- 1) *Evaluate information quality*: perform a small pre-experiment, with one sensing modality turned on at a time and all the others turned off. The data are then downsampled and compared with the original data to qualify the information loss as frequency drops.
- 2) *Select the optimal frequency ratio*: Given a preset

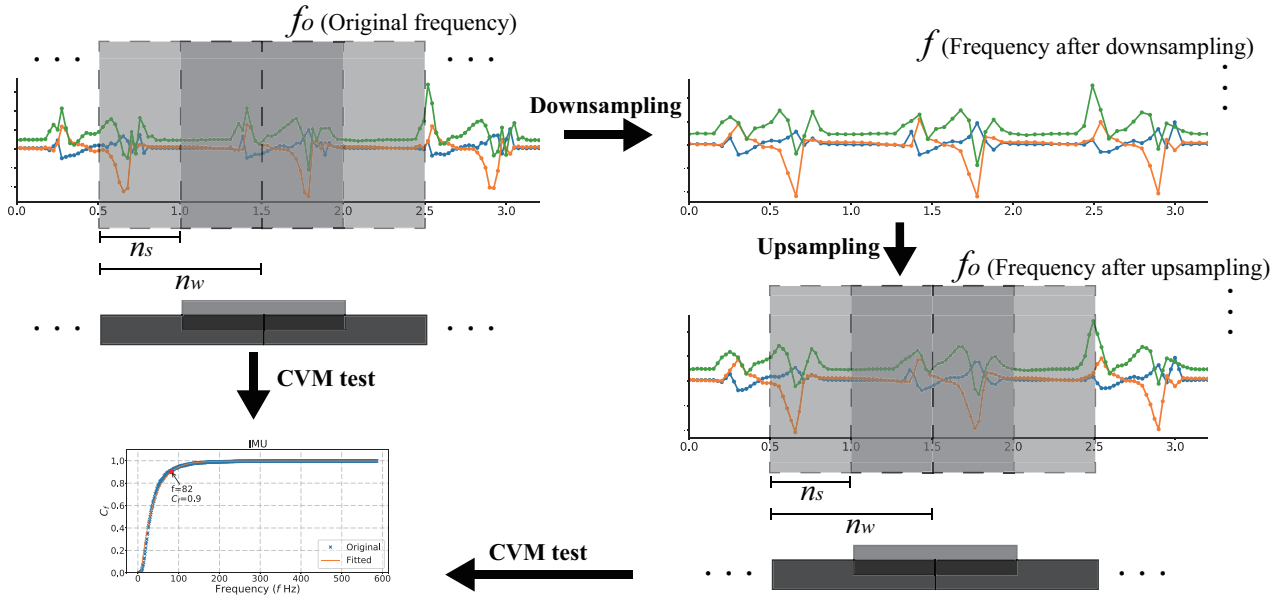


Fig. 3. Data test process (IMU as an example)

frequency ratio, calculate or measure the frequency of each sensing modality. Based on how much information loss can be endured, find the lowest frequency border of each sensing modality. This defines the common space, where the optimal frequency ratio can be found.

- 3) *Configure the embedded system*: Given the found optimal frequency ratio, configure the system. This step is system dependent. We will demonstrate how to do it by modifying the sample process.

This workflow requires no pre-knowledge on signal quality and is not limited to the domain of activity recognition, where the recognition rate can be the measure for information loss.

B. Evaluate Information Quality

Given a multi-modality system, the first step is to find out whether the default frequency ratio (e.g. the commonly practiced 1:1 setting) is good enough for keeping information. We introduce the following approach (also shown in Fig. 3).

The original data are downsampled to frequency f and upsampled back to f_o , using linear interpolation. The original data and the upsampled data are segmented using the sliding window approach. Cramer-Von Mises (CVM) Test [17] is performed on each segments and the results are averaged. The CVM test statistically measures the integral of the squared distance between the empirical cumulative distribution function and the target cumulative distribution function. It is more suitable for demonstrating the overall difference, compared to the K-S test [13], which measures the largest distance between two cumulative distributions. The averaged CVM test result (C_f as defined in Eq. 2) is a value between 0 and 1, where 1 means no information loss. The concrete mathematic expressions are given below, taking IMU data as the example.

First define all the variables:

- N : number of samples;
- d : dimension index, $i \in [1, 12]$;
- j : sample index, $j \in [1, N]$;
- f_o : original sampling frequency;
- f : frequency after downsampling ($f < f_o$);
- n_w : window size;
- n_s : step length;
- W : total number of windows, $W = \lceil (N - n_w) / n_s \rceil$;
- i_w : window index, $i_w \in [0, W - 1]$;
- j_w : $j_w = i_w \times n_s$, each window's starting index.
- $D_o(d, j)$: the dataset, at the original frequency f_o ;
- $D_u(d, j, f)$: the dataset, after downsampling to frequency f , and upsampling back to frequency f_o ;

A two-sample CVM test is performed for each pair of windows, the returned p-value from Python function `scipy.stats.cramervonmises_2samp` is taken as the confidence of CVM test (definition of p-value in [18]):

$$p(d, i_w, f) = \text{cramervonmises_2samp}([D_o(d, j_w), D_o(d, j_w + 1), \dots, D_o(d, j_w + n_w - 1)], [D_u(d, j_w, f), D_u(d, j_w + 1, f), \dots, D_u(d, j_w + n_w - 1, f)]) \quad (1)$$

Finally, the test values of each window are averaged to obtain the CVM test value for one axis of data, and then the 12 axes of data are averaged to obtain the final test value C_f :

$$C_f = \frac{1}{12W} \sum_{d=1}^{12} \sum_{i_w=0}^{W-1} p(d, i_w, f) \quad (2)$$

With $C_f \in [0, 1]$, when $C_f = 0$, it means that the two samples are not similar at all, while when $C_f = 1$, it means

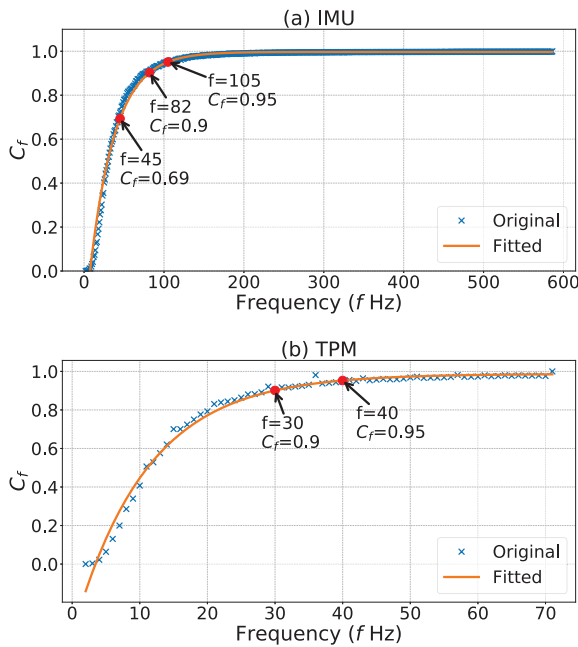


Fig. 4. CVM test result, (a)IMU, (b)TPM

that the samples are identical, and the closer C_f is to 1 the higher the similarity of the two samples' data. The TPM data test process is similar and only requires replacing the 12-axis IMU data with 448-point pressure data.

We performed the CVM test on the smart shoe data with the following strategy:

- *Two IMUs*: All IMU channels are considered equally important, CVM test is performed on each channel and the average is calculated (result given in Fig. 4 a).
- *Pressure matrix*: CVM test is performed on each channel (28×16 in total) and the average is calculated (result given in Fig. 4 b).

Other options remain, for example, calculate the CVM test for the overall acceleration and angular speed ($2 C_f$'s for each IMU), then calculate the average; calculate the CVM test for the overall weight of the TPM. Weight can also be given to certain sensing modalities and/or channel(s) to emphasize its/their importance. For example, for indoor navigation, IMU sampling frequency plays an important role, while the pressure matrix is only an aiding tool, then the IMU can be given to a higher weight; in gait analyses, to emphasize how the foot strikes the ground, a higher weight can be given to the z-axis acceleration of the IMU on the heel.

We use 0.90 and 0.95 as the C_f thresholds to decide where to set the proper sampling frequency (for CVM test, 1 means that the two distributions are identical). To overcome the small glitches (which is more visible on the TPM result in Fig. 4 b)), the curves are fitted. This ends up with 82 Hz and 105Hz for IMUs, and 30Hz and 40Hz for the TPM.

It is specially worthwhile to check the IMU result. While Khan A et al. [13] studied 5 datasets and draw the conclusion

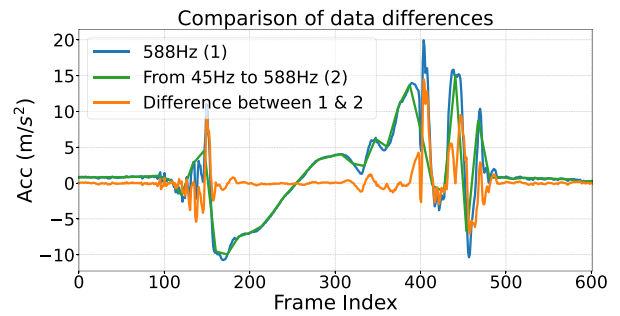


Fig. 5. Comparison of the difference in data between the right forefoot x-axis acceleration of 45hz and 588hz, (1) is the original data (sampled at 588Hz), (2) is the recovered data (downsampled to 45Hz then back to 588Hz)

that 45Hz is proper for activity recognition, our analyze demonstrates that this is definitely not a good choice for obtaining gait details. In Fig. 4 a), the C_f drops to 0.69, indicating significant information loss. To demonstrate the distortion, we provide in Fig. 5 the original data (acceleration on x-axis sampled at 588Hz), the recovered data (downsampled to 45Hz then back to 588Hz) and their difference. The amplitude difference can be as high as 70% of the original signal. As IMU is widely used in IoT systems, it is worthwhile for the developers to take a second thought on its sampling frequency.

If the common practiced 1:1 frequency ratio is taken, then the shoe will sample at 68Hz (detailed analysis in section IV-C). For IMU, C_f drops to 0.85 and for the TPM, C_f drops to 0.98. The IMU data are still seriously distorted. This confirms our assumption, that the commonly practiced 1:1 ratio might not be a good choice. There does exist hope for improvement, as the TPM consumes a large percent of the bandwidth with its 28×16 channels, a small sacrifice of its sampling frequency can already release enough bandwidth to largely increase the IMU's sampling frequency and enhance its information quality.

C. Select the Optimal Frequency Ratio and Configure the Embedded System

Now that it is found out that 1:1 setting is not the optimum, the task is then to find out the optimum ratio. Theoretically, given a fixed bandwidth and a frequency ratio, the frequencies of each sensing modality can be directly deduced. In practice, the data acquisition process is influenced by more factors. Below we will first demonstrate how to estimate the real sampling frequencies, then provide the method to select the optimum ratio.

As IMU data quality is heavily influenced by the 1:1 setting, and the matrix consumes a high percent of the bandwidth, we adopt a slice sampling method. At each interrupt, data are read from the two IMUs and only $\lceil n_{row}/k \rceil$ rows of the TPM, where $n = 28$ is the total number of rows. After k repeats, the whole matrix is sampled for 1 time and the IMUs for k times, resulting in a 1 : k frequency ratio. These data

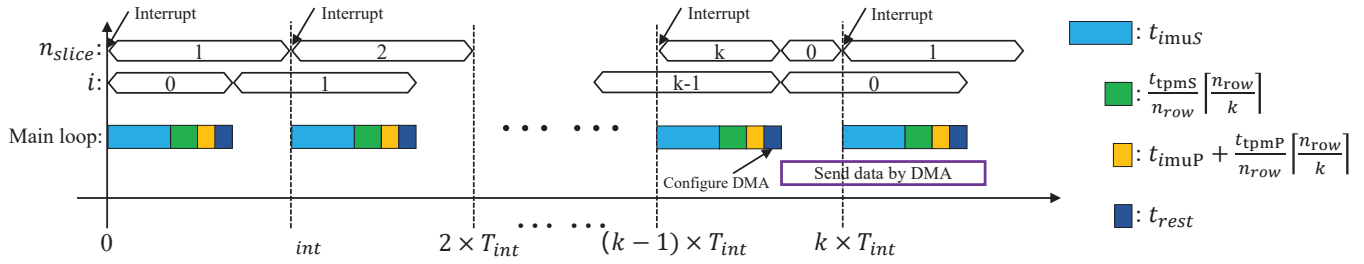


Fig. 6. Slice sampling Gantt chart. Where T_{int} is the set timer interrupt interval, n_{slice} is the current index of the slice to be sampled, and i is the index of the current sampled slice.

are then sent out via DMA through Bluetooth. The general sequence diagram is given in Fig. 6.

The variables are defined as below:

- t_{imuS} : the time needed for sampling the two IMUs;
- t_{imuP} : the time needed for packing the IMUs' data into the final send package;
- t_{tpmS} : the time needed for obtaining data from the entire TPM;
- t_{tpmP} : the time needed for packing the data from the entire TPM;
- n_{row} : the number of TPM rows;
- k : the total number of slices, and also the frequency ratio of the IMUs over the TPM;
- $\lceil n_{row}/k \rceil$: the number of rows sampled each time (at the last round, the rows sampled might be smaller than this value);
- t_{rest} : time consumption of the rest of the codes in a loop (include time spent configuring DMA), which is a constant value.

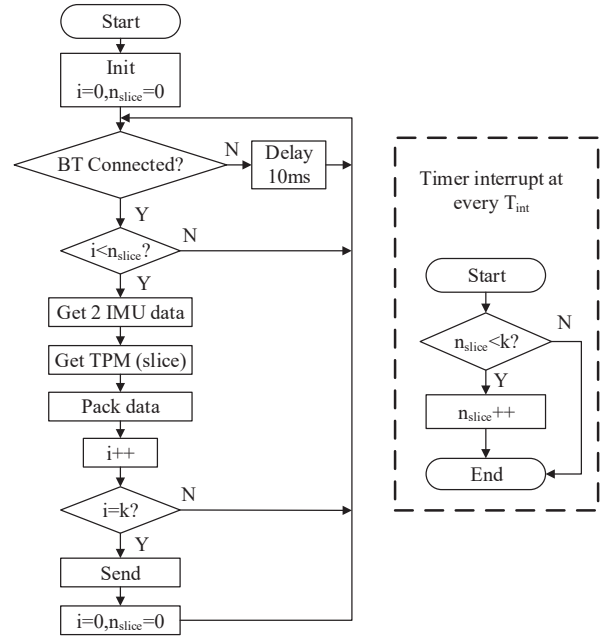


Fig. 7. Slice sampling flow chart

The flow chart for the embedded program is given in Fig. 7.

If the transmission (or storage) bandwidth were not limited, then the time needed for each sample t_{sample} (including two IMUs and part of the matrix) can be estimated as:

$$t_{sample} = t_{imuS} + t_{imuP} + \frac{t_{tpmS} + t_{tpmP}}{n_{row}} \lceil \frac{n_{row}}{k} \rceil + t_{rest} \quad (3)$$

This time is slightly less than the interruption interval (T_{int} in Fig. 7). As the time of interrupt callback is several orders of magnitude less than the existing time, it is not considered here.

The theoretic sampling frequencies of the IMUs and the matrix are then given by:

$$\begin{cases} f_{imu,theory} = 1/t_{sample} \\ f_{tpm,theory} = f_{imu,theory}/k \end{cases} \quad (4)$$

Since the sampling frequency of the TPM is always $1/k$ of the sampling frequency of the IMU, only the equation for the IMU sampling frequency is given below. The sampling

frequency is also limited by the total bandwidth, given as below:

$$f_{imu,limit} = \frac{Bandwidth}{B_{imu} + \lceil n_{row}/k \rceil \times B_{tpm}/n_{row}} \quad (5)$$

The new variables and their values for the smart shoe are given below:

- $Bandwidth$: the bandwidth limit, 48KB/s, due to the usage of Bluetooth transmission;
- B_{tpm} : data size of a whole matrix: 28×16 channels, 12-bit resolution, 672 Bytes/frame;
- B_{imu} : data size of two IMUs per sample: 12 channels, 16-bit resolution, 24 Bytes/sample.

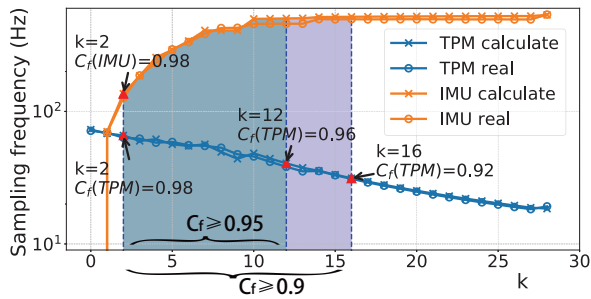


Fig. 8. Sampling frequencies of TPM and IMU at different k's

The real sampling frequency can go beyond neither of the above, thus:

$$\begin{cases} f_{imu} = \min(f_{imu,theory}, f_{imu,limit}) \\ f_{tpm} = f_{imu}/k \end{cases} \quad (6)$$

For the smart shoe, the sample frequencies at different k values can be deduced and measured, the result is provided in Fig. 8. As k grows, the sampling frequency of the IMUs grows and the matrix drops. Already at $k = 2$, $f_{IMU} = 136$, IMUs reach a C_f value of 0.98 (while the C_f of TPM also 0.98). At $k = 12$, $f_{IMU} = 504$, there is almost no information loss for IMUs ($C_f = 0.997$), while the information from the TPM drops to $C_f = 0.95$. It is thus highly recommended to set k to 2 and the embedded program of the smart shoe shall be configured accordingly before gait analyze experiments.

The demonstration of the whole workflow of information quality evaluation, optimal frequency ratio selection and system configuration is finally finished.

V. CONCLUSION AND DISCUSSIONS

The information loss in the digitalization process can not be recovered later by the software. Frequency selection for sensing systems is thus very important. For multi-modality IoT systems, we propose a general workflow, 1) to answer the question whether frequency balancing is needed, 2) to find out the proper frequency ratio and 3) to configure this ratio into the embedded system. We demonstrate how this workflow functions with a concrete example, a smart shoe equipped with two IMUs and a pressure sensing matrix. The sensor configuration of this example system (multiple sensors together with a sensing matrix) and the corresponding workflow are general enough to be expanded also to other multi-modality IoT systems.

We find out that the commonly practiced 1:1 frequency ratio might not be a good choice for multi-modality IoT systems. Specially, IMU frequency might need to be raised to over 100Hz if the goal goes beyond activity recognition. It is highly recommended to follow the frequency balancing workflow before carrying out large-scale data acquisition, so that the most information can be kept, ensuring a sound dataset for information retrieval algorithms and also for future use.

REFERENCES

- [1] Y. Kang, D. Y. Chen, M. Lawo, and S. J. Xia Hou, "A wearable swallowing detecting method based on nanometer materials sensor," in *Advances in Science and Technology*, vol. 100. Trans Tech Publ, 2017, pp. 120–129.
- [2] F. Seoane, J. Ferreira, J. J. Sanchez, and R. Bragós, "An analog front-end enables electrical impedance spectroscopy system on-chip for biomedical applications," *Physiological measurement*, vol. 29, no. 6, p. S267, 2008.
- [3] J. Cheng, O. Amft, G. Bahle, and P. Lukowicz, "Designing sensitive wearable capacitive sensors for activity recognition," *IEEE sensors journal*, vol. 13, no. 10, pp. 3935–3947, 2013.
- [4] M. Rofouei, M. A. Ghodrat, Y. Huang, N. Alshurafa, and M. Sarrafzadeh, "Improving accuracy in e-textiles as a platform for pervasive sensing," in *2013 IEEE International Conference on Body Sensor Networks*. IEEE, 2013, pp. 1–6.
- [5] T. Thamaraimanalan and P. Sampath, "A low power fuzzy logic based variable resolution adc for wireless ecg monitoring systems," *Cognitive Systems Research*, vol. 57, pp. 236–245, 2019.
- [6] R. Paradiso, G. Loriga, and N. Taccini, "A wearable health care system based on knitted integrated sensors," *IEEE transactions on Information Technology in biomedicine*, vol. 9, no. 3, pp. 337–344, 2005.
- [7] Y.-D. Lee and W.-Y. Chung, "Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring," *Sensors and Actuators B: Chemical*, vol. 140, no. 2, pp. 390–395, 2009.
- [8] P. Vepakomma, D. De, S. K. Das, and S. Bhansali, "A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities," in *2015 IEEE 12th International conference on wearable and implantable body sensor networks (BSN)*. IEEE, 2015, pp. 1–6.
- [9] C. N. Teague, J. A. Heller, B. N. Nevius, A. M. Carek, S. Mabrouk, F. Garcia-Vicente, O. T. Inan, and M. Etemadi, "A wearable, multimodal sensing system to monitor knee joint health," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 323–10 334, 2020.
- [10] Y. Cai, X. Qian, H. Cao, J. Zheng, W. Xu, and M.-C. Huang, "mhealth technologies toward active health information collection and tracking in daily life: A dynamic gait monitoring example," *IEEE Internet of Things Journal*, 2022.
- [11] S. Taleb, H. Hajj, and Z. Dawy, "Entropy-based optimization to trade-off energy and accuracy for activity mobile sensing," in *2013 4th Annual International Conference on Energy Aware Computing Systems and Applications (ICEAC)*. IEEE, 2013, pp. 6–11.
- [12] A. Allik, K. Pilt, D. Karai, I. Fridolin, M. Leier, and G. Jervan, "Optimization of physical activity recognition for real-time wearable systems: effect of window length, sampling frequency and number of features," *Applied Sciences*, vol. 9, no. 22, p. 4833, 2019.
- [13] A. Khan, N. Hammerla, S. Mellor, and T. Plötz, "Optimising sampling rates for accelerometer-based human activity recognition," *Pattern Recognition Letters*, vol. 73, pp. 33–40, 2016.
- [14] N. Anish, G. Bhat, J. Park, H. G. Lee, and U. Y. Ogras, "Sensor-classifier co-optimization for wearable human activity recognition applications," in *2019 IEEE International Conference on Embedded Software and Systems (ICCESS)*. IEEE, 2019, pp. 1–4.
- [15] Z. G. Xiao and C. Menon, "An investigation on the sampling frequency of the upper-limb force myographic signals," *Sensors*, vol. 19, no. 11, p. 2432, 2019.
- [16] G. Lei, S. Zhang, Y. Fang, Y. Wang, and X. Zhang, "Investigation on the sampling frequency and channel number for force myography based hand gesture recognition," *Sensors*, vol. 21, no. 11, p. 3872, 2021.
- [17] T. W. Anderson, "On the distribution of the two-sample cramer-von mises criterion," *The Annals of Mathematical Statistics*, pp. 1148–1159, 1962.
- [18] S. Community, "Two-sample cramer-von-mises criterion," https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.cramervonmises_2samp.html, 2022.

Enhanced Photon Detection Probability Model for Single-Photon Avalanche Diodes in TCAD with Machine Learning

Xuanyu Qian
Electrical and Computer Engineering
McMaster University
Hamilton, ON, Canada
qianx27@mcmaster.ca

Wei Jiang
Electrical and Computer Engineering
McMaster University
Hamilton, ON, Canada
jiangw35@mcmaster.ca

M. Jamal Deen
Electrical and Computer Engineering
& School of Biomedical Engineering
McMaster University
Hamilton, ON, Canada
jamal@mcmaster.ca

Abstract—Accurate photon detection probability (PDP) modeling is important for the optimized design of single-photon avalanche diodes (SPADs) using modern standard CMOS technologies. To ensure a planar active region of a SPAD, the edge of the depletion region must have a lower electric field, so a lower doping concentration is needed. However, this edge effect may have a negative impact on the total PDP, especially for small-sized SPADs. In this paper, we proposed an enhanced PDP modeling process by combining the Technology Computer-Aided Design (TCAD) simulations with machine learning (ML) techniques. Using this ML-TCAD PDP model, we investigated the influence of the edge effect on the PDP of SPADs by varying the diameter of the SPADs from 1.75 μm to 8.75 μm . After generating the sample simulation data, Gaussian process regression (GPR) and deep neuron network (DNN) are applied to train the model. With the application of principal component analysis (PCA), the accuracy of the trained models was significantly improved. Overall, this ML-TCAD PDP model provides an optimized and accelerated design process for SPADs, thus saving simulation time and reducing the design iterations required in the traditional design process of SPADs.

Keywords—Photon detection probability (PDP), single-photon avalanche diodes (SPADs), edge effect, machine learning (ML), TCAD, principal component analysis (PCA)

I. INTRODUCTION

Single-photon avalanche diodes (SPADs) are p-n junctions that are reversely biased above their breakdown voltages (V_{BD}), thus leading to extremely high electric fields across the depletion regions. In such high electric fields, avalanching can be triggered even by a single photon. Due to the high sensitivity, SPADs are widely used in many optical applications including biomedical imaging [1]–[4], communications [5], and light detection and ranging [6]. With the development of silicon-based fabrication technologies, many SPADs were designed and fabricated using standard complementary metal-oxide-semiconductor (CMOS) technology due to their lower manufacturing cost and the potential to be integrated with other CMOS read-out and signal processing circuits to implement complete imaging systems to achieve a strong overall performance [7]–[11]. However, CMOS SPADs usually suffer from the low photon detection probability (PDP) due to two main reasons [12], [13]. First, the thin

This work was supported in part by the Natural Science and Engineering Research Council of Canada, the Canada Research Chair Program, and CMC Microsystems. 978-1-6654-8684-2/22/\$31.00 2022 IEEE

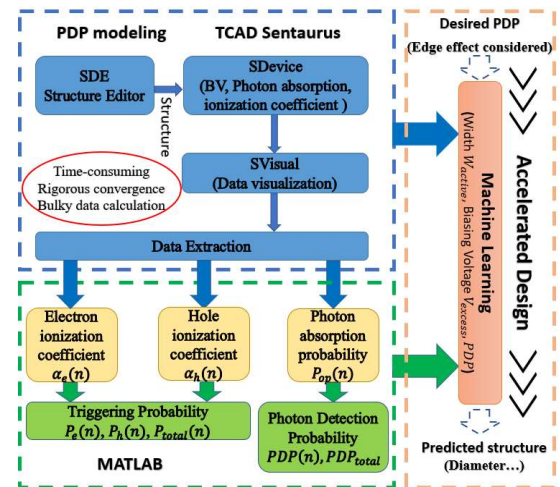


Fig. 1. Design flow of enhanced 2-D PDP simulation with ML-TCAD.

depletion regions of CMOS SPADs lead to a low photon absorption probability. Second, the reflection and absorption of inter-metal dielectric layers and passivation layer result in only a portion of incident photons arriving the active region of SPADs. Traditionally, to optimize the PDP performance, the initial prototype of SPADs needs to be designed, fabricated, and measured. Then, the SPAD designs will be revised based on the initial measured results, then the entire process is repeated, so the final design may need several iterations. This traditional optimization process is costly and time-consuming.

In this paper, to accelerate this optimization process, we developed a PDP model using TCAD simulation tools. With this model, the PDP performance can be easily simulated when varying the SPAD design parameters such as layer doping concentrations, guard ring structure, junction depth, and active region diameter, thus significantly reducing the time and cost for the SPAD PDP optimization. In addition, we investigated the edge effect for SPADs with different diameters and found that the edge effect has more impact on the total PDP in small-sized SPADs. We applied machine learning (ML) techniques, including the Gaussian process regression (GPR) and the deep neural network (DNN), to further improve the PDP modeling. The accuracy of diameter predictions is significantly improved after we introduced the principal component analysis (PCA) into the models. The proposed

ML-TCAD PDP modeling process is shown in Fig 1. The whole process is divided into two parts: PDP modeling and ML. In the PDP modeling part, SPAD’s structure is created and then simulated using TCAD. The data such as photon absorption probability, electron and hole ionization coefficients will be extracted from TCAD directly and processed in MATLAB to calculate the PDP. For the ML part, the diameters and the final PDPs with edge effect of SPADs with different sizes will be used to train ML models to further improve the efficiency and accuracy of the SPAD PDP model.

II. PDP MODELING OF SPAD

The PDP modeling process of the SPADs mainly includes 4 steps: simulating breakdown voltage, simulating triggering probability, simulating photon absorption probability, and calculating the final integration. The simulations were performed using TCAD Synopsys Sentaurus [14], [15] because this tool provides various models for avalanche simulation.

A. Breakdown Voltage

Considering that SPADs are reverse-biased above the breakdown voltage with typical critical electric fields beyond 10^5 V/cm, the avalanche model we applied is the Okuto-Crowell model [15], [16]. The model is implemented as a “local model” so that the ionization coefficients will only be determined by the local electric field. The structure of the SPAD was built using Sentaurus Structure Editor. Fig. 2 shows the cross-sectional view of the general structure of a p⁺/n-Well SPAD with p-Well guard rings. The doping concentrations of the n⁺ and p⁺ regions are set to be 10^{19} cm⁻³ while doping concentrations of the p-Well guard ring and the n-Well are set to be 10^{17} cm⁻³. Shallow trench isolations (STIs) are applied to isolate the n⁺ and p⁺ regions according to the process-related requirements. The Ohmic contacts are tied directly to the highly doped n⁺ or p⁺ regions. With n-type contact tied to the ground, negative sweeping voltages are applied on the p-type contacts to find the avalanche breakdown point. When diodes are reverse-biased above V_{BD} , the reverse current will increase sharply, usually several orders of magnitude [17], [18]. The simulated current-voltage (*I-V*) characteristic is shown in Fig. 3. Upon the breakdown, the reverse current increases almost vertically, from a typical value below 10^{-10} A to around 10^{-6} A. Additional break criteria were added to stop the avalanching process, thus avoiding the convergence problem. In addition to the *I-V* characteristic, we also performed avalanche analysis to calculate the ionization integrals, provided by Sentaurus

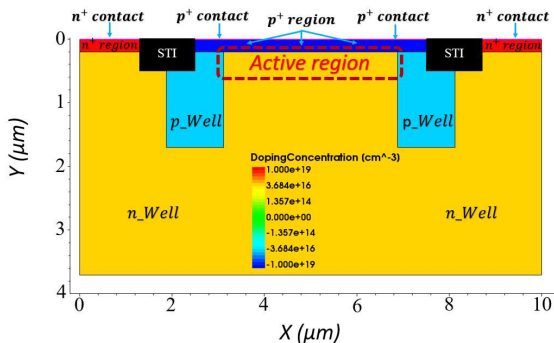


Fig. 2. Schematic of the SPAD cross-sectional view.

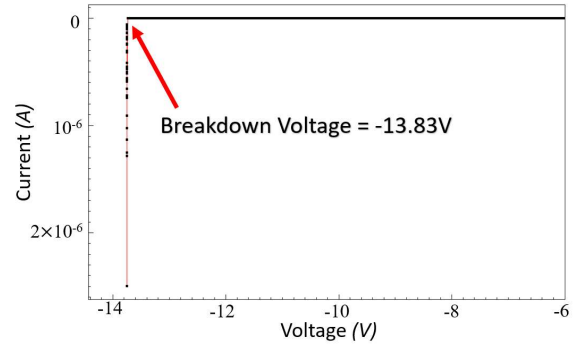


Fig. 3. Current-voltage (*I-V*) characteristics of the SPAD in Fig. 2. Breakdown voltage $V_{BD}=-13.8$ V obtained.

SDevice. When the ionization integral calculated by TCAD using Eq. 1 and Eq. 2 reaches 1, the V_{BD} is determined.

$$I_E = \int_0^d \alpha_e(x) e^{-\int_x^d (\alpha_e(x') - \alpha_h(x')) dx'} dx \quad (1)$$

$$I_h = \int_0^d \alpha_h(x) e^{-\int_x^d (\alpha_h(x') - \alpha_e(x')) dx'} dx \quad (2)$$

The terms d , α_e and α_h in Eq. 1 and Eq.2 represent the integration path length, electron ionization coefficient and hole ionization coefficient, respectively. Combining the two methods, the accurate V_{BD} of the SPAD was determined to be 13.83 V as shown in Fig. 3.

B. Avalanche Triggering Probability

Avalanche triggering probability is defined to be the probability that carriers induce an avalanching process at specific positions. The widely used McIntyre model is applied in our work as it is easy to be extended in 2-D calculations [19]. The differential equations for calculating the triggering probability are shown in Eq. 3 and Eq. 4.

$$dP_e/dx = (1 - P_e) \times \alpha_e \times (P_e + P_h - P_e P_h) \quad (3)$$

$$dP_h/dx = -(1 - P_h) \times \alpha_e \times (P_e + P_h - P_e P_h) \quad (4)$$

The terms $P_e(x)$ and $P_h(x)$ represent the triggering probabilities of electrons and holes, respectively. The terms α_e and α_h are electron and hole ionization coefficients, respectively. In the previous work, the ionization coefficients are generally either referenced from the literature [20] or calculated through an external MATLAB routine [21]. However, in our work we directly used the simulated results of the ionization coefficients using TCAD. To simulate the 2-D PDP with the edge effect, some researchers chose to use a two-step discrete way [21]. They calculated the triggering probability through the lines following the electric field’s direction first, and then did the integration along the line. However, this method is not convenient and is time-consuming. Therefore, we implemented the simulations of the electric field, ionization coefficients and optical absorption simultaneously in our work, which facilitates the simulation process. Electric field distributions of SPADs with 3 different diameters are shown in Fig 4. SPADs with different diameters are all biased at 1V beyond the V_{BD} (-

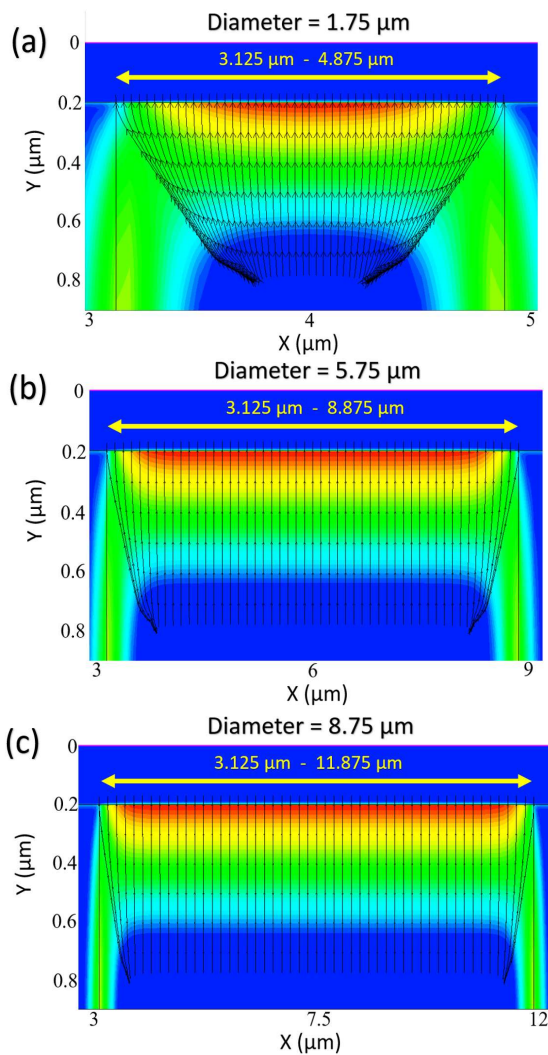


Fig. 4. Electric field distribution of SPAD with diameter of (a) 1.75 μm. (b) 5.75 μm. (c) 8.75 μm, 51 lines along electric force lines are created.

14.83 V). For the SPADs with different diameters, 51 lines following the electric field's direction were generated, and the lengths of each line were specified by applying the boundary conditions of the depletion region. Since the 51 lines were generated through the simulation directly, and no discretization or approximation was applied, our method provided more accurate ionization coefficients and distances for the integration when calculating the avalanche triggering probability. The typical simulated ionization coefficients of electrons and holes of a SPAD with a 5.75 μm diameter are shown in Fig. 5. In Fig. 5, the highest ionization coefficients are in the central part and a decreasing trend of ionization coefficients is shown when the lines are approaching to the edges. After simulation, the coefficients were extracted from TCAD and then imported to MATLAB. These coefficients were used as accurate parameters for solving the differential equations in Eq. 3 and Eq. 4 with the additional boundary conditions in Eq. 5 and Eq. 6.

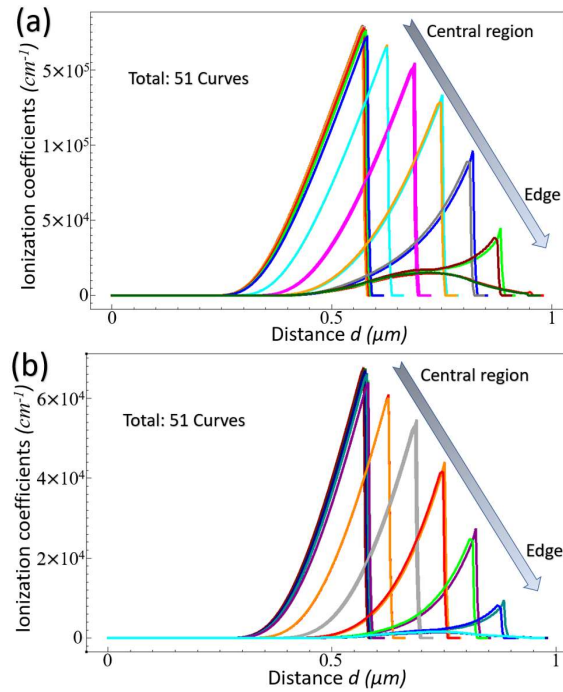


Fig. 5. Ionization coefficients along 51 lines of (a) electrons and (b) holes, based on a SPAD of a 5.75 μm diameter.

$$P_e(0) = 0 \quad (5)$$

$$P_h(d) = 0 \quad (6)$$

The term d is the length of each line. Since the length of each line is determined by specifying the width of the depletion region, additional definitions of the depletion region's length in MATLAB are not needed when solving the differential equations. The triggering probability of each line was calculated using the BVP4C solver in MATLAB. The calculated results show that the lines in the central region have higher triggering probabilities while the lines close to edges show negligible triggering probabilities. For example, the central line (line 26) has a triggering probability of 0.55 while the line closest to edges (line 1 and line 51) has a triggering probability of almost 0. These results clearly show the influence of the edge effect on avalanche triggering probability.

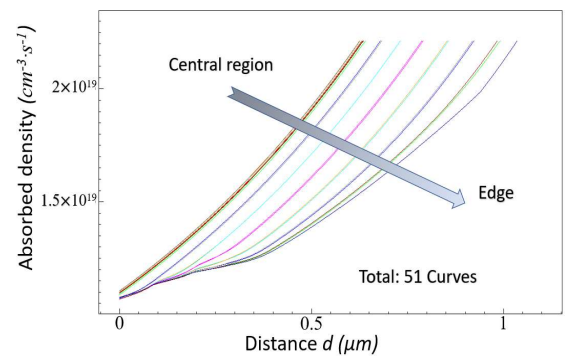


Fig. 6. Absorbed photon density along 51 lines based on a SPAD of 5.75 μm diameter, illuminated by a 500-nm light source

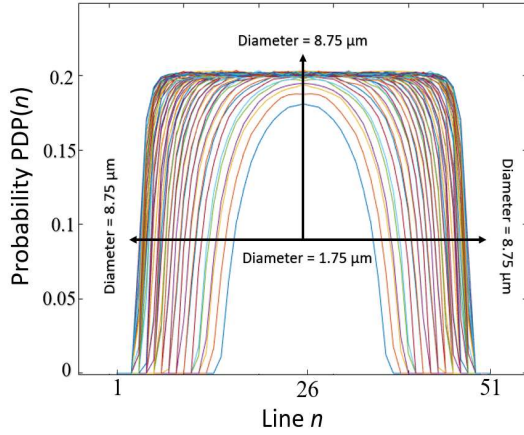


Fig. 7. Simulated photon detection probability of SPADs with diameters ranging from 1.75 μm to 8.75 μm .

C. Photon Absorption Probability

The photon absorption probability (PDP) determines the available number of carriers in the depletion region to possibly trigger avalanche events. Compared to the triggering probability, the photon absorption probability of a fixed material mainly only depends on the depth, following an exponential decay. In this work, the material is silicon. The exact values of the absorbed photon density along the 51 lines we created were first simulated using a vertical illumination of a 500-nm light source, as shown in Fig. 6. Then, the photon absorption probabilities were calculated using Eq. 7, with the intensity $I_{op} = 1 \text{ mW/cm}^2$, Planck constant $h = 6.626 \times 10^{-34} \text{ J}\cdot\text{s}$, speed of light in vacuum $c = 2.998 \times 10^8 \text{ m/s}$, and the virtual thickness $t_v = 1 \mu\text{m}$.

$$P_{op} = \text{Absorbed density} / \frac{I_{op} \times \lambda \times t_v}{hc} \quad (7)$$

D. Photon Detection Probability

The $PDP(n)$ ($n = 1, 2, \dots, 51$) in the depletion region was then calculated according to Eq. 8 with the actual length of each line, where P_{trig} is the total triggering probability defined in Eq. 9.

$$PDP(n) = \int_0^{d_n} P_{op}(x) \times P_{trig}(x) dx \quad (8)$$

$$P_{trig}(x) = P_e(x) + P_h(x) - P_e(x)P_h(x) \quad (9)$$

Next, we varied the diameters of the SPADs to investigate the edge effect on different-sized SPADs. Fig. 7 shows the $PDP(1) - PDP(51)$ of each SPAD with the diameters ranging from 1.75 to 8.75 μm . As expected, a significant edge effect is shown on the PDP for SPADs with smaller diameters. Similar results can be found in [21]. However, the actual PDP modeling or measurement process is time-consuming since it needs careful settings for the simulation conditions, as mentioned in Fig. 1. The simulation of the accurate breakdown voltage, extraction of ionization coefficients, manual operation to generate lines for integration, and the final calculation of $PDP(n)$ require a large amount of time and expertise in optoelectronic device design and programming. One

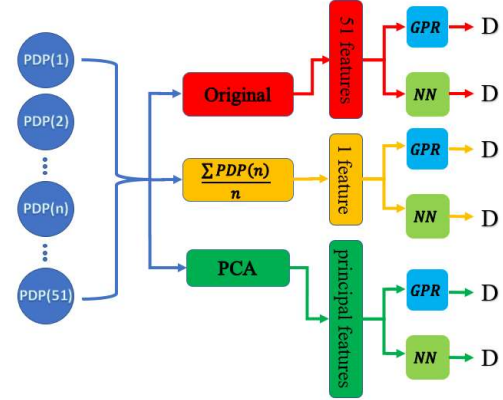


Fig. 8. Proposed process for predicting the diameter D by using the calculated 2-D PDP values.

promising method to improve the efficiency of the optimization process for SPAD design is to combine ML techniques with PDP modeling, which will be discussed in detail in Section III.

III. PREDICTION OF SPADs' DIAMETER

ML was recently applied in the field of TCAD device modeling due to its high efficiency, accuracy, and low cost. Typical applications of the ML-TCAD combined approaches include the prediction of the FinFET defect location [22], process variation analysis of vertical FET [23], breakdown estimation [24], and even circuit design [25].

The effectiveness of the trained model mainly relies on the number of samples, data processing, and algorithms [24]–[27]. However, many works are only based on the basic $I-V$ characteristics of the devices. In our work, we used the complicated input attributes of the PDP of different SPADs to train our model. Our proposed ML route for the diameter predictions is shown in Fig 8. The simulated $PDP(n)$ are preprocessed in three different ways: original values, average value, and PCA applied. After that, processed data will be used to train two different models: Gaussian process regression (GPR) and deep neural network (DNN). The details of each part are stated as follows.

First, GPR and DNN are used to train the model with the inputs $PDP(n)$ and the output diameter D . GPR is believed to be suitable for solving problems with uncertain complexity due to its ability to find the correlation between the samples and the data. The DNN layers are activated using the common rectified linear unit (ReLU) activation function [28]. Fig. 9 shows that GPR trained model gives a good prediction with the root-mean-square error (RMSE) of 0.137 μm while the DNN trained model has presented some unreasonable predictions. This is probably caused by overfitting when using DNN to train the model. Some small variations of the $PDP(n)$ may lead to large differences in the outputs of the model. To alleviate the possible overfitting, we summed up the $PDP(n)$ of each sample and calculated the mean values to train the two models. As expected, the overfitting phenomenon was improved, with RMSE of 0.113 μm and 0.107 μm for the GPR model and the DNN model, respectively.

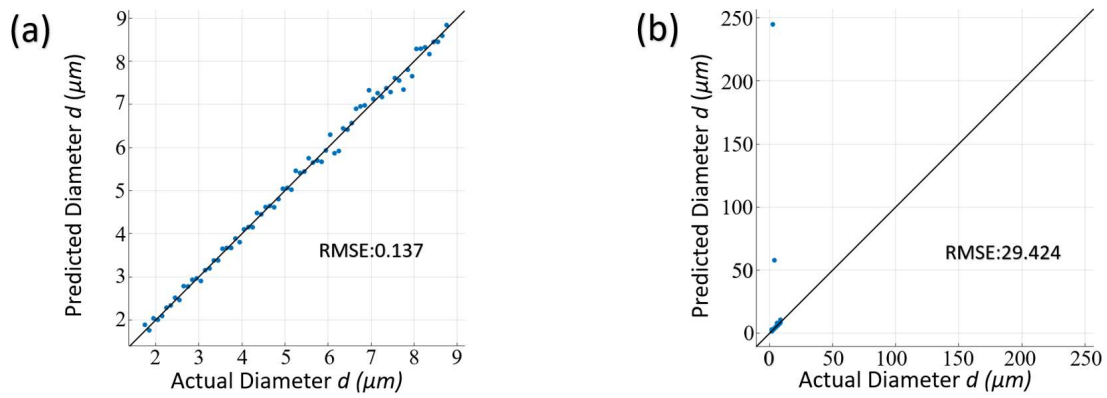


Fig. 9. Predictions versus actual diameters with original PDP used. (a) gaussian process regression. (b) neural network.

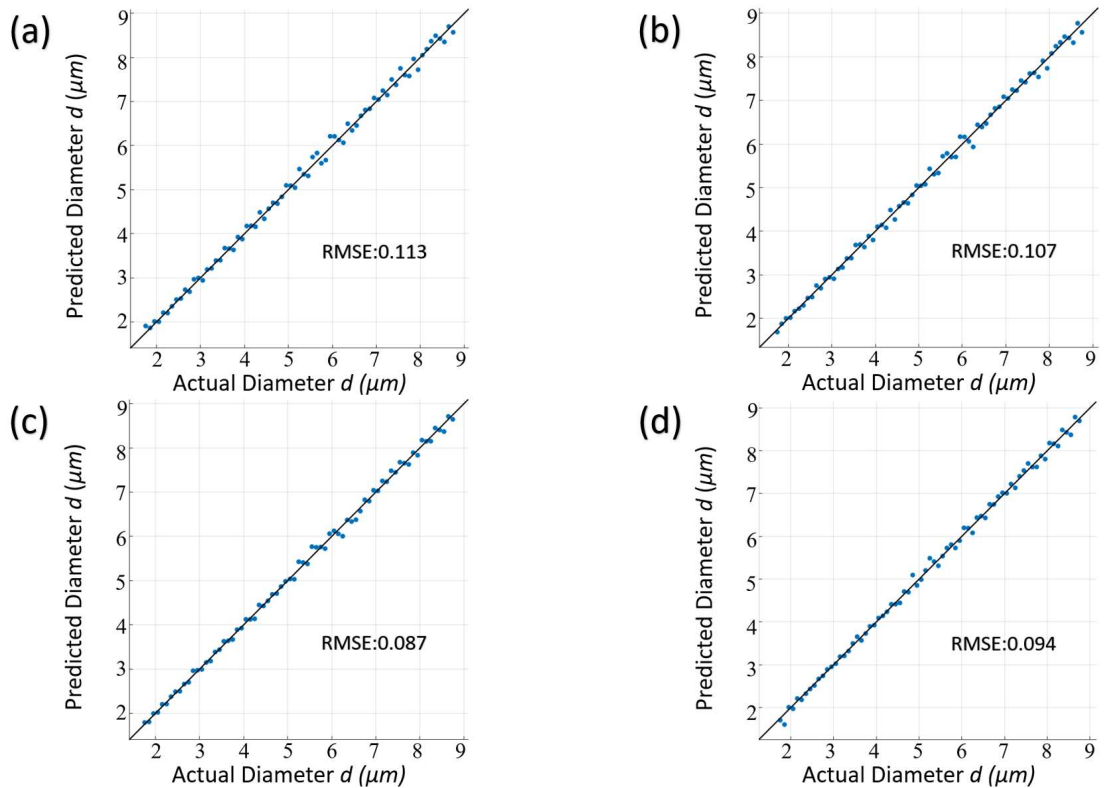


Fig. 10. Predictions versus actual diameters with (a) Gaussian process regression using PDP mean value. (b) Neural network using PDP mean value. (c) Gaussian process regression using PCA. (d) Neural network using PCA

We further improved the effectiveness of the ML-TCAD PDP model by applying the principal component analysis (PCA) to the original PDP(n) along each line. Originally, the input features of each sample are 51. The PCA can help to reduce the dimensions for the training of the model by identifying the most important components. Fig. 10 shows the predictions made by using mean value and PCA, respectively. The models that are combined with the PCA provide the optimal predictions, with RMSE of 0.087 μm for the GPR and 0.094 μm for the DNN model, respectively. The comparison of different models is shown in Table I.

TABLE I. COMPARISON OF MODELS' EFFECTIVENESS

Data Inputs	Features used	Model type	Root-mean-square error (μm)
Original PDP(n) ($n = 1, 2, \dots, 51$)	51	GPR	0.137
		NN	29.424
Mean value of PDP(n)	1	GPR	0.113
		NN	0.107
PCA processed PDP(n)	3	GPR	0.087
		NN	0.094

IV. CONCLUSION AND FUTURE WORK

In conclusion, we proposed a new design process for SPADs that combines the 2-D PDP model that takes edge effects into consideration with the machine learning (ML)

algorithms. We optimized the simulation of 2-D PDP in the depletion region based on simulated results without needing any discrete approximations. Furthermore, the introduction of the principal component analysis (PCA) into the ML model training effectively improves prediction accuracy. To our best knowledge, this work is the first one to improve the accuracy of PDP modeling for SPADs by using ML techniques. The results show that this ML-TCAD PDP model will save the time, cost and effort for the optimization process of the SPAD design. Our future work will include more variables and other physical processes like the biasing voltage, doping concentration and dark count rate to develop a comprehensive SPAD model that can predict all parameters for the SPAD.

REFERENCES

[1] Z. Li, M. J. Deen, Q. Fang, and P. R. Selvaganapathy, "Towards a Portable Raman Spectrometer using a Concave Grating and a Time-Gated CMOS SPAD," *Optics Express*, Vol. 22, Issue 15, pp. 18736-18747, vol. 22, no. 15, pp. 18736-18747, Jul. 2014, doi: 10.1364/OE.22.018736.

[2] W. Jiang, Y. Chalich, and M. J. Deen, "Sensors for Positron Emission Tomography Applications," *Sensors 2019*, Vol. 19, Page 5019, vol. 19, no. 22, p. 5019, Nov. 2019, doi: 10.3390/S19225019.

[3] D. P. Palubiak and M. J. Deen, "CMOS SPADs: Design Issues and Research Challenges for Detectors, Circuits, and Arrays," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 409-426, Nov. 2014, doi: 10.1109/JSTQE.2014.2344034.

[4] M. Alayed and M. J. Deen, "Time-Resolved Diffuse Optical Spectroscopy and Imaging Using Solid-State Detectors: Characteristics, Present Status, and Research Challenges," *Sensors 2017*, Vol. 17, Page 2115, vol. 17, no. 9, p. 2115, Sep. 2017, doi: 10.3390/S17092115.

[5] J. Kosman *et al.*, "29.7 A 500Mb/s -46.1dBm CMOS SPAD Receiver for Laser Diode Visible-Light Communications," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2019-February, pp. 468-470, Mar. 2019, doi: 10.1109/ISSCC.2019.8662427.

[6] O. Kumagai *et al.*, "A 189x600 Back-Illuminated Stacked SPAD Direct Time-of-Flight Depth Sensor for Automotive LiDAR Systems," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 64, pp. 110-112, Feb. 2021, doi: 10.1109/ISSCC42613.2021.9365961.

[7] W. Jiang, Y. Chalich, R. Scott, and M. J. Deen, "Time-Gated and Multi-Junction SPADs in Standard 65 nm CMOS Technology," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 12092-12103, May 2021, doi: 10.1109/JSEN.2021.3063319.

[8] E. Charbon, H. J. Yoon, and Y. Maruyama, "A Geiger Mode APD Fabricated in Standard 65nm CMOS Technology," *Technical Digest - International Electron Devices Meeting, IEDM*, 2013, doi: 10.1109/IEDM.2013.6724705.

[9] J. Rhim *et al.*, "Monolithically-Integrated Single-Photon Avalanche Diode in a Zero-Change Standard CMOS Process for Low-Cost and Low-Voltage LiDAR Application," *Instruments 2019*, Vol. 3, Page 33, vol. 3, no. 2, p. 33, Jun. 2019, doi: 10.3390/INSTRUMENTS3020033.

[10] W. Jiang, R. Scott, and M. J. Deen, "Differential Quench and Reset Circuit for Single-Photon Avalanche Diodes," *Journal of Lightwave Technology*, vol. 39, no. 22, pp. 7334-7342, Nov. 2021, doi: 10.1109/JLT.2021.3111119.

[11] W. Jiang, R. Scott, and M. J. Deen, "High-Speed Active Quench and Reset Circuit for SPAD in a Standard 65 nm CMOS Technology," *IEEE Photonics Technology Letters*, vol. 33, no. 24, pp. 1431-1434, Dec. 2021, doi: 10.1109/LPT.2021.3124989.

[12] W. Jiang and M. J. Deen, "Random Telegraph Signal in n+/p-Well CMOS Single-Photon Avalanche Diodes," *IEEE Transactions on Electron Devices*, vol. 68, no. 6, pp. 2764-2769, Jun. 2021, doi: 10.1109/TED.2021.3070557.

[13] W. Jiang, R. Scott, and M. Jamal Deen, "Improved Noise Performance of CMOS Poly Gate Single-Photon Avalanche Diodes," *IEEE Photonics Journal*, vol. 14, no. 1, Feb. 2022, doi: 10.1109/JPHOT.2021.3128055.

[14] "Sentaurus™ Structure Editor User Guide Version R-2020.09, September 2020."

[15] "Sentaurus™ Device User Guide Version R-2020.09, September 2020."

[16] Y. Okuto and C. R. Crowell, "Threshold Energy Effect on Avalanche Breakdown Voltage in Semiconductor Junctions," *Solid-State Electronics*, vol. 18, no. 2, pp. 161-168, Feb. 1975, doi: 10.1016/0038-1101(75)90099-4.

[17] C. L. Forrest Ma, M. J. Deen, L. E. Tarof, and J. C. H. Yu, "Temperature Dependence of Breakdown Voltages in Separate Absorption, Grading, Charge, and Multiplication InP/InGaAs Avalanche Photodiodes," *IEEE Transactions on Electron Devices*, vol. 42, no. 5, pp. 810-818, 1995, doi: 10.1109/16.381974.

[18] A. Bandyopadhyay, M. Jamal Deen, L. E. Tarof, and W. Clark, "A Simplified Approach to Time-Domain Modeling of Avalanche Photodiodes," *IEEE Journal of Quantum Electronics*, vol. 34, no. 4, pp. 691-699, Apr. 1998, doi: 10.1109/3.663452.

[19] R. J. McIntyre, "On the Avalanche Initiation Probability of Avalanche Diodes Above the Breakdown Voltage," *IEEE Transactions on Electron Devices*, vol. 20, no. 7, pp. 637-641, 1973, doi: 10.1109/T-ED.1973.17715.

[20] A. Panglosse, P. Martin-Gonthier, O. Marcelot, C. Virmontois, O. Saint-Pé, and P. Magnan, "Modeling, Simulation Methods and Characterization of Photon Detection Probability in CMOS-SPAD," *Sensors*, vol. 21, no. 17, Sep. 2021, doi: 10.3390/s21175860.

[21] C. H. Liu, C. A. Hsien, and S. di Lin, "2-D Photon-Detection-Probability Simulation and a Novel Guard-Ring Design for Small CMOS Single-Photon Avalanche Diodes," *IEEE Transactions on Electron Devices*, 2021, doi: 10.1109/TED.2021.3119264.

[22] C. W. Teo, K. L. Low, V. Narang, and A. V. Y. Thean, "TCAD-Enabled Machine Learning Defect Prediction to Accelerate Advanced Semiconductor Device Failure Analysis," *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD*, vol. 2019-September, Sep. 2019, doi: 10.1109/SISPAD.2019.8870440.

[23] K. Ko, J. K. Lee, M. Kang, J. Jeon, and H. Shin, "Prediction of Process Variation Effect for Ultrascaled GAA Vertical FET Devices Using a Machine Learning Approach," *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4474-4477, Oct. 2019, doi: 10.1109/TED.2019.2937786.

[24] J. Chen *et al.*, "Powernet: SOI Lateral Power Device Breakdown Prediction with Deep Neural Networks," *IEEE Access*, vol. 8, pp. 25372-25382, 2020, doi: 10.1109/ACCESS.2020.2970966.

[25] N. Kandpal, A. Singh, and A. Agarwal, "A Machine Learning Driven PVT-Robust VCO with Enhanced Linearity Range," *Circuits, Systems, and Signal Processing*, pp. 1-18, Mar. 2022, doi: 10.1007/S00034-022-02001-X/TABLES/3.

[26] H. Y. Wong *et al.*, "TCAD-Machine Learning Framework for Device Variation and Operating Temperature Analysis With Experimental Demonstration," *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 992-1000, 2020, doi: 10.1109/JEDS.2020.3024669.

[27] K. Mehta, S. S. Raju, M. Xiao, B. Wang, Y. Zhang, and H. Y. Wong, "Improvement of TCAD Augmented Machine Learning Using Autoencoder for Semiconductor Variation Identification and Inverse Design," *IEEE Access*, vol. 8, pp. 143519-143529, 2020, doi: 10.1109/ACCESS.2020.3014470.

[28] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," Mar. 2018, doi: 10.48550/arxiv.1803.08375.

Design of an automated manure collection system for the production of biogas through biodigesters

Iraiz Lucero Quintanilla-Mosquera
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
70347500@continental.edu.pe

Jhon Rodrigo Ortiz-Zacarias
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
71689110@continental.edu.pe

Sliver Ivan Del Carpio-Ramirez
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
71902502@continental.edu.pe

Jesús Eduardo Rosales-Fierro
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
76560426@continental.edu.pe

Yadhira S. Valenzuela-Lino
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
73105932@continental.edu.pe

Carlos Coaquira- Rojo
Department of Mechatronic
Engineering
Universidad Continental
Huancayo, Perú
ccoaquira@continental.edu.pe

Nabilt Moggiano
Engineering Research Unit,
Faculty of Engineering
Universidad Continental
Huancayo, Perú
nmoggiano@continental.edu.pe

Abstract— *This research presents the design of an automated system for obtaining biogas, mainly methane gas, through biodigesters. The anaerobic microorganisms present in the manure generate methane through their digestion, which is the main object of this study, since without this process it would not be possible to obtain natural gas. On the other hand, the system has 5 controlled processes that will complete the anaerobic digestion, namely: evaluation and positioning of manure, transmission belts and actuators that together separate the organic matter and begin with the filling of the biodigesters for the extraction of natural gas; additionally, there will be an adequate temperature control for the development of microorganisms, since at a higher temperature in its ideal environment, the gas will be obtained in less time, so the system will count and turn on heaters or fans based on the results. The objective of this process is to avoid exposure to the contagion of diseases in people and reduce the emission of greenhouse gases.*

Keywords— *Automated system, methane, collection, biodigesters, storage.*

I. INTRODUCTION

Livestock is one of the main polluting activities, since manure is one of the environmental problems that develop in livestock complexes [1], nitrous oxide (N₂O) and methane (CH₄) are important gases involved in the effect greenhouse produced by these animals during the process of manure management and enteric fermentation, the largest amounts of emissions produced are caused by inefficient systems in the rearing, management and production of manure in developing countries due to limited technical and economic resources, for the which the use of mitigation tools is recommended to reduce these gases in the atmosphere [2].

More than 1 million tons of manure is produced in the livestock sector from the digestion of feed in the rumen per

year which has a GHG potential of 21 and 310 times greater than CO₂ by producing methane and N₂O respectively, where burning 1 ton of methane is equivalent to eliminating 24 tons of CO₂ [3]; furthermore, enteric methane accounts for 17% of global CH₄ emissions, which is produced mostly by ruminant animals, and their excretion contributes to 0.4% and 2% of global GHG and methane emissions in that order [4]. Indonesia is committed to reduce 29% of domestic greenhouse gas emissions with its own effort and 41% of international support by 2030, because the anaerobic digester captures CH₄ from manure for use in bioenergy to replace non-renewable energy such as kerosene and LPG, This converts manure into methane-rich biogas, indirectly reducing GHG emissions produced by animals, since the livestock sector contributes between 18% and 51% and the agricultural sector between 10% and 12% of total GHG emissions, digesters produce a concentration of 60% to 80% of methane from pig, poultry and cattle slurry [3].

One of the main facts when considering strategies for methane reduction is methanogens, given that methanogens are the only organisms capable of producing methane, other strategies that are investigated are vaccines and defaunation; other studies take advantage of the potential produced by bioenergy from cow manure in Comarca Lagunera of 21 GW/year and in Mexico of 410 GW/year with a reduction of greenhouse gases of 21 Gg per year in CO₂ equivalent, according to the International Energy Agency the bioenergy potential of cow manure will have an increase of 250% with respect to bioenergy production from biogas, but there is little implementation of the design and operation of biogas plants [6].

An analysis performed on the gas components showed a significant volume of methane in sheep manure with 62.1%

and cow manure with 66.9%, resulting that cow manure is one of the best for obtaining biogas, compared to sheep and pig, since the latter is less productive for biogas, these alternative energy resources obtained will allow us to replace the depletion of fossil fuels [7]. This study aims to optimize time and resources to reduce methane in the environment.

II. MATERIALS AND METHODS

The design proposal aims to reduce greenhouse gas emissions produced in livestock farms, as mentioned above, and reduce the time it takes to obtain biogas. [12] The automated system means that fewer people are exposed to the treatment of waste, thus avoiding diseases. The team used the bio digestion method and the design is called batch biodigester, which is simple and economical unlike other industrial processes [8, 13, 16]. There are 3 main gases that make up the Methane fermentation residue component: and they are nitrogen, phosphorus and potassium which are different according to the manure of each animal and are shown in Table 1. [17].

TABLE 1. The nutrient content in the different manure [17].

Compost	N %	P%	K%
Cow	1	1.4	1.5
Horse	1	1.8	2
Chicken	2.3	2.5	3.4
Food Waste	5.66	1.92	0.52

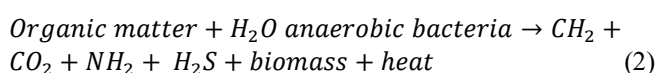
To estimate a saturation point between methanogens (catalysts) and manure we define that these organisms use H2 and CO2 as substrate for biogas production during their growth, considering also a suitable environment for their development or cultivation of these organisms and having a stable methane production or methanogenesis. The production of CH4 by methanogens requires the consumption of large amounts of sodium formate even for a modest cell growth of these organisms, which is the saturation point in the methanogenesis process, since the large amounts of sodium formate significantly accumulate NaOH (1), modifying the pH of the environment and turning it into an alkaline one, thus causing cell lysis and rapid death of the methanogens [18].



A. Biogas production process

The composition of biogas is mostly methane between 50% to 60% and it is produced thanks to the organic matter through anaerobic digestion, in addition once the natural gas was obtained we also have a residue that can be used as fertilizer, since this residue improves the solubility of mineral compounds in the soil [5, 13, 14].

On the other hand, anaerobic digestion is the interaction of microorganisms in the absence of oxygen, the chemical reaction shown in (2) [15].



Four stages are necessary for the process: hydrolysis, acidogenesis, acetogenesis and methanogenesis, which are consecutive phases and in each one different microorganisms participate; that is, in the first stage these beings decompose the larger wastes and when they are smaller, other types of microorganisms can take charge of their treatment [8].

There are factors that make the development of digestion more feasible and most of them are environmental, methanogens are susceptible to these changes and one of the parameters that has greater influence is the temperature since it alters the activity of enzymes, so a mesophilic temperature of between 30°C to 40°C should be maintained, which accelerates the fermentation process [15].

B. Digester system design

Through the simulation performed in the Factory I/O software we have the inlet of the tanks which we will call biodigesters, there are also the discharge ducts of these, to refer to the mixture of waste and water we use the word affluent, we also find the outlet valve for the biogas.

For the implementation of the biodigesters, the total volume to be housed was considered and is given in (3) where V is the useful volume and Vg is the volume of biogas produced.

$$V_T = V_L + V_G \quad (3)$$

For an optimal retention it is necessary to know the time this variable will be called HRT and are the days required for the organic mass or influent to be consumed by the bacteria, then the design proposes a daily tank filling, so we have an Q_{affluent} (m³/day) then to know the useful volume we use (4).

$$V_D = V_{Useful}(m^3) = TRH \text{ (days)} \cdot Q_{Influent} \left(\frac{m^3}{day} \right) \quad (4)$$

The biodigesters will have a pressure gauge in order to activate the tank discharge process, since the retention time will not always be met, this can be less and will work with a range: if the pressure is very low, the discharge ducts will be opened with a 10-minute advance notice; the same will happen once the pressure of the storage tank and the biodigester are 80% full [14, 15].

On the other hand, it was important to analyze various parameters of the raw material, for which Table 2 was taken as a reference; in addition, it should be noted that the optimum hydraulic retention time is in the range of 25 to 30 days [19].

TABLE 2. Raw material design parameters [19].

Parameter	Value
Average daily generation rate	564 kg/dia
Volumetric feed flow rate	0,4 m3 /dia
Total Solids (TS)	18,9 %
Volatile Solids (VS) (% TS)	80 %
Moisture content	81,1%
In Situ Density	1410kg/m3
Optimum Hydraulic Holding Time (HRT)	25-30 dias
Optimal organic loading rate	5-10 Kg SV/m3

As for the return on investment, the financial viability of biogas to produce biomethane at different scales needs to be looked at, the results are expected to help investors and governments to make decisions. The results obtained show that 'there are 10 potential areas of biomethane produced in Thailand, with a total biogas production capacity of 6,000 m³/day or more, with the condition that the pipeline does not exceed 50 km. Compressed biomethane plants with a capacity of less than 6 tons/day should be financed by the Government with 30% of the total investment capital with a payback period of 5-6 years. Plants producing more than 6 tons/day offer a good return on investment even without public capital [20].

First of all, to establish the development of the process in Fig. 1, it was essential to start the system by evaluating the position of the manure, on which depends the actuation of the pistons and belts, which in turn will move the manure to the extraction zone, thus initiating the methane extraction process; additionally, according to the ideal temperature, the heater or agitator is activated to finally store the manure in the tank.

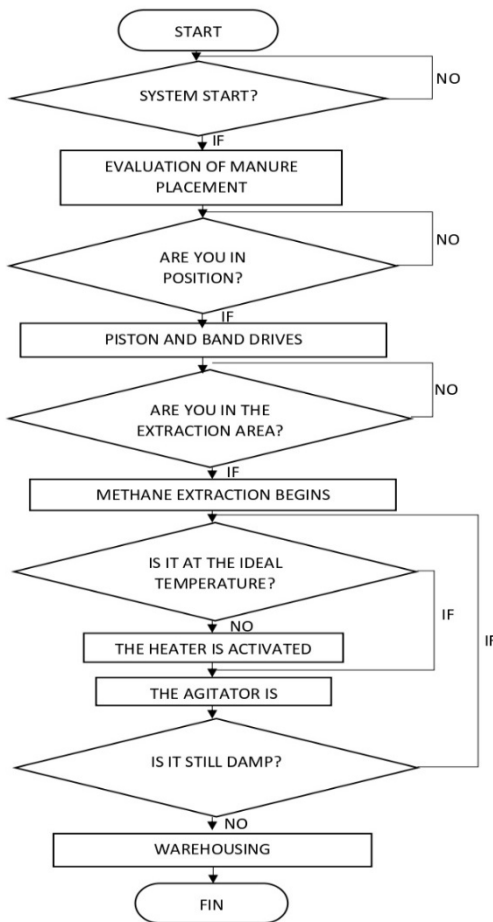


Fig. 1. System flow diagram.

III. RESULTS

Fig. 2 shows in the drivers the sensors and actuators used in the system by means of the Factory IO software; in addition to using the connection to the OPC server and

performing the simulation in the TIA Portal software. On the other hand, infrared sensors, pushbuttons (start and stop), conveyor belts, electric pistons and a 2 GDL robotic arm were used.

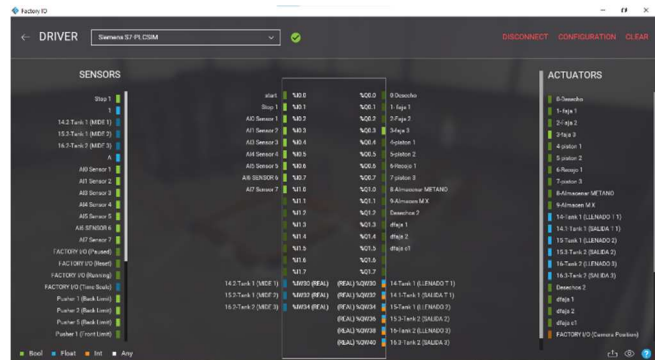


Fig. 2 PLC inputs and outputs in Factory IO software.

In Fig. 3, the programming of the 3 tanks of the system was carried out, which were set within the range of 0 to 10 volts (v); therefore, in the first place, a determined value of 6v was set, which indicates that the input flow rate is 60%. For the second tank, a determined value of 4v was set, which represents 40% of the input flow rate. Finally, the third tank worked with an aggregate value of 2v which revealed that the inlet flow rate is 20%.

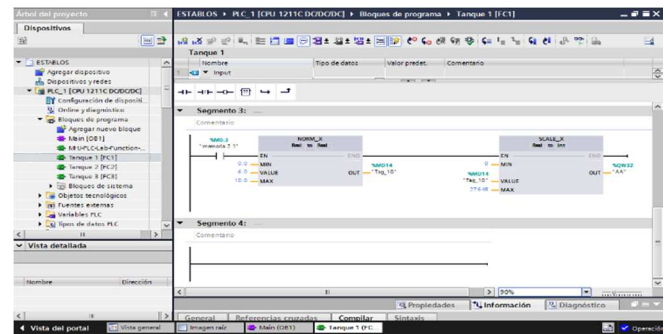


Fig. 3. Block programming of the tanks.

Fig. 4 shows the wet manure store, since it stores the wet feces, which will be detected by the infrared sensor, which will then drive the electric piston to move the feces on the conveyor belt, and the second sensor detects the feces, which generates the drive of the second piston to finally take the feces to the extractor.

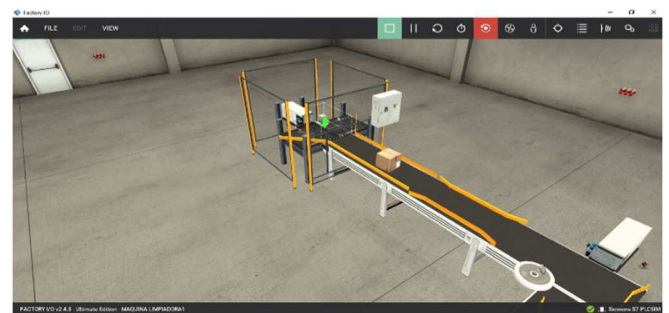


Fig. 4. Top view of the wet manure storage area.

Fig. 5 shows the manure detected by the third sensor which actuates the robotic arm, this represents the manure extractor, thus starting the filling of tank 1.

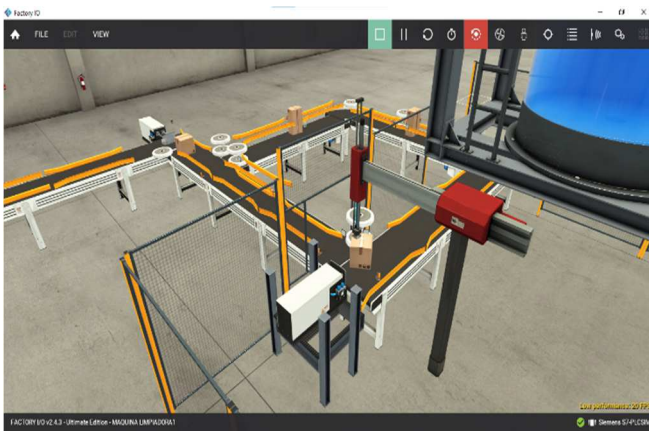


Fig. 5. Isometric view of the manure extractor.

Fig. 6 shows the filling of the tanks with methane in different percentages; for example, in the first process methane was obtained with a mixture of manure. In the second process, a filtering process is carried out to obtain a higher percentage of methane. Finally, in the third tank, a higher proportion of methane gas is obtained through a second filtration.

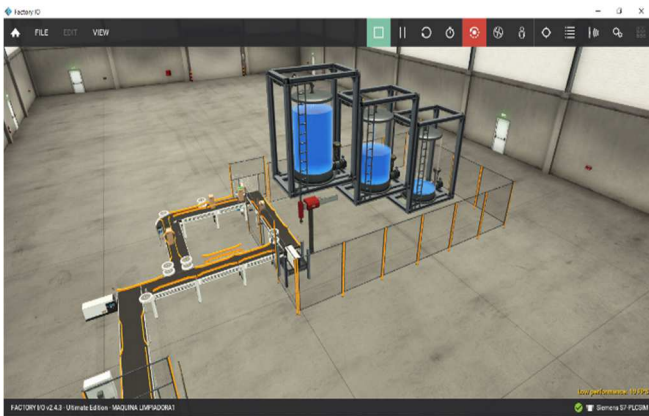


Fig. 6. Methane tanks.

As shown in the following image, an interface is implemented for a better representation of the automated system, where the process of the second and third tank is represented, for which the first tank stores the manure of livestock animals and through a filter extracts only the liquid to be processed by methanogenic organisms, which as mentioned above are responsible for producing methane [5], in addition these must be in constant movement and in a temperature range of 30° to 40° so that at higher temperature there is an acceleration of methane [15], a motor will be used that can be connected to blades to move the liquid, in addition to having a system to maintain an ideal heat and cooling for methanogenic organisms. Subsequently, as the methane is stored, the percentage of the third tank will be observed, but the pressure sensor will let the methane pass only when the second tank is greater than 90% of its capacity and then the second tank will be emptied. When more than 95% of the capacity of the third

tank is stored, the system stops completely to remove the methane gas.

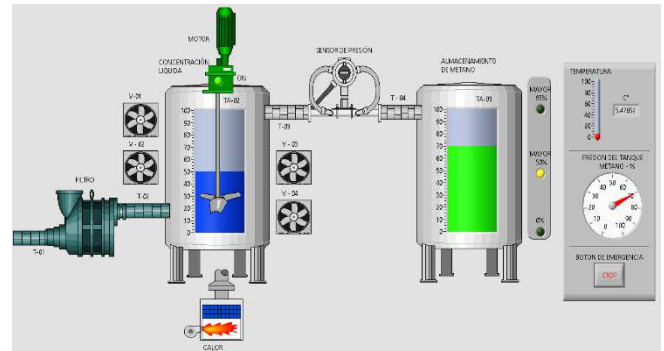


Fig. 7. Representation of the extraction of tank 2 and tank 3.

Subsequently, the programming for the process is shown, first, for the second tank starts with the filling through the filter, when it is over 40% of its capacity the engine is turned on to mix the liquid, then when it passes 90% of its capacity the filling and filtering is stopped, also oscillating temperatures are programmed to see the heat and cooling system for which a range of 35° to 40° was set so that the organisms produce the methane in less time. Secondly, when the second tank reaches 90% the sensor lets the methane pass to the tank which is stored and is reflected in the indicators of the third tank, once it reaches over 95% the system stops to avoid damage to the tank and the gas can be extracted without any inconvenience for its use.

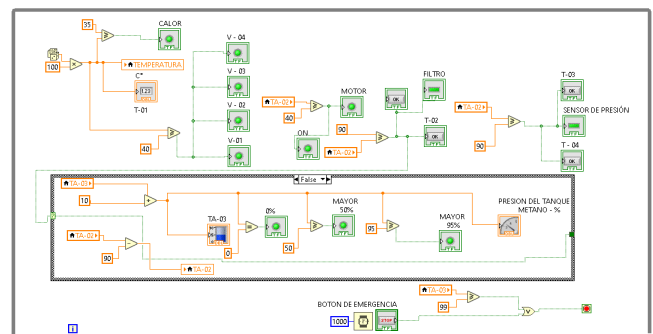


Fig. 8. LabVIEW programming.

In Fig. 9, sensor 4 detects the moisture in the manure, which is directed to the system's feedback belt, where the manure returns to the robotic arm (methane extractor) for extraction. Finally, the dregs are directed to the dregs tank.

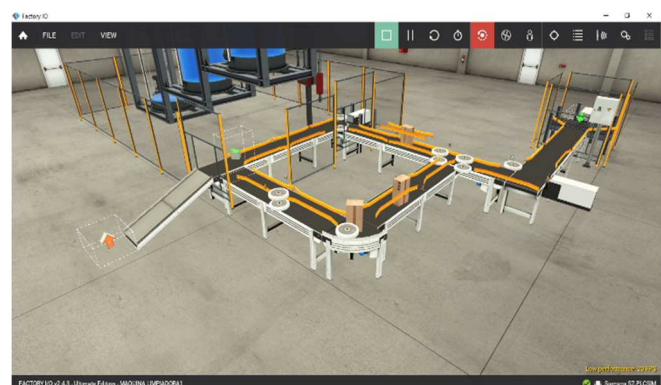


Fig. 9. Wet manure feedback.

In Fig. 10, the HMI environment visualizes the emergency start-up and shutdown of the system and the simulated methane extraction at various levels for each tank.

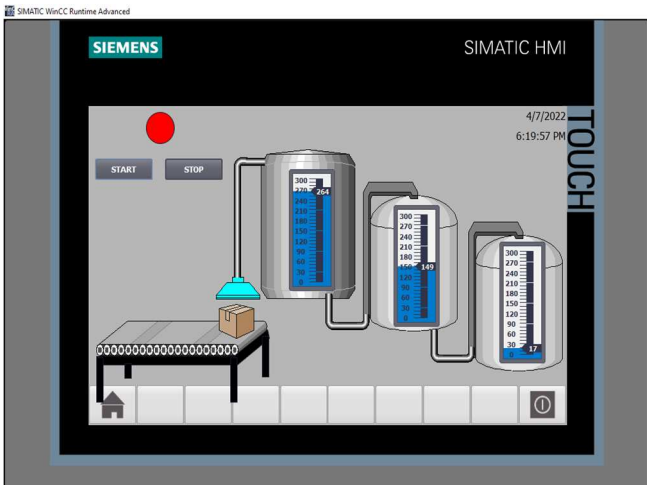


Fig. 10. HMI environment of the system.

Fig. 11 shows the execution of the system programming, where the actuation of the start control and emergency stop control is shown.

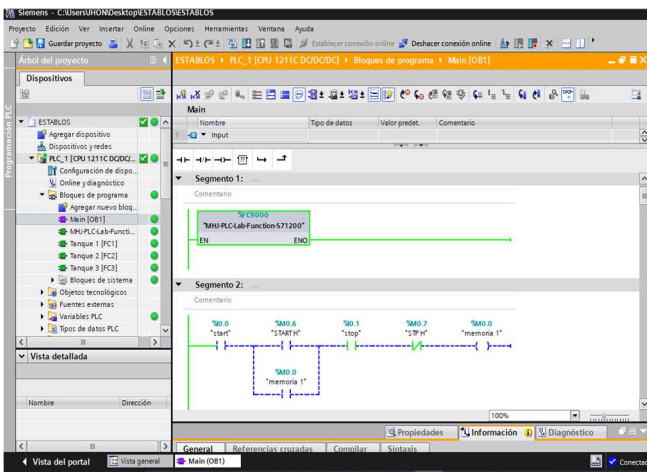


Fig. 11. Compilation of the PLC program in Ladder language.

In Fig. 12, a mushroom button was included on the control panel to interrupt the process or perform an emergency stop. There is also an on/off selector switch for the entire system.

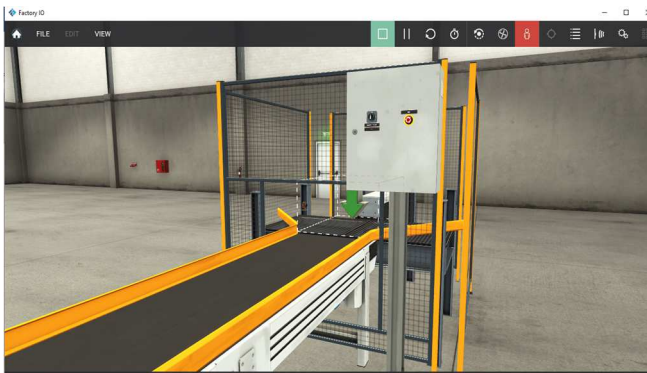


Fig. 12. Control panel design.

IV. DISCUSSIONS

A vacuum depression methane gas enrichment and separation (PSA) system was presented, demonstrating that this type of system can improve the efficiency of methane absorption, separation, and recovery in a coal mine ventilation system. For the development of this system, 2 absorption towers, vacuum pump, sensors and a PLC module were used to process the data from the control, measurement and piping network subsystems and import this data to a computer. A simulation was performed in Aspen Adsim software showing that the proportion of methane decreases with increasing feed concentration with a pressure and adsorption time of 210 kPa and 120 s respectively in the simulation; while in the field experiment, it increased about 30 s the adsorption time and methane concentration by 0.3%. They concluded that this system can reduce the investment cost and energy consumption, significantly improve the adsorption, separation and recovery of methane gas and thus protect the environment [1].

On the other hand, the research "Solution of environmental problems of livestock farming through the intensification of waste treatment processes" developed the design of a bioreactor, introducing bacteria for methane formation. It was carried out in a laboratory using a 50-liter bioreactor and later more components were added to these for improvement. The process starts by introducing microorganisms and manure into the bioreactor with a moisture content between 80 - 85% and continuously recirculating and fermenting for 28 days in mesophilic mode until biogas is formed and repeating the same procedure successively. The results of the research showed that the maximum percentage of methane in the biogas is reached on the seventh day (68%), and a stable mode of biogas formation is established on the 9th-10th day of fermentation. It was concluded and recommended that this bioreactor should be further upgraded to increase productivity and reduce HRT by using a leaching layer and immobilization of microorganisms [8].

When comparing the design presented in this research and the system of enrichment and separation of methane gas by vacuum depression oscillation (PSA) used 2 absorption towers for methane collection which does not have an automated process while the system presented in this research is automated and has 3 tanks that are at different levels, thus separating the processes in each tank for better quality of biogas, including a feedback process to obtain more methane from the manure that has already passed through each tank; This reduces and optimizes the time and amount of methane collected. Consequently, better environmental and economic benefits are obtained. Regarding the bioreactor design on the solutions of environmental problems of livestock farming through the intensification of waste treatment processes this design is not automated and invests considerable time in methane collection, since the research was developed in a laboratory and the bioreactor was not fully developed this would be limited for industrial processes, while the design presented in this article meets these requirements of control processes and automated design.

V CONCLUSIONS

Today it is known that one of the main pollutants is livestock; with which more than 1 million tons of manure is produced per year and having a potential of 21 times with CO₂ producing methane and 310 times more N₂O; furthermore,

enteric methane constitutes 17% of the global emissions produced by these animals. Environmental problems develop due to the manure of livestock animals due to inefficient systems in the respective control and management. This problem, in developing countries, is due to the limitation of economic and technical resources.

In the present investigation, the results obtained showed a control system focused on the collection of manure in cattle stables, since this contributes to avoid diseases and reduces the methane produced by the manure; in addition to optimizing the collection and storage time, it also reduces the percentage of methane gas emission by the fermentation process to the environment generated by the cattle.

The research presents a human-machine interface (HMI) system that makes it reliable and useful for users, from the evaluation process in real time to the identification of a problem or error in the system as a whole. For this, the Factory IO and TIA Portal software were used to simulate the process, which begins with the positioning of the manure to be transferred through a conveyor belt, in order to reach the gas extraction process and then pass by 3 cylinders that are distributed in such a way that the extraction time is accelerated, where the first cylinder will have an inlet flow of 60%, for the second 40% and the last 20%, the connection of the first to the second cylinder will have a filter separating the solids from the liquids, once the liquid is obtained at 40% level, it will be stirred by a motor connected with vanes, at the same time the temperature will be controlled in the range of 35 °C to 40 °C so that the microorganisms work in greater gas production, in an excess of temperature the fans will be activated and otherwise a heater will be activated. Finally, the manure will continue with the help of a conveyor belt to an exit control where it will be stored.

The design presented optimizes manure collection and storage time and automates manpower in the process, since it minimizes the exposure of livestock workers to diseases, since the precision of manure transport is maintained, avoiding spills and human contact.

The motivation for presenting this design was to support farms that produce milk in large quantities; These farms make cheese, yogurt and different products derived from milk. The objective is to avoid the diseases that are generated by exposure to the manure of bovine animals; In addition, given that there is a lack of information on the subject, a collection and storage system was developed based on the problems faced by agricultural workers.

REFERENCES

[1] T. Zhu *et al.*, "Enrichment and Separation of Methane Gas by Vacuum Pressure Swing Adsorption," *Adsorpt. Sci. Technol.*, vol. 2021, 2021, doi: 10.1155/2021/5572698.

[2] A. Pramono, T. A. Adriany, H. L. Susilawati, Jumari, and I. F. Yuniarti, "Alternate wetting and drying combined farmyard manure for reducing greenhouse gas while improving rice yield," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 950, no. 1, pp. 1–8, 2022, doi: 10.1088/1755-1315/950/1/012012

[3] J. Zhou, G. Liang, T. Deng, S. Zhou, and J. Gong, "Coalbed methane production system simulation and deliverability forecasting: Coupled surface network/wellbore/reservoir calculation," *Int. J. Chem. Eng.*, vol. 2017, 2017, doi: 10.1155/2017/8267529.

[4] F. Forabosco, Z. Chitchyan, and R. Mantovani, "Methane, nitrous oxide emissions and mitigation strategies for livestock in developing

countries: A review," *South African J. Anim. Sci.*, vol. 47, no. 3, pp. 268–280, 2017, doi: 10.4314/sajas.v47i3.3.

[5] Sarah, H. L. Susilawati, and A. Pramono, "Quantifying the potency of greenhouse gas emission from manure management through anaerobic digester in Central Java," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 648, no. 1, 2021, doi: 10.1088/1755-1315/648/1/012111.

[6] J. A. Silva-González, I. O. Hernández-De Lira, A. Rodríguez-Martínez, G. A. Ruiz-Santoyo, B. Juárez-López, and N. Balagurusamy, "Design of a centralized bioenergy unit at comarca lagunera, Mexico: Modeling strategy to optimize bioenergy production and reduce methane emissions," *Processes*, vol. 9, no. 8, 2021, doi: 10.3390/pr9081350.

[7] P. E. Kiat, M. A. Malek, and S. M. Shamsuddin, "Artificial intelligence projection model for methane emission from livestock in Sarawak," *Sains Malaysiana*, vol. 48, no. 7, pp. 1325–1332, 2019, doi: 10.17576/jsm-2019-4807-02.

[8] Z. K. Bakhov, K. U. Sultangalieva, N. B. Zhumadilova, V. E. Drevin, and G. L. Gizzatova, "Solution of environmental problems of livestock breeding through intensification of waste treatment processes," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 659, no. 1, 2021, doi: 10.1088/1755-1315/659/1/012008.

[9] S. E. Hook, A. D. G. Wright, and B. W. McBride, "Methanogens: Methane producers of the rumen and mitigation strategies," *Archaea*, vol. 2010, pp. 50–60, 2010, doi: 10.1155/2010/945785.

[10] S. Hu *et al.*, "Methane Extraction from Abandoned Mines by Surface Vertical Wells: A Case Study in China," *Geofluids*, vol. 2018, 2018, doi: 10.1155/2018/8043157.

[11] [M. Luqman and T. Al-Ansari, "A novel solution towards zero waste in dairy farms: A thermodynamic study of an integrated polygeneration approach," *Energy Convers. Manag.*, vol. 230, no. October 2020, p. 113753, 2021, doi: 10.1016/j.enconman.2020.113753.

[12] F. K. Gemechu, "Evaluating the Potential of Domestic Animal Manure for Biogas Production in Ethiopia," *J. Energy*, vol. 2020, pp. 1–4, 2020, doi: 10.1155/2020/8815484.

[13] Y. Wang, Y. Zhang, J. Li, J. G. Lin, N. Zhang, and W. Cao, "Biogas energy generated from livestock manure in China: Current situation and future trends," *Journal of Environmental Management*, vol. 297, p. 113324, Nov. 2021, doi: 10.1016/J.JENVMAN.2021.113324.

[14] J. Zhang *et al.*, "Enhancing biogas production from livestock manure in solid-state anaerobic digestion by sorghum-vinegar residues," *Environ. Technol. Innov.*, vol. 26, p. 102276, 2022, doi: 10.1016/j.eti.2022.102276.

[15] M. Palù *et al.*, "In-situ biogas upgrading assisted by bioaugmentation with hydrogenotrophic methanogens during mesophilic and thermophilic co-digestion," *Bioresour. Technol.*, vol. 348, no. December 2021, p. 126754, 2022, doi: 10.1016/j.biortech.2022.126754.

[16] M. A. Barrera Gurbillón, F. Cubas Alarcón, W. Gosgot Angeles, C. M. Ordinola Ramírez, J. Rascón Barrios, and M. Huanes Mariños, "Sistema de producción de biogás y bioabonos a partir del estiércol de bovino, Molinopampa, Chachapoyas, Amazonas, Perú," *Arnaldoa*, vol. 26, no. 2, pp. 725–734, 2019.

[17] N. Santi, R. K. Dewi, Y. Suganuma, T. Iikubo, H. Seki, and M. Komatsuzaki, "Methane fermentation residue compost derived from food waste to aid komatsuna (*Brassica rapa*) growth," *Horticulturae*, vol. 7, no. 12, pp. 1–12, 2021, doi: 10.3390/horticulturae7120551.

[18] F. Long, L. Wang, B. Lupa, and W. B. Whitman, "A Flexible System for Cultivation of Methanococcus and Other Formate-Utilizing Methanogens," *Archaea*, vol. 2017, 2017, doi: 10.1155/2017/7046026.

[19] I. Valela, E. Muzenda, and E. Muzenda, "Design of a Biodigester to Treat Cow Dung in Botswana," *Proc. 2019 7th Int. Renew. Sustain. Energy Conf. IRSEC 2019*, Nov. 2019, doi: 10.1109/IRSEC48032.2019.9078244.

[20] P. Tonrangklang, A. Therdyothin, and I. Preechawuttipong, "The financial feasibility of compressed biomethane gas application in Thailand," *Energy, Sustainability and Society*, vol. 12, no. 1, pp. 1–12, Dec. 2022, doi: 10.1186/S13705-022-00339-3/FIGURES/11.

Design of an Automated Feeding and Drinking System for Turkeys in Different Stages of Development

Yadhira S. Valenzuela-Lino
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
73105932@continental.edu.pe

Jesús Eduardo Rosales-Fierro
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
76560426@continental.edu.pe

Jhon Rodrigo Ortiz-Zacarias
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
71689110@continental.edu.pe

Nabilt Moggiano
 Engineering Research Unit, Faculty of
 Engineering
 Universidad Continental
 Huancayo, Perú
nmoggiano@continental.edu.pe

Carlos Alberto Coaquira-Rojo
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
ccoaquira@continental.edu.pe

Deyby Huamanchahua
 Department of Mechatronic
 Engineering
 Universidad Continental
 Huancayo, Perú
dhuamanchahua@continental.edu.pe

Abstract—Over the years, various activities within the poultry industry have automated their processes to optimize the breeding time of birds; Therefore, this research presents the design of an automated food and beverage system for turkeys in the different stages of its development. To obtain the sequence of the process, a flow chart was developed, the same one that structures the various phases of the system; In addition, to present the nutrients and additives present in the turkey feed, the Factory IO software was used as a simulation tool; on the other hand, the programming of the process was carried out in the TIA Portal software. As a result of the above, three different types of rations were obtained, which will be distributed according to the weights of the turkeys that are within the ranges (0-500) g, (500g-10kg), and (10kg-20kg) respectively; Additionally, each ration contains different Selenium (Se) additives, which is used to enhance the growth of turkeys. Finally, an automated feeding and drinking system for turkeys at different stages of their development was designed by structuring the flow diagram in each phase, which allowed us to optimize the feed supply of turkeys efficiently.

Keywords—Automated system, turkeys, poultry, breeding, optimization.

I. INTRODUCTION

Poultry farming is an economic activity that consists of raising birds such as chickens, ducks, turkeys, and geese to obtain meat and/or obtain eggs for human consumption [1]. However, the conventional poultry rearing system employs traditional free-range rearing; that is, poultry are raised outdoors; therefore, they roam freely during the day. It is evident that such activity requires a higher workload for farmers since the birds need to be fed regularly to be more productive [2]; furthermore, the conventional method is based on a manual feeding system, which requires human labor; therefore, farmers cannot guarantee that all poultry receive similar amounts of feed [3].

On the other hand, the breeding process for turkeys has still been using the conventional method, since it does not delve into automated feeders for this species; on the other hand, turkeys are usually raised on average for up to 18-20 weeks in the case of females and 24 weeks for males, where it is sought to supplement their diet with leaf meal to raise zinc

levels in meat [4]; In addition, as production increases, it is more difficult to maintain clean spaces for birds, therefore they tend to suffer the effects of mycotoxins, especially in the first 6 weeks of life which leads to reduce the performance of animals [5]. Another harmful pollutant for free-range turkeys is *Aspergillus* since it generates a high rate of morbidity and mortality in birds since it produces an acute suppression of the immune system which leads to lower production [6-14].

From another perspective, the integration of automation and new sensory technologies to poultry breeding allows the optimization of time [7] and in turn makes effective the feeding process in which different additives can be added to improve the meat [8].

Faced with this problem, the design of an automated food and drink system for turkeys was proposed at the different stages of its development, this system aims to evaluate by the weight of the turkeys the type of ration that corresponds to it, in this way the various additives are classified, for example, selenium nanospheres (SeNP), organic selenium (Sel-Plex®) and inorganic selenium (Se (IV)), which will be used in their diet [9-15]. In addition, to observe the proper functioning of the system before its implementation in a real environment, its simulation was essential which was carried out through the Factory IO software and TIA Portal.

The distribution of the article is as follows. In section II, the flowchart for the system process was made to order the progress stages of the system, as well as the ingredients, nutrients, and additives of the food on the tables; on the other hand, in section III the results of the simulation are presented using the Factory IO software; in addition to presenting various graphic environments. In Section IV, discussions on the automated system were presented to finally give way to the conclusions of the investigation in Section V.

II. MATERIALS AND METHODS

The process begins by introducing the number of turkeys that will be fed in an HMI graphic environment, in which programming will obtain the respective calculations according to the type of food to be dispensed; therefore, the system was established in the initial, intermediate, or adult stages having as a reference an analog input that converts the

potential differential into grams (g) and kilograms (kg). These references are expressed in ranges (0 -500g), (500g-10kg) and (10kg-20kg) respectively.

According to studies carried out regarding the methodology of feeding and improvement of this process, certain breeds of turkeys undergo a partial amputation of the beak, in which the lower beak is usually left longer than the upper beak, to obtain a better reception of food and avoid possible diseases that may develop throughout the stages of growth of these [10]. On the other hand, to favor the design shown in this article, it was sought that the turkeys obtain a better reception of food which reduces the waste of this in each process to optimize the control of feed rations in daily quantities.

First, to develop the process it was essential to establish the start of the system by identifying the weight of the turkey, which will then go through the second water dispenser process so that it finally goes through the third process of food distribution in rations according to the range established and calculated in the programming as shown in Fig. 1.

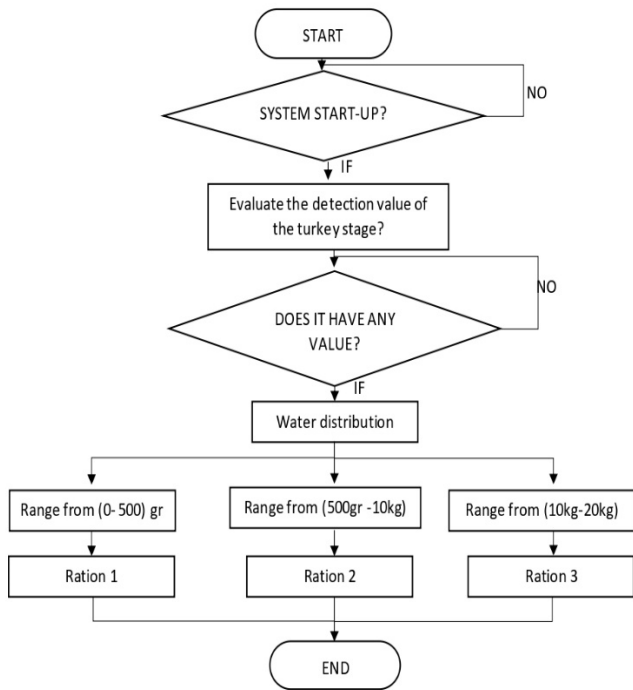


Fig.1 System flowchart.

In automated systems and the representation of these, another study in cattle feeding has used the software (TIA Portal and Factory I / O) in this system the dispensing and transport of food was done in 3 processes, these processes were established according to the amount of feed required by the cattle based on their weights entered, a transformation of a difference in potential to kilograms was made. In this design you could choose 3 portions of feed already balanced with the nutrients that the cattle may require these were 1/2 kg, 1 kg, and 2 kg, leaving three conveyor belts and being mixed later in a single transport belt reaching a scale that through an HMI screen monitored the amount of food based on the weight of each cow and thus ending with the feeding process the Food is transported to feeding points [11]. Developing this design

in a simulation environment with the software makes our research methodology more reliable during the execution of our processes.

Table I is a reference to turkey diets, which were formulated from the recommended nutrient requirements for poultry according to the National Research Council (NRC) was established in 1960 and made constant changes until 1994 [12].

Table I. Ingredients and nutrients were calculated for turkey growth and breeding [9].

Ingredients, %	melancholic 1-2 weeks	growing 2.- 8 weeks
Yellow corn	50	69
Soybean meal, 44% PB	39	20
Pisces, 64%	10	10
PB Dicalcium Phosphate	-	0.3
ground limestone	0.4	0.1
DL-methionine	-	0.3
L-lysine	0.1	0.1
premezclal	0.25	0.1
NaCl	0.25	0.25
Total	100	100
Calculated nutrient levels		
AME, Kcal/Kg	2830	3000
Crude protein, %	27.5	20.87
Extract any, %	3.15	3.72
Crude fiber, %	3.9	2.9
calcium, %	0.79	0.72
Phosphorus available, %	0.43	0.42
Lysine, %	1.8	1.38
Methionine, %	0.78	0.8
Methionine + Cysteine, %	0.9	0.82

Table II shows the effect of different biologically produced forms of Selenium (Se) on growth performance in experimental diets, in which initial body weight, body weight gain, and food consumption are evidenced. This study was developed for 6 weeks with values that express as a measure (±) standard error; In addition, the groups were called: selenium nanospheres (SeNP), and organic selenium (Sel-Plex®), and inorganic selenium (Se (IV)). On the other hand, feed conversion rate (FCR) was also considered [9].

Table 2. The effect of different forms of selenium on the feeding and growth of turkeys [9].

Article	Experimental diets				
	Control	Simple	Selplex	It is (IV)	
Corporate Starting Weight, g	56,79 ± 1,45	57,16 ± 1,25	56,87 ± 1,17	56,97 ± 1,17	0.82
final body, g	4182,35 ± 274,54	4236,90 ± 202,23x	4037,50 ± 261,09	4179,75 ± 290,74	0.153
body weight, g	4125,56 ± 274,63	4179,75 ± 225,61	3980,63 ± 261,47	4122,77 ± 291,10	0.155
feed, g/6 weeks FCR,	14,288,00 ± 16,92un	13,752,00 ± 16,92b	13,483,00 ± 16,92C	13,736,00 ± 16,92b	0.0001
g/g	3,48 ± 0,22un	3,22 ± 0,18b	3,46 ± 0,24un	3,28 ± 0,23b	0.001

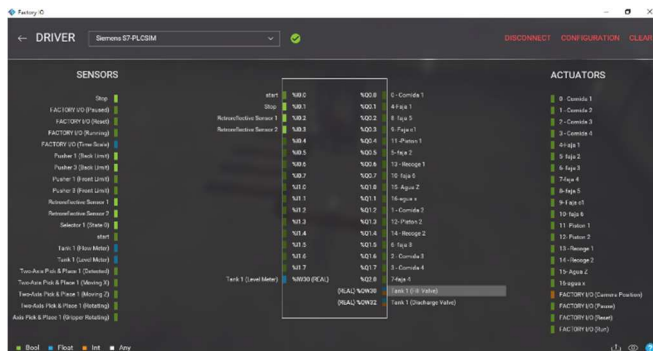
According to the percentage of ingredients and nutrient levels already calculated in Table I; In addition, in conjunction with the sources of selenium in Table II, a

nutritious ration was formed, which stimulates the accelerated metabolism and growth of the turkey.

On the other hand, the automated food and drink system for turkeys at different stages of its development consists of a process for the dispensing of 3 types of food which each attaches various additives, for example, ration 1 contains selenium nanospheres (SeNP), ration 2 includes organic selenium (Sel-Plex®) and ration 3 incorporates the inorganic selenium (Se (IV)) together each serving integrates constant water through the dispenser.

In another order of ideas, in this research, an ON/OFF control system [16] was applied, to monitor the feed distribution in the feed and water dispenser for the turkeys. In addition, when the analog inputs are entered as data to the HMI screen, the programming in Ladder language performs the conversions previously entered to obtain the value in kg of turkeys, consequently, such actuation activates and deactivates the actuators of the system.

Fig. 2 shows the drivers used in the Factory IO software, which will be used for connection through the OPC server; in addition to including the TIA Portal software for the automation of the system, which consists of 6 tapes, infrared sensors, water tank, a 2 GDL arm which simulates a water dispenser, electric piston and an HMI (Human-Machine Interface).



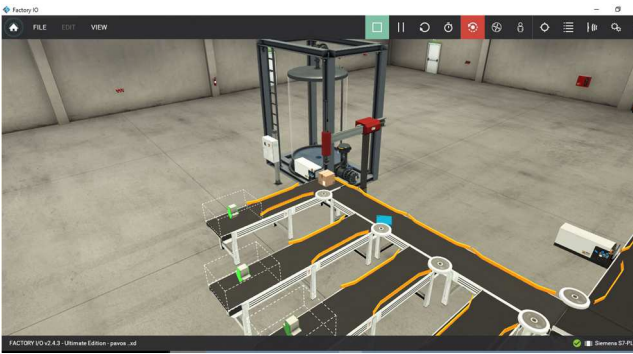


Fig. 6. Representation of the girdles in the Factory IO environment.

To experiment in Fig. 7, an analog input of 13500 was introduced, where weight of 9830g was obtained, which is located within the second established range (500g -10kg).

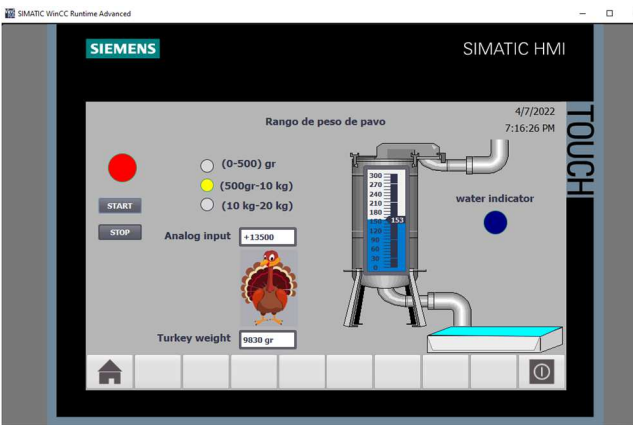


Fig. 7. HMI graphic environment.

In Fig. 8 the process was executed according to the data entered in Fig. 7; in addition, the green box represents the "2nd serving" of turkey food.



Fig. 8. Top view of the girdles in the Factory IO environment.

Finally, to experiment with the last established range, in Fig. 9 an analog input of 26000 was introduced, where a weight of 18932g was obtained, which is located within the third established range (10kg -20kg).

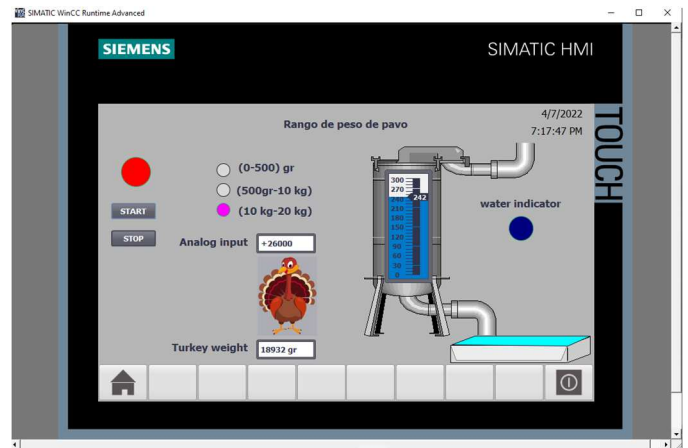


Fig. 9. HMI graphic environment.

Fig. 10 shows the process according to the data entered in Fig. 9; in addition, the gray box represents the "3 servings" of turkey food.

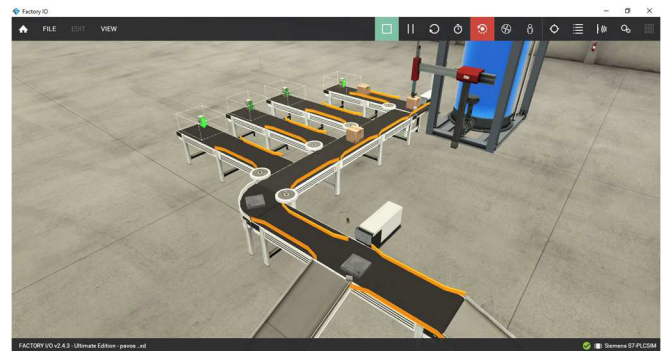


Fig. 10. Upside view of the girdles in the surroundings of Factory IO.

Fig. 11 reflects the virtual environment of the girdles and the water dispensing tank, which is the first to be activated independently of the various analog inputs introduced through the HMI graphic environment since it is indispensable for the diet of poultry (turkeys).

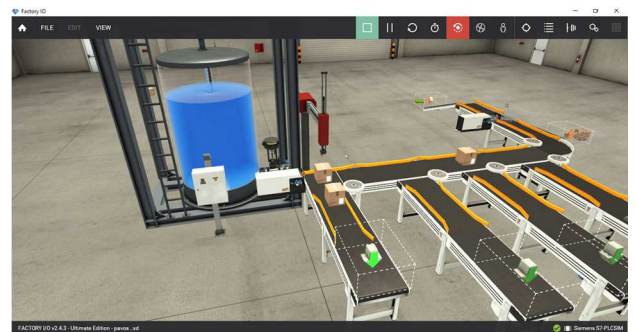


Fig. 11. Graphic representation of the water dispenser.

In Fig. 12 you can see the two environments together since both are related through the data entered in the analog inputs; that is, according to the information entered, the weight of the turkey will be located within the corresponding range and then start the process of water and food supply in the Factory IO environment.

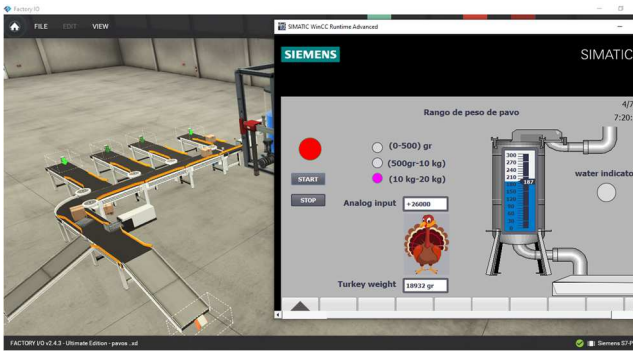


Fig. 12. HMI environment and simulation of the dispensing belts through Factory IO.

Fig. 13 shows the execution of the system programming; where system boot control and emergency control are displayed; in addition, a similar interface was used for each march.

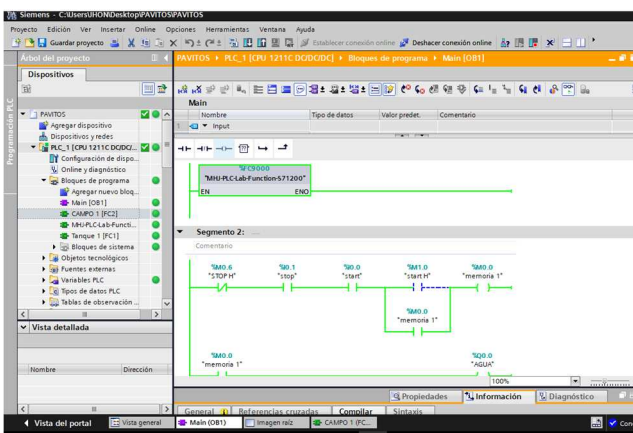


Fig. 13. Compilation of PLC programming in TIA Portal software.

Finally, in case of incidents with the actuation of the sensors or accumulation of the "rations" of food, there is a control box that includes a mushroom button whose purpose is to interrupt the processor also called an emergency stop. On the other hand, there is an on and off selector switch for the entire system as evidenced in Fig. 14.



Fig. 14. Design of the control box of the water and food dispenser.

IV. DISCUSSIONS

According to the research [13], an intelligent system for monitoring the feeding environment of livestock and poultry-based on Bluetooth with minimum energy was designed and implemented where data is monitored and collected with multiple wireless sensor networks. The system was

developed on a pig farm with an effective test time of 276 hours, testing the distance and stability of communication and packet loss rate of the gateway and obtaining a distance of 20.33 m that meets the requirements for monitoring. It was concluded that the gateway design, priority queue, and feedback mechanism are introduced in innovative ways, which effectively solves communication distance problems for Bluetooth devices, the method used is effective in high concurrency data, processing, and response to abnormal environmental changes.

According to the study [2], they developed a feeder system for poultry using a line-following robot to replace humans in feeding poultry by moving along. The development of the system is divided into the bird feeder, the design of the feeder robot, and the alert of notifications by the buzzer. The experiment shows how the feeder robot followed the line and stopped at each intersection to expel the food. During the test, the sensitivity of the IR sensor can be adjusted using the potentiometer to test the efficiency of the IR sensors to detect the lines. It was concluded that the performance of the proposed system has been experienced in a real environment and makes it easier for users to interact with low development costs.

Regarding the design and implementation of an intelligent gateway system for monitoring the feeding environment of livestock and poultry-based on Bluetooth, the design of this research has an automated and selective process through conveyor belts for feeding turkeys which include rations enriched with selenium (Se) and a constant water supply. e, to provide a complete diet; therefore, labor costs in transportation are reduced. Considering the system of a feeder for poultry using a line-following robot this system is the linear movement of the robot and lacks a supply of liquid being this indispensable for the development of the birds, by depending on a battery the activities and efficiency of the robot are reduced according to the level of a load of this, and compared to the design presented which has a constant process of transport, supply of food and drink, this would be more efficient and would reduce the cost considerably.

V. CONCLUSIONS

According to the simulation made with the TIA Portal software, linked to the Factory I/O, our design of an automated feeding and drinking system for turkeys in the different stages of its development shows an optimization in the supply and transfer of these to the feed points correctly and efficiently; Considering these aspects, it is concluded that this design reduces the cost of labor and the time to feed the turkeys according to their stage of development.

In the same way, the human-machine interface HMI facilitates the user's activities in the feeding process and selection of the type of food to be portioned, also avoiding possible errors that may occur in the process or external elements that harm the normal operation of the system.

In addition, this research shows through results a correct dosage of the 3 types of foods enriched with sources of selenium (Se)," using conveyor belts with timers and sensors, making the respective pauses for proper rationing thus avoiding surplus or lack of food.

On the other hand, implementation in the control system with a PID is recommended, since it would optimize the census and control of bird weights, in addition to rationing feed and drink with greater precision. Additionally, algorithms for temperature and humidity monitoring can be added to this type of control to offer a healthier and more comfortable environment for the birds; therefore, consequently, obtaining a better quality of turkey meat and increasing production through the automation of the processes.

Finally, it was concluded that the proposed system not only considerably reduces labor costs and time invested in the turkey feeding process, but also contributes favorably to industrialized processes, where parameters such as the individual weight gain of each turkey and the consumption of food, the periods desired by the poultry industry are estimated; since with this design, daily quantities of 3 types of food can be estimated with more precise rationing, avoiding surpluses that could cause possible monetary losses. That said, the system can compete in a national or international market in the poultry industry.

REFERENCES

[1] K. Sinduja, S. S. Jenifer, M. S. Abishek, and B. Sivasankari, "Automated Control System for Poultry Farm Based On Embedded System," *Int. Res. J. Eng. Technol.*, 2016, Accessed: Apr. 08, 2022. [Online]. Available: www.irjet.net.

[2] N. S. Ahmad et al., "Development of a poultry feeder system using line follower robot," *Int. J. Eng. Trends Technol.*, no. 1, pp. 173–179, Aug. 2020, doi: 10.14445/22315381/CATI3P227.

[3] siswanto imam santoso, T. A. Sarjana, and A. Setiadi, "Income Analysis of Closed House Broiler Farm with Partnership Business Model," *Bul. Peternak*, vol. 42, no. 2, pp. 164–169, May 2018, doi: 10.21059/BULETINPETERNAK.V42I2.33222.

[4] A. Sharma et al., "Effect of dietary supplementation of sea buckthorn and giloe leaf meal on the body weight gain, feed conversion ratio, biochemical attributes, and meat composition of turkey poults," *Vet. World*, vol. 11, no. 1, p. 93, Jan. 2018, doi: 10.14202/VETWORLD.2018.93-98.

[5] J. E. N. Tilley et al., "Efficacy of feed additives to reduce the effect of naturally occurring mycotoxins fed to turkey hen poults reared to 6 weeks of age," *Poult. Sci.*, vol. 96, no. 12, pp. 4236–4244, Dec. 2017, doi: 10.3382/PS/PEX214.

[6] C. Yang, G. Song, and W. Lim, "Effects of mycotoxin-contaminated feed on farm animals," *J. Hazard. Mater.*, vol. 389, p. 122087, May 2020, doi: 10.1016/J.JHAZMAT.2020.122087.

[7] Z. H. C. Soh, M. H. Ismail, F. H. Otthaman, M. K. Safie, M. A. A. Zukri, and S. A. C. Abdullah, "Development of automatic chicken feeder using Arduino Uno," *2017 Int. Conf. Electr. Electron. Syst. Eng. ICEESE 2017*, vol. 2018-January, pp. 120–124, Feb. 2018, doi: 10.1109/ICEESE.2017.8298402.

[8] M. T. Gilang Prasetyo Utomo, M. Munadi, "International Journal of Recent Technology and Engineering (IJRTE)," *Int. J. Recent Technol. Eng.*, vol. Volume-8, 2019, doi: 10.35940/ijrte.B3453.078219.

[9] S. E. Ibrahim, M. H. Alzawqari, Y. Z. Eid, M. Zommara, A. M. Hassan, and M. A. O. Dawood, "Comparing the Influences of Selenium Nanospheres, Sodium Selenite, and Biological Selenium on the Growth Performance, Blood Biochemistry, and Antioxidative Capacity of Growing Turkey Pullets," *Biol. Trace Elem. Res.* 2021, pp. 1–8, Aug. 2021, doi: 10.1007/S12011-021-02894-W.

[10] S. Grün et al., "Welfare and Performance of Three Turkey Breeds—Comparison between Infrared Beak Treatment and Natural Beak Abrasion by Pecking on a Screenshot Grinding Wheel," *Anim.* 2021, Vol. 11, Page 2395, vol. 11, no. 8, p. 2395, Aug. 2021, doi: 10.3390/ANI11082395.

[11] I. L. Q. Mosquera, J. E. R. Fierro, J. R. O. Zacarias, J. B. Montero, S. A. C. Quijano, and D. Huamanchahua, "Design of an Automated System for Cattle-Feed Dispensing in Cattle-Cows," *2021 IEEE 12th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2021*, pp. 671–675, 2021, doi: 10.1109/UEMCON53757.2021.9666491.

[12] M. A. Grashorn, "NUTRITIONAL REQUIREMENTS OF BROILERS WITH DIFFERENT GROWTH CAPACITY Summary."

[13] [Y. Du, G. Sun, B. Zheng, and Y. Qu, "Design and Implementation of Intelligent Gateway System for Monitoring Livestock and Poultry Feeding Environment Based on Bluetooth Low Energy," *Inf.* 2021, Vol. 12, Page 218, vol. 12, no. 6, p. 218, May 2021, doi: 10.3390/INFO12060218.

[14] Cheng, L., Qin, Y., Hu, X., Ren, L., Zhang, C., Wang, X., Wang, W., Zhang, Z., Hao, J., Guo, M., Wu, Z., Tian, J., & An, L. (2019). Melatonin protects in vitro matured porcine oocytes from toxicity of Aflatoxin B1. *Journal of Pineal Research*, 66(4), e12543. <https://doi.org/10.1111/JPL.12543>

[15] Oliveira, T. F. B., Rivera, D. F. R., Mesquita, F. R., Braga, H., Ramos, E. M., & Bertechini, A. G. (2014). Effect of different sources and levels of selenium on performance, meat quality, and tissue characteristics of broilers. *Journal of Applied Poultry Research*, 23(1), 15–22. <https://doi.org/10.3382/JAPR.2013-00761>

[16] A. H. Shatti, H. A. Hasson, and L. A. Abdul-Rahaim, "Automation conditions of mobile base station shelter via cloud and IoT computing applications," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 5, pp. 4550–4557, Oct. 2021, doi: 10.11591/IJECE.V11I5.PP4550-4557.

[17] S. I. Del Carpio Ramirez, J. R. O. Zacarias, J. B. M. Vazquez, S. A. C. Quijano and D. Huamanchahua, "Comparison Analysis of FIR, ARX, ARMAX by Least-Squares Estimation of the Temperature Variations of a Pasteurization Process," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0699-0704, doi: 10.1109/UEMCON53757.2021.9666649.

[18] A. H. Uribe, J. Brayan Macuri Vasquez, A. C. Miranda Yauri, and D. Huamanchahua, "Control and Monitoring System of Hydraulic Parameters for Rainbow Trout Culture," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0780-0784, doi: 10.1109/UEMCON53757.2021.9666512.

[19] J. P. A. Misajel, S. A. C. Quijano, D. M. C. Esteban, S. R. T. Rojas, D. Huamanchahua and R. A. M. Grados, "Design of a Prototype for Water Desalination Plant using Flexible, Low-Cost Titanium Dioxide Nanoparticles," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0286-0289, doi: 10.1109/UEMCON53757.2021.9666586.

Automated design of a cleaning machine and an environmental temperature controller for guinea pig houses

Yadhira S. Valenzuela-Lino
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 73105932@continental.edu.pe

Yossef Rojas-Tapara
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 72448224@continental.edu.pe

Jhon Rodrigo Ortiz-Zacarias
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 71689110@continental.edu.pe

Sliver Ivan Del Carpio-Ramirez
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 71902502@continental.edu.pe

Frank William Zarate-Peña
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 fzaratep@continental.edu.pe

Carlos Coaquira-Rojo
Department of Mechatronic Engineering
Universidad Continental
 Huancayo, Perú
 ccoaquira@continental.edu.pe

Abstract— Guinea pig breeding is an important livestock activity in Peru, since its consumption provides great nutritional value to humans; however, its production has been affected mainly by two factors: first, the lack of hygiene in the guinea pig sheds generates salmonella, which enters the guinea pig's body through the consumption of contaminated food and generates a rate of mortality from 15% to 18%; secondly, the excesses of temperature in the environment generate stress in the guinea pigs, where they end up dying and therefore the offspring are affected. For these two reasons, the automated design of a cleaning machine and an environmental temperature controller for guinea pig sheds was proposed, with the aim of optimizing the time of farmers and obtaining an ideal habitat for raising guinea pigs. The results obtained from the system are the programming and simulation of the cleaning machine through the Factory IO software, where the process ends with the storage of its excrement, which can then be used as fertilizer; On the other hand, a temperature controller was established that ranges between 24°C and 26°C, which were determined as ideal temperatures of the environment in which guinea pigs should be raised, said implementation was carried out through programming and simulation in the LabVIEW software.. Then, this research presents a novel solution to the various problems mentioned above.

Keywords— Guinea pigs, sheds, optimization, cleaning machine, temperature, humidity.

I. INTRODUCTION

The guinea pig is a hybrid rodent species that provides great nutritional value to the body if ingested as a food source. It is also a species with a higher productive efficiency in feed conversion compared to other animals in the livestock sector such as cows, horses and pigs [1]. For people involved in livestock farming, the raising of this species is very important, since it is being developed with great success, since its meat is export-oriented [2].

On the other hand, the young of the guinea pig are neonatal animals that are born in an advanced state of maturation; however, their mortality rate can reach 38% to 56% in family farms and in technified farms it can reach 23% [3], [4], the causes can be diverse; However, there is not enough research that analyzes the guinea pig breeding environment, since not frequently cleaning the sheds where they are raised is a reason

for the generation of various diseases that attack them and therefore the breeding of this species is affected [5].

Due to this context, it is important to study this problem, since there are very few studies oriented to these situations. For example, salmonella is a frequent infection due to inadequate house hygiene. A particular case is evident in Peru, where the mortality rate of this species is between 15% and 18%, which generates economic losses of up to 53% due to morbidity and 95% due to mortality [6], [7]. On the other hand, it is known that most mammals have homeothermy as a characteristic [8], [9]; however, in the case of guinea pigs, temperature is a factor that affects breeding, since they experience increases in body temperature when stressed [10], which is another of the most frequent causes of mortality in this species.

From another perspective, the research [11] presents a patent of a scraper type manure cleaning machine for livestock, which is connected with 2 support rods at the ends; On the other hand, this research aims to propose an automated design of a cleaning machine and an environmental temperature controller for guinea pig sheds, in order to optimize the time of farmers and obtain a better production through the breeding of these animals, also emphasis was given to the main causes of mortality of guinea pigs; finally the machine will move the excrement of the guinea pigs so that it can be stored and used as fertilizer later.

The structure of the research is as follows: in section II, the flow diagram of the various processes of the system was made, in addition to showing the percentage of bacterial species in the sheds. On the other hand, section III presents the results obtained; firstly, the results of the house cleaning machine are shown, and secondly, the results of the environmental temperature controller are visualized, and the simulations were carried out using Factory IO and LabVIEW software, respectively. Section IV presents the discussions regarding the research proposal proposed in this study and finally the conclusions in Section V.

II. MATERIALS AND METHODS

First, to establish the development of the process in Fig. 1, it was essential to start up the system which first evaluates the position of the guinea pig manure, on which depends the drive of the pistons and the conveyor belt, which in turn will

move the manure to the storage area where the robotic arm of 2GDL is located, which is programmed to deposit the manure from the conveyor belt to the storage box; in addition, as a result, an ideal habitat is obtained as shown in Fig. 1.

On the other hand, since the aim is to control the temperature, it is important to detect the humidity in the environment; if an excess of humidity is detected, the fan is activated; otherwise, if low humidity levels are detected, the heater is activated to obtain an ideal habitat.

Finally, the temperature levels are detected; if there is an excess of high temperature in the environment, the fan is activated; on the other hand, if there are low temperature levels, the heater is activated in order to obtain an ideal habitat.

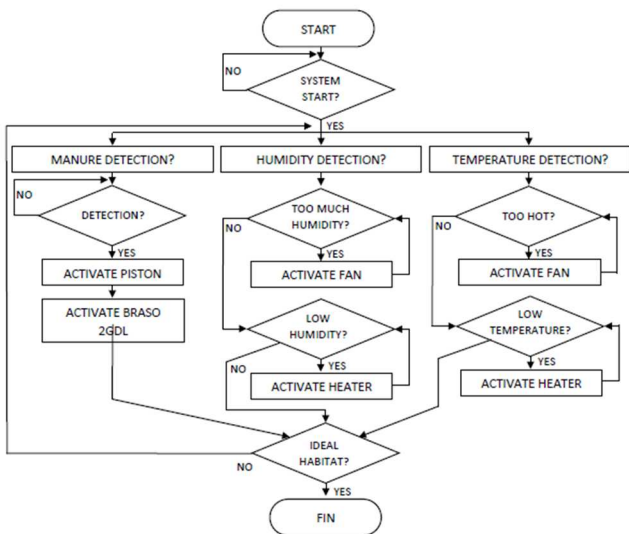


Fig.1 System flow diagram.

On the other hand, the design presented aims to reduce mortality of guinea pigs, either by the bacterial disease "salmonellosis" which most frequently affects their production in Peru, so it is mentioned that the causes of death of guinea pigs during the first week of lactation are 46.7%. Salmonellosis is transmitted through the digestive tract, that is, by ingesting contaminated water or food, which mainly affects young guinea pigs and old of them. Table 1 shows the percentages of bacterial species identified from the liver, spleen, lung and intestine of the guinea pigs that were affected by salmonellosis [3].

Table 1. Percentages of bacterial species [3].

Bacterial species	Positives		Negatives	
	N° of animals	Percentage (%)	N° of animals	Percentage (%)
<i>Escherichia coli</i>	78	40.84	113	59.16
<i>Salmonella spp.</i>	75	39.27	116	60.73
<i>Citrobacter freundii</i>	35	18.32	156	81.68
<i>Citrobacter diversus</i>	27	14.14	164	85.86
<i>Enterobacter aerogenes</i>	21	10.99	170	89.01
<i>Serratia liquefaciens</i>	21	10.99	170	89.01
<i>Proteus vulgaris</i>	14	7.33	177	92.67
<i>Citrobacter amalonaticus</i>	5	2.62	186	97.38
<i>Providencia alcalifaciens</i>	5	2.62	186	97.38
<i>Providencia stuartii</i>	5	2.62	186	97.38

<i>Serratia marcescens</i>	5	2.62	186	97.38
<i>Hafnia alvei</i>	3	1.57	188	98.43
<i>Proteus mirabilis</i>	3	1.57	188	98.43
<i>Edwardsiella tarda</i>	2	1.05	189	98.95
<i>Morganella morganii</i>	2	1.05	189	98.95
<i>Klebsiella pneumonia</i>	1	0.52	190	99.48
<i>Providencia rettgeri</i>	1	0.52	190	99.48

In another order of ideas, a temperature and humidity control were implemented in the cleaning machine, since high temperature levels are another factor that increased the mortality rate of guinea pigs [5].

Fig. 2 shows in the controllers the sensors and actuators used in the system using the Factory IO software; in addition to using the connection to the OPC server and performing the simulation in the TIA Portal software. On the other hand, various inputs and outputs were used for the system, for example: the fuzzy sensor, push buttons (start, stop and emergency), conveyor belt, electric pistons and a 2 GDL robotic arm.

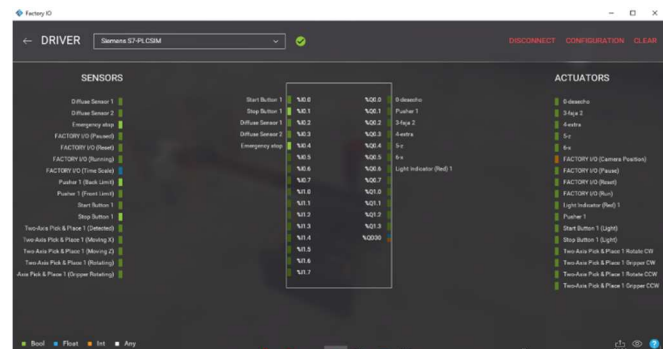


Fig. 2 System conformation controllers in Factory IO software.

In Fig. 3, the 2 GDL robotic arm was programmed to simulate the storage of guinea pig manure, a fuzzy sensor 2 was used to detect the manure and subsequently activate the robotic arm, on the other hand, different timers such as TP and TOM were added for the activation of each degree of freedom and the arm's aspirator.

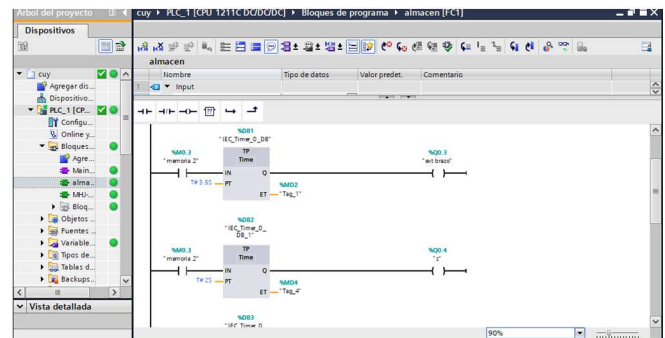


Fig.3 Programming of manure storage in the PLC.

III. RESULTS

A. Cleaning machine for guinea pig houses

Fig. 4 shows the control box, where there is a green button (on), a red button (off), a mushroom-shaped button (emergency stop) and a red pilot light, which lights up when the system is started.

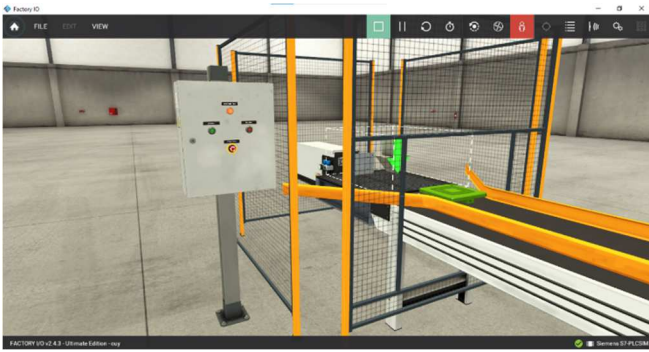


Fig. 4. System control box.

Fig. 5 shows the fuzzy sensor 1, which detects the manure (green boxes) to then activate the electric piston to then send the manure to the belt where it will go to storage, the objective of the research is to obtain a constant cleaning as it is a very important factor for the guinea pigs to feed healthily and not to be in contact with their own excrement.

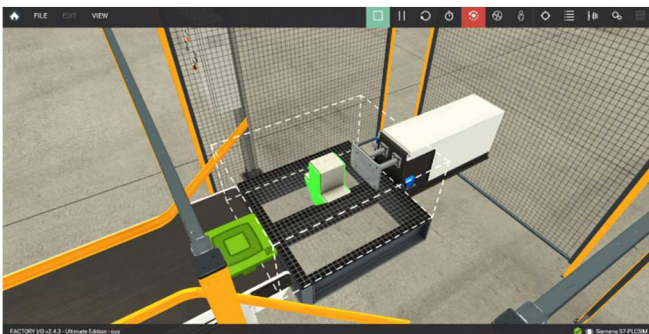


Fig. 5. Representation of manure output in the Factory IO environment.

In Fig. 6 we can see on the fuzzy sensor 2 where it activates the robotic arm to send the guinea pig manure to the warehouse, the purpose of this is to obtain organic fertilizer in x amount of time.

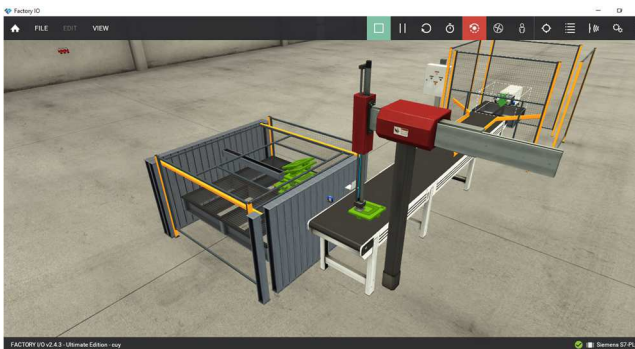


Fig. 6. Robotic arm interaction.

Fig. 7 shows that after some time the warehouse is gradually filling up, which indicates that the project is constantly being worked on in order to observe the potential of what is proposed.

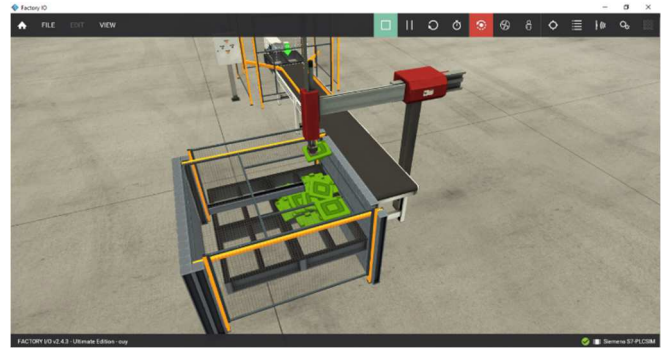


Fig. 7. Manure storage environment.

Fig. 8 shows the execution of the ladder programming of the system, where the optimal operation of the start-up control and the emergency control is shown.

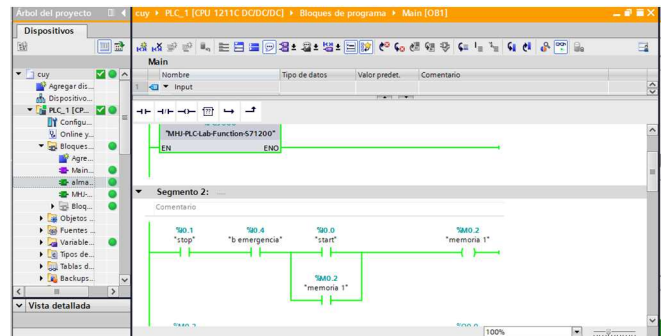


Fig. 8. Compilation of PLC programming in TIA Portal software.

B. Environmental temperature controller for guinea pig houses.

LabVIEW software was used to represent the automation process to control the temperature and humidity in the guinea pig houses; in addition, with the help of sensors and respective actuators, we have the appropriate measurement to control these two factors and thus have an ideal habitat for these animals. Additionally, being an automated process, each factor is controlled independently to avoid conflicts according to the environmental measurements taken by the sensors, in which it was determined that at higher values of the ideal standard the fans will be turned on and at lower values the heaters will be turned on, as shown in Fig. 9.

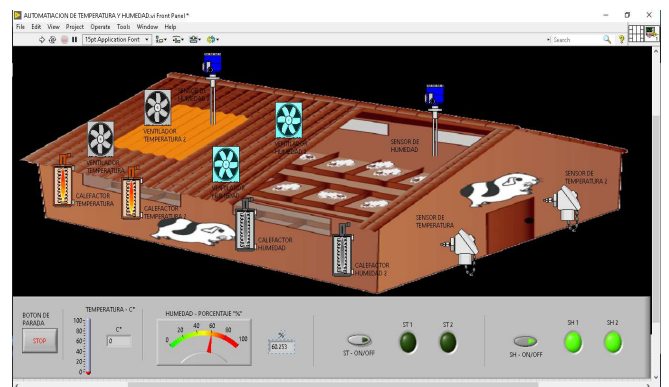


Fig. 9. Graphical representation of the automation of the temperature and environment of the sheds.

In order to have an adequate programming, a temperature range of 23C° to 26C° was used in the case of humidity in the environment, for which a range of 40% to 50% was used in order to have an ideal habitat in the sheds [5]. On the other hand, Fig. 10 shows the start of the process, where the actuators will respond once the sensors are sensed or are

turned on to prevent errors in the process, meanwhile the value of 0C°; in addition, random values were set to visualize the reaction of the actuators and indicators in the control table and thus activate or deactivate the actuators and indicators in the control table.

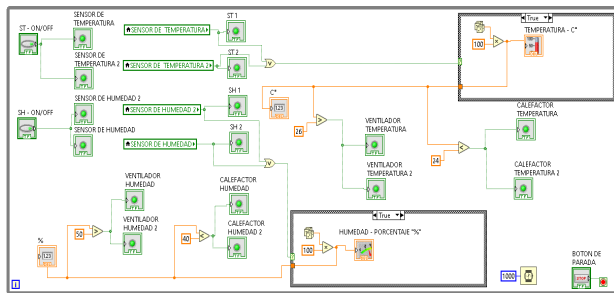


Fig. 10. Compilation of the programming in the LabVIEW software.

IV. DISCUSSIONS

Over the years some machine designs were made to have a space cleaning control in order to automate and facilitate such process, in which one of them focuses on sheep cleaning and is composed of upper and lower blades and with a rail provided at the top that allows moving the blades from one side to another to perform effective work, whereby it is also composed of gears, bearings, gear wheel and rack and pinion coupling [12]. Other designs include a water tank, manure scraping device and a controller, the bottom of the water tank is provided with a driving wheel is electrically connected with the controller, the bottom section is grooved on the front side the water tank; in addition, two vertical motorized rails are installed, these are electrically connected with the controller and the manure scraping device is provided with the front of the motorized rails in the groove; therefore, this model is more related to the field of chicken breeding equipment thus helping in farms [13].

Among all these designs a machine that covers more generally manure cleaning for livestock is characterized by having a shovel at the front of the self-propelled equipment for cleaning excrement behind a support beater that are installed as a helical blade to the groove of the shovel this is connected to the main shaft to the power system with a driving mechanism, therefore, this design is intended to facilitate this tedious, dirty and painful activity [14], other cleaning machines focus on mixing the supplies in the receiving device and transporting them along a filter to a conveyor unit; whereby, this is separated from the manure by shaking / sliding the filter element, so that the manure is shaken, pressed and slid through the orifices, dropping into a sand collecting container or onto a stable outlet and/or base, a separate claim is also included for a mobile cleaning device, with the object of separating litter from manure in horse stables [15].

Sends along the plate from one direction of plate movement and barbed wires. The device allows the manure removal to be gentle and plays a good protective role for livestock, in addition, it is simple and easy to manufacture, with low production cost [8]; however, all these machines focus only on cleaning for which is a very important but not the only feature to have an ideal environment for animals. For this reason, the present research presented an automated design of a cleaning machine for guinea pig houses, as well as a control system for air conditioning the environment through a range of ideal temperature and humidity

determined in the guinea pig houses to optimize production and reduce mortality rates.

V. CONCLUSIONS

From the research carried out, it is concluded that the present work will allow an optimal cleaning of the sheds that will prevent the guinea pigs from contracting diseases such as salmonellosis, which affected 46%; the latter, in turn, affects the owners of these animals due to the economic loss of 53% due to morbidity and 95% due to mortality. In addition, with this project the temperature and humidity of the environment will also be controlled, since cases of guinea pig mortality due to excessive amounts of these factors were found. In this way, the temperature and humidity were regulated by sensors allowing a range of 24C° to 26C° and 40% to 50% respectively.

For a detection of lower temperature and humidity, the heaters will be activated to control both factors and in case of excess, the turbines will be activated to cool the environment and lower the humidity in the sheds; on the other hand, in the case of cleaning, once the guinea pig manure is detected, the pistons will be activated and in turn will work in parallel with the activation of a 2GDL arm, as shown in Fig. 1. Finally, to complete the process, all the factors mentioned will be evaluated and the process will be completed until returning to the beginning of the system. With this project, it is possible to have an ideal environment for the guinea pigs, at the same time reducing economic expenses in the loss of production and/or bad control of temperature and humidity and even in the worst of the scenarios to lose the breeding of these animals for not having the respective care.

APPENDIXES

<https://drive.google.com/file/d/12SOkeBewymZkQEeUw5BGsRcbFkng3tK/view?usp=sharing>

REFERENCES

- [1] V. G. Santos, "Importancia del cuy y su competitividad en el mercado | Semantic Scholar," 31 December, 2007. <https://www.semanticscholar.org/paper/Importancia-del-cuy-y-su-competitividad-en-el-Santos/d6ddcc6031494478a72ab53b3045001ae1ab2d2e> (accessed Apr. 30, 2022).
- [2] M. Chacra Emprendedora -Haku Wiñay, "MANUAL TÉCNICO Crianza de cuyes."
- [3] C. Chuquizuta R. and S. I | Morales C., "Identificación de agentes bacterianos aislados de gazapos muertos de cuyes en una granja de crianza intensiva en Lima, Perú." <https://www.redalyc.org/articulo.oa?id=63654640041> (accessed Apr. 30, 2022).
- [4] I. Lilia and C. De Zaldívar, *Producción de cuyes (Cavia porcellus)*, Vol. 138. Food & Agriculture Org., 1997.
- [5] C. Pang, F. An, S. Yang, N. Yu, D. Chen, and L. Chen, "In vivo and in vitro observation of nasal ciliary motion in a guinea pig model," *Exp. Biol. Med. (Maywood)*, vol. 245, no. 12, pp. 1039–1048, Jun. 2020, doi: 10.1177/1535370220926443.
- [6] O. Gabriela Ortega, A. Ronald Jiménez, G. Miguel Ara, and C. Siever Morales, "La Salmonellosis como factor de riesgo de mortalidad en cuyes," *Rev. Investig. Vet. del Perú*, vol. 26, no. 4, pp. 676–681, 2015, doi: 10.15381/RIVEP.V26I4.11203.
- [7] D. D. C. Miguel Ara G., Ronald Jiménez A., Amparo Huamán C., Fernando Carcelén C., "Desarrollo de un índice de condición corporal en cuyes: relaciones entre condición corporal y estimados cuantitativos de grasa corporal," *Rev. investig. vet. Perú* v.23 n.4 Lima dic., 2012. http://www.scielo.org.pe/scielo.php?pid=S1609-91172012000400004&script=sci_arttext (accessed May 01, 2022).
- [8] I. H. Levy, N. Di Girolamo, and K. A. Keller, "Rectal temperature is a prognostic indicator in client-owned guinea pigs," *J. Small Anim.*

- Pract.*, vol. 62, no. 10, pp. 861–865, Oct. 2021, doi: 10.1111/JSAP.13388.
- [9] J. R. Spotila and D. M. Gates, “Body Size, Insulation, and Optimum Body Temperatures of Homeotherms,” pp. 291–301, 1975, doi: 10.1007/978-3-642-87810-7_17.
- [10] A. E. Snow and A. Horita, “Interaction of Apomorphine and Stressors in the Production of Hyperthermia in the Rabbit1,” vol. 22, no. 2.
- [11] W. Shunxi, L. Shiguan and X. che, “Cattle-expelling device for scraper-type manure cleaning machine”, CN102550427B China, January 12, 2012.
- [12] W. Tan, “Automate the sheep hurdle of cleaning”, CN105409784B China, November 29, 2015.
- [13] L. Huiping, L. Jiajun, G. Yuan, X. Jian and C. Cheng, “A kind of chicken farm cleaning equipment”, CN207100118U China, August 25, 2017.
- [14] Z. Ruihua, “Cattle ranch dung cleaner”, CN201088056Y China, October 7, 2007.
- [15] H. Ullstein, “Reinigungsgerät für Pferdeställe”, EP2243354A1 Alemania, October 27, 2010.

Machine Learning Performances for Covid-19 Images Classification based Histogram of Oriented Gradients Features

1st Yessi Jusman*

Department of Electrical Engineering,
Faculty of Engineering
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia
*Corresponding
Email:yjusman@umy.ac.id

2nd Wikan Tyassari

Department of Electrical Engineering,
Faculty of Engineering,
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia

3rd Difa Nisrina

Department of Electrical Engineering,
Faculty of Engineering,
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia

4th Fahrul Galih Santosa

Department of Electrical Engineering,
Faculty of Engineering,
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia

5th Nugroho Abdi Prayitno

Department of Electrical Engineering,
Faculty of Engineering,
Universitas Muhammadiyah
Yogyakarta
Yogyakarta, Indonesia

Abstract— Coronavirus disease (Covid-19) is an infectious disease that attacks the respiratory area caused by the severe acute respiratory syndrome (SARS-CoV-2) virus. According to the World Health Organization (WHO) as of April 2022, there were more than 500 million cases of Covid-19, and 6 million of them died. One of the tools to detect Covid-19 disease is using X-ray images. Digital X-ray images implementation can be developed classification method using machine learning. By using machine learning, the diagnosis of this disease can be faster. This study applied a features extraction method using the Histogram of Oriented Gradients (HOG) algorithm and the Linear Support Vector Machine (SVM), K-Nearest Neighbor (KNN) Medium and Decision Tree (DT) Coarse Tree classification methods. The study can be used in the diagnosis of Covid-19 disease. The best method among the classification methods is features extraction from HOG algorithm and DT Coarse Tree. The highest values of accuracy, precision, recall, specificity, and F-score were 83.67%, 96.30%, 78.79%, 98.25, and 76.48%.

Keywords—Covid-19, Histogram of Oriented Gradients, Support Vector Machine, K-Nearest Neighbor, Decision Tree.

I. INTRODUCTION

In the last three years, the world has been hit by a global pandemic that has affected the entire world community. Coronavirus disease (Covid-19) is an infectious disease that attacks the respiratory area caused by the Severe Acute Respiratory Syndrome (SARS-CoV-2) virus [1]. According to data from the World Health Organization (WHO) as of April 2022, there were more than 500 million cases due to Covid-19, and 6 million of them died [2].

This Covid-19 disease is quite dangerous apart from being deadly because the transmission of Covid-19 is very fast. With its rapid spread and very easy transmission, fast

and accurate detection is needed to prevent its spread. There are several methods of detecting Covid-19 disease, one of which is detection using x-ray images [3]. Although diagnosis using Polymerase Chain Reaction (PCR) is more widely used, detection with x-ray images is cheaper because it can detect a large number of patients and is not limited by test equipment [4]. However, the diagnosis of Covid-19 disease using x-ray images can only be carried out by the relevant expert or doctor. So it takes quite a long time and is limited to the number of experts or doctors available [5] [6].

To speed up the diagnosis of Covid-19 disease, it is carried out on a computer-based basis. The development of computer-based Covid-19 disease detection is one of the studies that has been widely developed. In recent years, studies [7] [8] have used computers to help in the early diagnosis of Covid-19.

Machine Learning (ML) is well known for its high-performance image processing and its application in certain image classifications. Thus, ML is an alternative for automatic detection of Covid-19 disease using x-ray images. The use of ML in the diagnosis of Covid-19 disease has been carried out by several studies [9] [10].

Image processing is an important aspect in the ML process. For an image to be analyzed using the ML method, the image must be extracted (feature). One feature extraction method that is quite often used is the Histogram of Oriented Gradients (HOG). HOG feature extraction focuses on the shape and angle of the feature gradient. Several related studies that use HOG as a feature extraction method in detecting Covid-19 images are. [5][11][12] yielded the highest accuracy values of 98.66%, 98.11%, and 98.5%.

To determine the image of Covid-19 and not the image of Covid-19, a classification method is used. Several image classification methods include Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree (DT). Several studies related to the diagnosis of Covid-19 disease using the SVM, KNN, and DT classification methods include [13] [12]. Several studies use other methods to be applied in the detection of Covid-19 disease [14].

With only a few studies on Covid-19 images combined with the HOG feature extraction method with the SVM, KNN, and DT classification methods. Therefore, in this study the HOG method, with the classification of SVM, KNN, DT is used in the diagnosis of Covid-19 disease with x-ray images.

II. METHOD

A. Data Collection

The image data used in this study is Covid-19 images obtained from open datasets and Machine Learning Projects Kaggle website with the type of lung X-ray images. The number of Covid-19 images used is 1184 images, which are divided into three types of images, namely Covid-19 images (class 1), Pneumonia images (class 2), and Normal lung images (class 3). The number of images for each class is 404 images for class 1, 390 images for class 2 and for 390 images for class 3. The data is divided into 2 types of data sets, namely training and testing. The amount of training data is 90% of the total image data, which is 1065 images, and for data set testing is 10% of the total image data, which is 119 images.

B. System Design

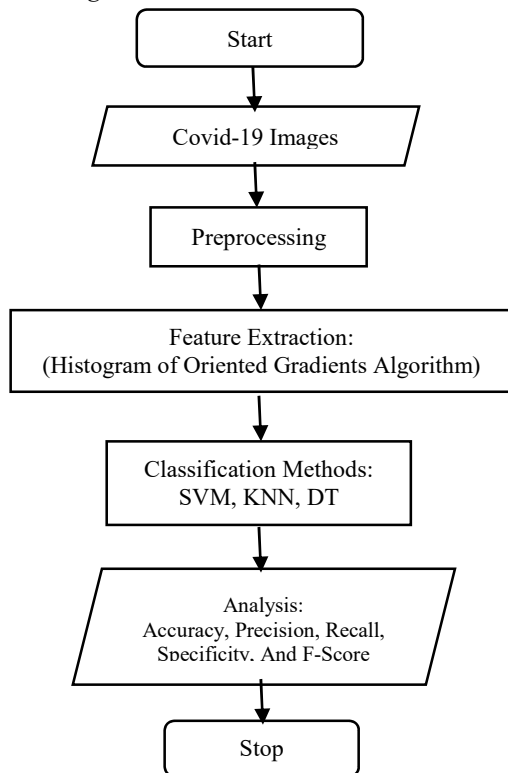


Fig. 1. System Design Flowchart

In this study, the analysis was carried out using the MATLAB software version R2020a. The computer specifications used during the Covid-19 image processing process are shown in Table 1.

TABLE I. COMPUTER SPESIFICATIONS

Hardware	Specification
Processor	Intel® Core i5 9400f
RAM Memory	16GB
GPU	Nvidia RTX 2060 6GB

This COVID-19 image classification goes through several steps, including pre-processing, feature extraction, classification, and analysis. The feature extraction used for this study is the Histogram of Oriented Gradients (HOG) algorithm. The classification method used is Linear Support Vector Machine (SVM), K-Nearest Neighbor (KNN) Medium and Decision Tree (DT) Coarse Tree. The whole system design is shown in Figure 1.

C. Pre-pocessing

This Covid-19 image is divided into a training dataset and a testing data set. The next process is preprocessing for the two datasets. The original image with the 2 datasets was flipped vertically as an effort to reproduce the image. To be able to perform feature extraction using the HOG algorithm of the image must be changed from an RGB image to a Grayscale. The image that has been converted to grayscale is converted to histogram to equalize the contrast value of the image.

D. Feature Extraction

The testing and training datasets are processed to obtain feature extraction. This feature extraction stage is carried out after the image has been pre-processed. In this study, the feature extraction used is the HOG algorithm, with the extraction results totaling 36 features.

E. Classification

At this stage, classification is carried out with the extracted data using HOG algorithm. Classification is carried out using 3 different methods, the classification methods used include Linear Support Vector Machine (SVM), K-Nearest Neighbor (KNN) Medium and Decision Tree (DT) Coarse Tree. These three methods were tested ten times running with 10% data validation of the training data. The results of the classification are analyzed based on the training computation time and training accuracy. These results are displayed on the Receiver Operating Characteristic (ROC) graph.

F. Analysis

The results of the Covid-19 image classification were analyzed using a performance matrix. The image dataset used in this analysis is a testing dataset. The performance matrix analysis used includes accuracy, precision, recall, specificity, and f-score. The results of the performance matrix analysis are compared to get the best method.

III. RESULTS AND DISCUSSIONS

A. Feature Extraction Result

The testing and training dataset images are feature extraction using HOG algorithm and produce a total of 36 features for each image, with different extraction values. The results of feature extraction with HOG algorithm are shown in Table 2. Table 2 shows the average extraction results of 10 images in each class.

TABLE II. FEATURE EXTRACTION RESULTS

Images	Histogram of Oriented Gradients		
	Class 1	Class 2	Class 3
1	0.142	0.144	0.149
2	0.149	0.145	0.149
3	0.144	0.146	0.139
4	0.149	0.136	0.154
5	0.143	0.125	0.143
6	0.150	0.135	0.137
7	0.144	0.127	0.141
8	0.147	0.133	0.145
9	0.145	0.133	0.142
10	0.151	0.139	0.146
Mean	0.147 ± 0.003	0.136 ± 0.007	0.144 ± 0.005

B. Classification Result

The extracted data obtained were used for classification using the Linear SVM, KNN Medium, and DT Coarse Tree methods. Classification was carried out 10 times running for each method. From the classification with these methods, the value of training accuracy and training time is obtained. The results of the classification, namely training accuracy and computational time are shown in Table 3 and the results of the Receiver Operating Characteristic (ROC) graph. The best ROC graph is shown in Figure 2.

TABLE III. CLASSIFICATION RESULTS

Run	SVM Linear		KNN Medium		DT Coarse Tree	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time(s)
1	69.40%	1.21	69.80%	0.79	63.70%	1.28
2	69.30%	1.23	69.50%	0.81	62.50%	0.80
3	69.30%	1.22	70.50%	0.76	63.20%	0.80
4	68.80%	1.23	69.90%	0.82	63.20%	0.83
5	68.50%	1.21	70.00%	0.77	62.70%	0.82
6	68.20%	1.22	70.00%	0.75	62.60%	0.82
7	68.70%	1.24	69.30%	0.76	63.00%	0.82
8	68.20%	1.28	70.00%	0.72	62.90%	0.86
9	68.80%	1.33	70.30%	0.81	63.60%	0.85
10	69.10%	1.30	69.80%	0.79	62.40%	0.84
Mean	68.83% ± 0.42%	1.25 ± 0.04	69.91% ± 0.33%	0.78 ± 0.03	63.05% ± 0.47%	0.88 ± 0.13

Based on the results of image classification in Table 3, the best training accuracy was obtained by KNN Medium with the highest accuracy value of 70.50% and the fastest time of 0.72 second. The average value of training accuracy obtained by KNN Medium is 69.91% ± 0.33% and the average computation time is 0.78 ± 0.03. In contrast to KNN Medium, the DT Coarse Tree method gets the lowest training accuracy value, which is 62.40% with a time of 0.84 seconds. The average training accuracy of the DT Coarse Tree method is 63.05% ± 0.47% and time 0.88 ± 0.13.

C. Analysis Result

At this stage, a testing dataset that has been extracted has been used as input. Testing data is tested using the best training classification results from each method. The analysis was carried out using a performance matrix comparison for each method to find out the best method for classifying Covid-19 images. The results of the performance matrix for testing are shown in Table 4. The results of the best confusion matrix of the three methods are shown in Figure 3, the best confusion matrix for SVM Linear is 1st running, KNN Medium is 3rd running, and DT Coarse Tree is 5th running.

Based on the data shown in Table 4, the best performance matrix in the classification of Covid-19 images is the DT Coarse Tree method. The accuracy values obtained by the DT Coarse Tree method are 78.10% for class 1, 83.67% for class 2, and 70.09% for class 3. The precision values are 76.48% for class 1, 96.30% for class 2, and 48.15% for class 3. The recall obtained were 63.41% for class 1, 63.41% for class 2, and 78.79% for class 3. The specificity results obtained were 87.54% for class 1, 98.25% for class 2, 66.67% for class 4. Results F- the score obtained is 69.33% for class 1, 76.48% for class 2, and class 3 is 59.78%. The results of the comparison of these three methods can be seen in the graph shown in Figure 4. From the graph, the performance of DT Coarse Tree tends to be higher than the other methods.

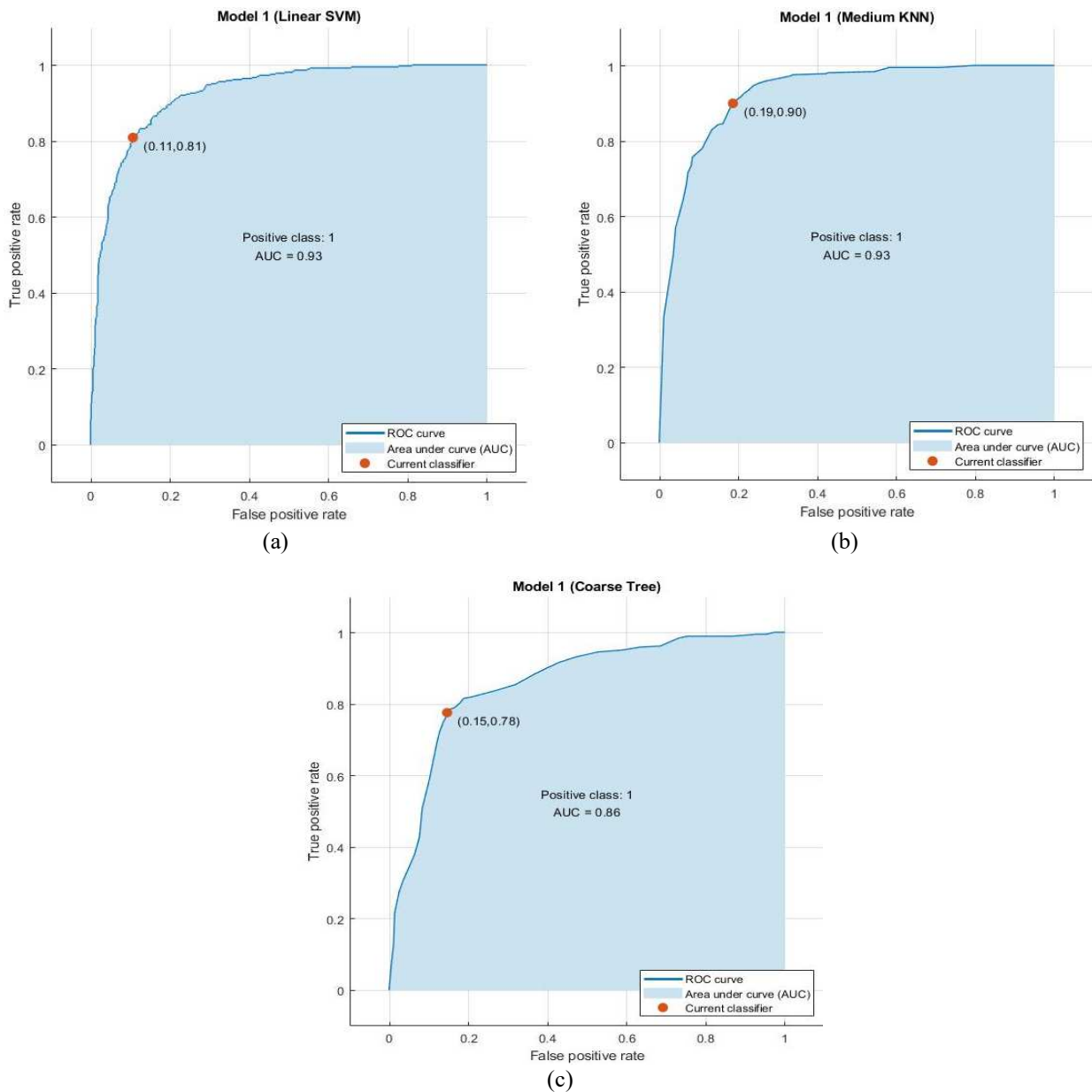


Fig. 2. Best of Receiver Operating Characteristic Graph (a) SVM, (b) KNN, (c) DT

TABLE IV. ANALYSIS PERFORMANCE MATRIX RESULT

Results	SVM Linear			KNN Medium			DT Coarse Tree		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Accuracy	70.37%	83.52%	66.09%	70.37%	81.72%	67.26%	78.10%	83.67%	70.09%
Precision	59.57%	96.55%	54.90%	58.18%	76.19%	76.19%	76.48%	96.30%	48.15%
Recall	68.30%	66.67%	63.64%	78.05%	82.05%	54.24%	63.41%	63.41%	78.79%
Specificity	71.64%	97.96%	67.61%	65.67%	81.48%	81.48%	87.54%	98.25%	66.67%
F-score	63.64%	78.87%	58.95%	66.67%	79.01%	63.37%	69.33%	76.48%	59.78%

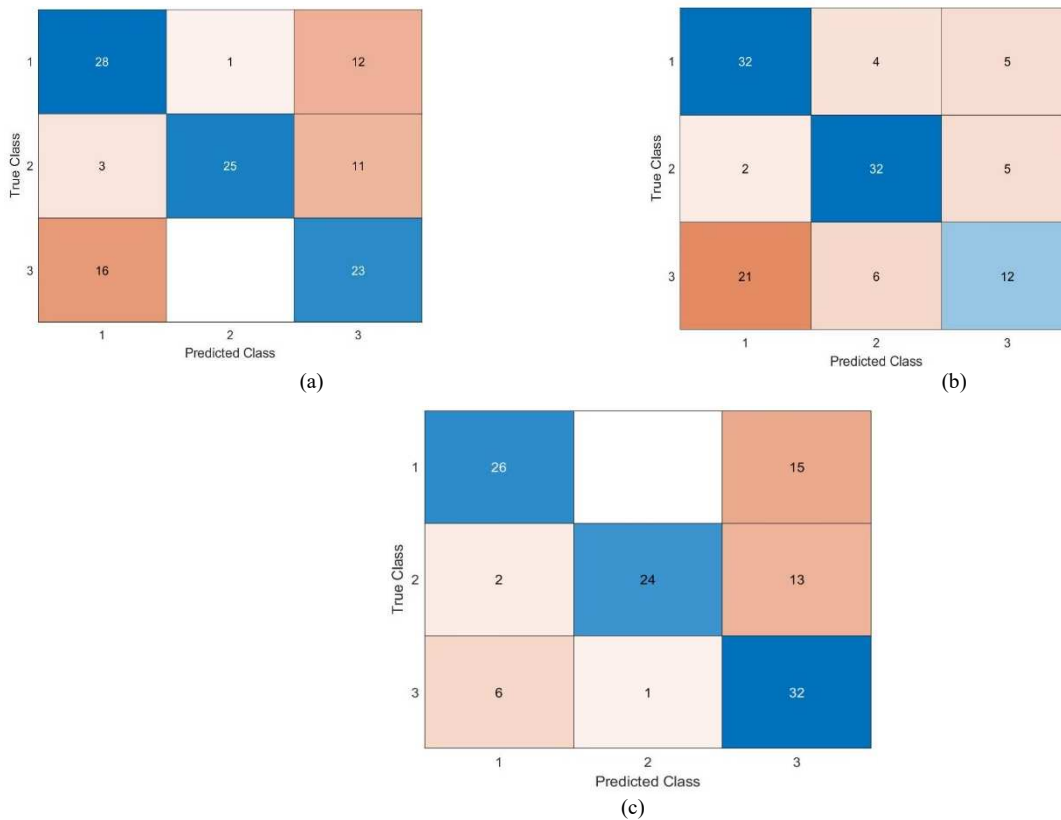


Fig. 3. Best Example of Confusion Matrix Testing (a) SVM, (b) KNN, (c) DT

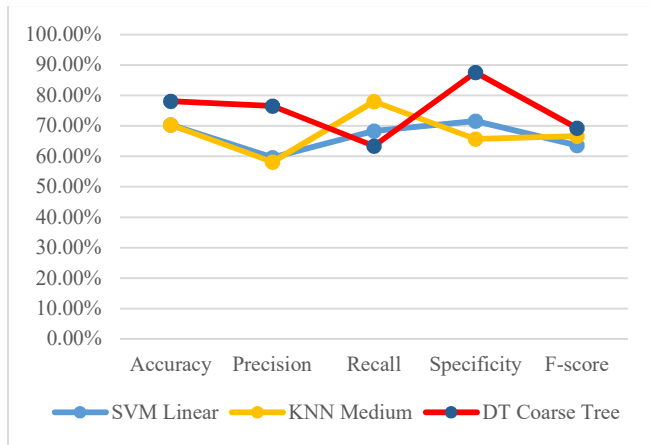


Fig. 4. Comparison Graph of Classification Models

IV. CONCLUSION

Covid-19 image classification using Linear Support Vector Machine (SVM), K-Nearest Neighbor (KNN) Medium and Decision Tree (DT) Coarse Tree with feature extraction by using Histogram of Oriented Gradients (HOG) algorithm can be developed to help to classify the Covid-19 disease based on X-ray images. It can be concluded that the DT Coarse Tree method is the method that produces the best performance in the classification of Covid-19 images. The results of the DT Coarse Tree produce accuracy values for class 1, class 2, and class 3 of 78.10%, 83.67%, 70.09%. The

precision values for class 1, class 2, and class 3 are 76.48%, 96.30%, and 48.15%. The recall scores for class 1, class 2, and class 3 were 63.41%, 63.41%, and 78.79%. The specificity results for class 1, class 2, class3 were 87.54%, 98.25, and 66.67%. The F-score results for class 1, class 2, and class 3 were 69.33%, 76.48%, and 59.78%. Based on these results, it is necessary to test using feature extraction and other classification methods as a comparison reference for the automatic detection method for Covid-19 disease.

ACKNOWLEDGMENT

This research is supported by Universitas Muhammadiyah Yogyakarta and a research project grant from the Ministry of Research and Technology of the Republic of Indonesia.

REFERENCES

- [1] R. Rehouma, M. Buchert, and Y. P. Chen, "Machine learning for medical imaging-based COVID-19 detection and diagnosis," *Int. J. Intell. Syst.*, p. 10.1002/int.22504, May 2021, doi: 10.1002/int.22504.
- [2] S. Kadry, V. Rajinikanth, S. Rho, N. Sri, and M. Raja, "Development of a Machine-Learning System to Classify Scan Images into Normal / COVID-19 Class," vol. 2, no. December, 2019.
- [3] L. Brunese, F. Martinelli, F. Mercaldo, and A. Santone, "Machine

- learning for coronavirus covid-19 detection from chest x-rays,” *Procedia Comput. Sci.*, vol. 176, pp. 2212–2221, 2020, doi: <https://doi.org/10.1016/j.procs.2020.09.258>.
- [4] M. Chiericato *et al.*, “A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data,” *Sci. Rep.*, vol. 12, no. 1, p. 4329, 2022, doi: 10.1038/s41598-022-07890-1.
- [5] A. Saygılı, “A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods,” *Appl. Soft Comput.*, vol. 105, p. 107323, 2021, doi: 10.1016/j.asoc.2021.107323.
- [6] M. R. Islam and M. Nahiduzzaman, “Complex features extraction with deep learning model for the detection of COVID19 from CT scan images using ensemble based machine learning approach,” *Expert Syst. Appl.*, vol. 195, no. February, p. 116554, 2022, doi: 10.1016/j.eswa.2022.116554.
- [7] D. Al-Karawi, S. Al-Zaidi, N. Polus, and S. Jassim, “Machine Learning Analysis of Chest CT Scan Images as a Complementary Digital Test of Coronavirus (COVID-19) Patients,” *medRxiv*, p. 2020.04.13.20063479, 2020, [Online]. Available: <https://www.medrxiv.org/content/10.1101/2020.04.13.20063479v1%0Ahttps://www.medrxiv.org/content/10.1101/2020.04.13.20063479v1.abstract>.
- [8] A. Zargari Khuzani, M. Heidari, and S. A. Shariati, “COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images,” *Sci. Rep.*, vol. 11, no. 1, p. 9887, 2021, doi: 10.1038/s41598-021-88807-2.
- [9] P. Saha, M. S. Sadi, and M. M. Islam, “EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers,” *Informatics Med. Unlocked*, vol. 22, p. 100505, 2021, doi: 10.1016/j.imu.2020.100505.
- [10] S. Samsir, J. H. P. Sitorus, Zulkifli, Z. Ritonga, F. A. Nasution, and R. Watrionthos, “Comparison of machine learning algorithms for chest X-ray image COVID-19 classification,” *J. Phys. Conf. Ser.*, vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012040.
- [11] A. M. Ayalew, A. O. Salau, B. T. Abeje, and B. Enyew, “Detection and classification of COVID-19 disease from X-ray images using convolutional neural networks and histogram of oriented gradients,” *Biomed. Signal Process. Control*, vol. 74, no. October 2021, p. 103530, 2022, doi: 10.1016/j.bspc.2022.103530.
- [12] J. N. Hasoon *et al.*, “COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images,” *Results Phys.*, vol. 31, no. December 2020, pp. 0–7, 2021, doi: 10.1016/j.rinp.2021.105045.
- [13] A. Mohammadi *et al.*, “Diagnosis / Prognosis of COVID-19 Chest Images via Machine Learning and Hypersignal Processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 37–66, 2021.
- [14] H. Yasar and M. Ceylan, “A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods,” *Multimed. Tools Appl.*, vol. 80, no. 4, pp. 5423–5447, 2021, doi: 10.1007/s11042-020-09894-3.

Multilayer Aperture Coupled Single Band Second Order Bandpass Patch Resonator

Mohan K N
 School of Electronics Engineering
 Vellore Institute of Technology
 Vellore, India
 narasimmamohan@gmail.com

Yogesh Kumar Choukiker
 School of Electronics Engineering
 Vellore Institute of Technology
 Vellore, India
 yogesh.ku.84@gmail.com

Abstract—Multilayer aperture coupled patch resonator antenna (M-ACRA) introduced to achieve single band second order frequency response. L shaped truncated patch at 12GHz designed to achieve circular polarization and minimize cross polarization at the receiver section. A theoretical approach for demonstrating floquet excitation are explained, furthermore, current distribution of proposed design are demonstrated. Based on analysis mode, the operation principle for equivalent circuit model about proposed designed also discussed. The simulated results and antenna parameters are to be consider for aperture coupled based frequency selective surfaces, reconfigurable frequency selective surfaces, reconfigurable intelligent surface, etc.

Index Terms—Aperture-coupled patch resonator, single-mode, dual order frequency

I. INTRODUCTION

Multilayer Aperture Coupled Resonator Antenna (M-ACRA) electrically compact size, low profile and low noise RF antenna. Aperture coupled based Frequency Selective surfaces (A-FSS) deals by various authors [1] [2]–[6], this type of aperture coupled resonator are applied to achieve higher order bandpass responses [3], [7], [8]. M-ACRAs maintain beam polarization stable, used to absorb transmitted EM waves in metamaterials based FSS structures, to minimize cross polarization and improve co-polarization towards receiver. A multilayer aperture coupled patch resonator introduced in 2D planar pattern on dielectric substrate, the approached design electrically small size when compared to aperture coupled dielectric resonator antennas. This type of design proposes single and dual bandpass responses in-terms of periodic structure implementation in [3], [9]. Various authors in the last decade used this types of structure to achieve higher order bandpass responses and transmission zero characteristics [2], [4], [8].

Aperture coupled mounted between two substrate acts like coupled resonator theory in filters [3]. Transmitted wave from the free space will be absorbed by aperture coupled layer and the waves are reflected back by ground patch. The ground patch similar geometry structure of top patch which is shown in Fig. 1. Given antenna excites by floquet excitation techniques which is discussed in [10], [11]. The primary sweep of reflection coefficient and insertion loss are shown in Fig. 2. Aperture coupled as non-resonant and it have

strong magneticfield in it. The field distribution towards top and bottom patches to achieve higher order bands.

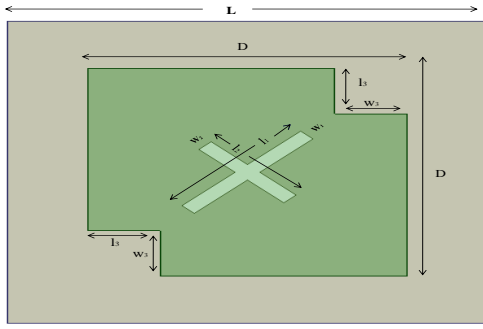
In this article proposed L shaped truncated multilayer aperture coupled patch resonator to achieve second order transmission zero filter response, where two transmission zeros are produced near to the passband and enhancing the frequency selectivity. So, this model can be consider for unit cell element model for Aperture coupled frequency selective surfaces. The resonant frequency at 12GHz, the bandwidth maintained by 2.2GHz from 12.01GHz to 14.8GHz. The L shaped truncated patch helped to achieve circular polarization towards receiver and all the antenna parameters like, axial ratio for circular polarization, returnloss for showing transmission zero characteristics, maximum gain towards desired direction are discussed. This type of aperture coupled resonator antennas can be useful in high frequency wireless applications like IOT devices, metamaterials based frequency selective surfaces, metasurfaces for Reconfigurable intelligent system, etc.

II. MULTILAYER APERTURE COUPLED RESONATOR

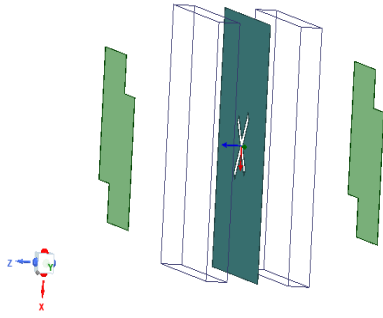
Multilayer aperture coupled resonator antenna shown in Fig. 1, the fundamental square patch for 12GHZ designed. L shaped truncated edges in both side to achieve circular polarization towards receiver and transmitter, the design equation for square patch referred in [12], the resonance mode equations are referred from same [12]. Rogers 4003 substrate with $\epsilon_r = 3.55$, $h = 0.813mm$ and $f_o = 12GHz$ to 14GHz used. As discussed, multilayer aperture coupled approached in this article, so h_1 and h_2 multi substrate with same properties are placed in between top and bottom of patches. The geometrical values are mentioned in Fig.1 and its corresponding returnloss shown in Fig.2. The excited \vec{E} and \vec{H} field intensities between top and bottom layer, which is leading to the cavity modes [12]. The reflection coefficient obtained from L shaped truncated patch from the bottom.

To understand the operation of mechanism, the incident Efield \vec{E}_x and \vec{E}_y along diagonal line from Fig. 1. The reflected \vec{E}_x and \vec{E}_y can be written as

$$\vec{E}_x^r = \frac{1}{2}(\vec{e}_y - \vec{e}_x) | \vec{E}_y^i | (S_{11})_x \quad (1)$$



(a)



(b)

Fig. 1: (a) Dimension of approached aperture coupled patch resonator, (b) multilayer substrate $h_1 = h_2 = 0.813mm$ with $\epsilon_r = 3.55$ and thickness of patch and aperture $t = 0.01mm$

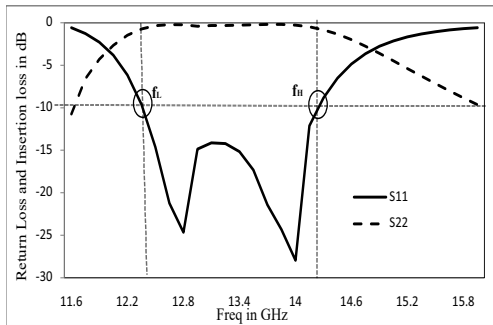


Fig. 2: Reflection coefficient S11 and Insertion loss s22 in dB for $L = 8mm, D = 5.3mm, l_1 = 2.8mm, w_1 = w_2 = 0.3mm, l_2 = 2mm, l_3 = 1.3mm, w_3 = 1.3mm$

$$\vec{E}_{\bar{y}}^r = \frac{1}{2}(\vec{e}_{\bar{y}} + \vec{e}_{\bar{x}}) | \vec{E}_{\bar{y}}^i | (S_{11})_{\bar{y}} \quad (2)$$

Here $(S_{11})_{\bar{x}} = | S_{11} |_{\bar{x}} e^{j(\beta_{11})_{\bar{x}}}$ and $(S_{11})_{\bar{y}} = | S_{11} |_{\bar{y}} e^{j(\beta_{11})_{\bar{y}}}$ are reflected coefficients of whole resonator in terms of $\vec{E}_{\bar{x}}^i$ and $\vec{E}_{\bar{y}}^i$ incident polarization respectively. So the total E-field vector \vec{E}^r can be written as,

$$\vec{E}^r = \vec{E}_{\bar{x}}^r + E_{\bar{y}}^r \quad (3)$$

Referred from 1 and 2 we can obtain $(S_{11})_y$ and $(S_{11})_x$. can be written as

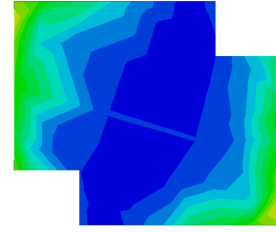


Fig. 3: Current distribution at 12GHz

$$(S_{11})_y = \frac{1}{2}((S_{11})_{\bar{y}} + (S_{11})_{\bar{x}}) \quad (4)$$

$$(S_{11})_x = \frac{1}{2}((S_{11})_{\bar{y}} - (S_{11})_{\bar{x}}) \quad (5)$$

Eq.4 and Eq. 5 are cross polarized $((S_{11})_y)$ and co polarized $((S_{11})_x)$ respectively.

III. SINGLE BAND SECOND ORDER RESONANCE

The equivalent circuit model shown in Fig.4 the designed equations are follows,

$$f_o = \frac{1}{2\pi\sqrt{LC}} \quad (6)$$

Here, $L = L + 2L_m$ and $C = C + 2C_m$ corresponding to [7], so the Eq. 6 becomes,

$$f_o = \frac{1}{2\pi\sqrt{(L + 2L_m)(C + 2C_m)}} \quad (7)$$

In Fig.4 introduced inductances and capacitances to achieve single band second order resonance frequency with respect to Fig.1 the corresponding S11 and S21 shown in Fig. 5. The Fig.5. shows the dual transmission zero in single mode frequency response. The inband maximum and minimum insertion loss also shown in the same figure.

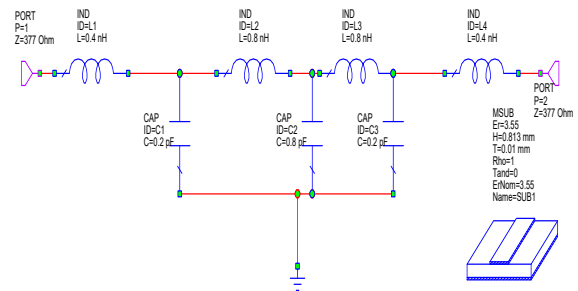


Fig. 4: Lumped circuit model for single band dual order, $L = 0.2pH, C = 0.205pF$ at free space impedance $Z_{in} = Z_O = 377\Omega$

Parameters	Bo Li <i>et al.</i> [13]	Abbaspour-Tamijani <i>et al.</i> [14]	Li <i>et al.</i> [15]	Xie <i>et al.</i> [7]	In this work
Number of metallic layer	-	3	5	3	3
Element size (λ)	0.21	0.63	0.25	0.72	0.13
Overall thickness (λ)	0.256	0.12	0.27	0.07	0.07
Element Structure	3D structure based on stacked slot lines	Antenna filter antenna	multilayer antenna filter antenna structure	Aperture coupled dual band patch resonator	Aperture coupled single band patch resonator
Polarization	Single	Single	Dual	Dual	Single
Number of transmission zero in band	1	1	0	1	1
Band ratio	2	1.37	1.48	1.23	1.18

TABLE I: Antenna parameters comparison by various authors

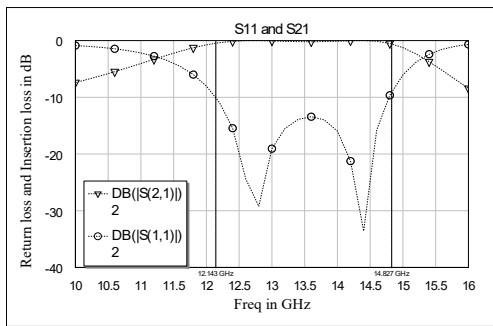


Fig. 5: Transmission zero responses in single band $L = 0.2pH$, $C = 0.205pF$ at free space impedance $Z_{in} = Z_O = 377\Omega$

IV. RESULTS AND DISCUSSION

Fig.2 illustrate reflection coefficient with two transmission zero between $12.14GHz$ and $14.8GHz$ in single band S22 shown in same Fig. 2.

The radiation pattern for approached design refer from Fig. 6. shown in Fig. carries $12dB$, the axial ratio for proposed design shown in Fig. 7. for circular polarization. The axial ratio maintained $2dB$ to $0dB$ for circular polarization, the maximum gain at desired direction have $12dB$. For single band response, the techniques [14] also proposed with second order and fourth order bandpass filter responses. The structures have polarization sensitive and angular range of operation.

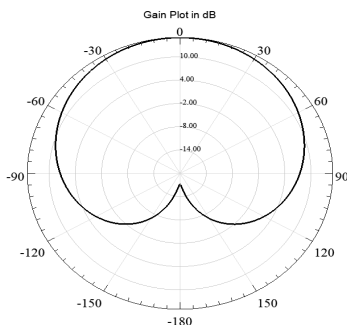


Fig. 6: Gain at $\phi = 90^\circ$, $\theta = -180^\circ$ to 180° at $f_o = 12GHz$

The wave ratio of proposed structure is ≤ 2 , the antenna parameters comparison Table 1 shows the previous works on higher order bandpass resonators, from the literature comparison, this work have key parameters characteristics to be consider for frequency selective surfaces.

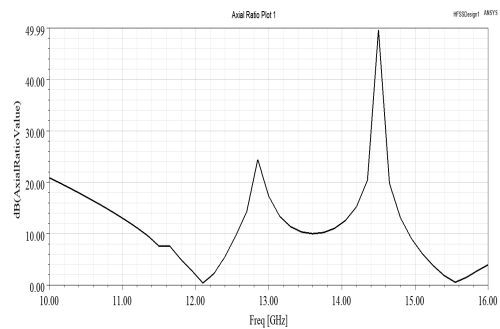


Fig. 7: Axial ratio Vs Freq in GHz at $f_o = 12GHz$

V. CONCLUSION

L shaped truncated unitcell patch along with aperture coupled multilayer resonator antenna proposed to achieve higher order frequency response. Here single band second order response discussed on details and equivalent circuit with analysis also explained. The proposed structure achieved higher order frequency response without introducing number of elements and size. The proposed design resonance at $12GHz$ to $14GHz$. Aperture truncated patch antenna can be used for high frequency applications like IOT based wireless devices, radar receiver where circular polarization on demand and higher order frequency selective surfaces.

REFERENCES

[1] G. Xu, G. V. Eleftheriades, and S. V. Hum, "Generalized synthesis technique for high-order low-profile dual-band frequency selective surfaces," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 11, pp. 6033–6042, 2018.

[2] R. Pous and D. Pozar, "A frequency-selective surface using aperture-coupled microstrip patches," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 12, pp. 1763–1769, 1991.

- [3] J.-M. Xie, B. Li, Y.-P. Lyu, and L. Zhu, "Dual-band frequency selective surface on aperture-coupled patch resonators with different polarization rotation," *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 12, pp. 2632–2636, 2019.
- [4] R. S. Elliott, "On discretizing continuous aperture distributions," *IEEE Transactions on Antennas and Propagation*, vol. 25, no. 5, pp. 617–621, 1977.
- [5] S. X. Ta, V. C. Nguyen, B.-T. Nguyen-Thi, T. B. Hoang, A. N. Nguyen, K. K. Nguyen, and C. Dao-Ngoc, "Wideband dual-circularly polarized antennas using aperture-coupled stacked patches and single-section hybrid coupler," *IEEE Access*, vol. 10, pp. 21 883–21 891, 2022.
- [6] J.-M. Xie, B. Li, and L. Zhu, "Dual-band circular polarizers with versatile polarization conversions based on aperture-coupled patch resonators," *IEEE Transactions on Antennas and Propagation*, pp. 1–1, 2022.
- [7] J.-M. Xie, B. Li, Y.-P. Lyu, and L. Zhu, "Single- and dual-band high-order bandpass frequency selective surfaces based on aperture-coupled dual-mode patch resonators," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 4, pp. 2130–2141, 2021.
- [8] J.-M. Xie, B. Li, L. Zhu, and H. Li, "High-order bandpass polarization rotator based on aperture-coupled patch resonators," *IEEE Antennas and Wireless Propagation Letters*, vol. 20, no. 9, pp. 1809–1813, 2021.
- [9] A. Pirhadi, H. Bahrami, and J. Nasri, "Wideband high directive aperture coupled microstrip antenna design by using a fss superstrate layer," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 4, pp. 2101–2106, 2012.
- [10] H.-T. Chou, "Floquet mode phenomena of infinite phased array antennas in near-field focus applications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 6, pp. 3060–3068, 2013.
- [11] J. I. Echeveste, M. A. González de Aza, and J. Zapata, "Shaped beam synthesis of real antenna arrays via finite-element method, floquet modal analysis, and convex programming," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 4, pp. 1279–1286, 2016.
- [12] C. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2016.
- [13] B. Li, L. Zhu, Y. Tang, Y. Chang, Y. Han, and Y. Lyu, "Wideband frequency selective structures based on stacked microstrip / slot lines," in *2018 International Conference on Microwave and Millimeter Wave Technology (ICMMT)*, 2018, pp. 1–3.
- [14] A. Abbaspour-Tamijani, K. Sarabandi, and G. Rebeiz, "Antenna-filter-antenna arrays as a class of bandpass frequency-selective surfaces," *IEEE Transactions on Microwave Theory and Techniques*, vol. 52, no. 8, pp. 1781–1789, 2004.
- [15] Y. Li, L. Li, Y. Zhang, and C. Zhao, "Design and synthesis of multilayer frequency selective surface based on antenna-filter-antenna using minkowski fractal structures," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 1, pp. 133–141, 2015.

ANALYSIS AND EVALUATION OF A EOD ROBOT PROTOTYPE

1st Silva Vidal Y.
Mechanical Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: ysilvav@unsa.edu.pe

2nd Elvis Supo C.
Electronic Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: esupo@unsa.edu.pe

3rd Milton Ccallata C.
Mechanical Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: mccallatac@unsa.edu.pe

4rd Jesus Mamani G.
Mechanical Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: jmamanigom@unsa.edu.pe

5rd M. Betancur P.
Mechanical Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: mbetancur@unsa.edu.pe

6rd Brunno Pino C.
Mechanical Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: bpino@unsa.edu.pe

7nd Pablo Lizardo Pari Pinto
Electronic Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: pparip@unsa.edu.pe

8nd Erasmo Sulla Espinoza
Electronic Engineering
 Universidad Nacional de San Agustin
 Arequipa, Perú
 Email: esullae@unsa.edu.pe

Abstract—In this document, a prototype of an explosive deactivator robot called JVC 0.1 was analyzed, based on the minimum requirements of the UDEX- Arequipa and the international standards given by the ASTM and NIST organizations, then it was evaluated using the Failure Modes and Effects Analysis (FMEA) matrix. In this way, design parameters are proposed for a next version of this prototype. It was obtained that the robot is oversized in material, dimensions and motors, has a reduced working space for the manipulator with low drive speeds and an inadequate gripper. Therefore, a redesign is necessary.

Index Terms—EOD Robot; Performance testing; FMEA; Overdimensioned.

I. INTRODUCTION

EOD (Explosive Ordnance Disposal) robots are used in many police units that deal with explosives disposal.

In the Arequipa region, the Unit for the Deactivation of Explosives (UDEX-Arequipa) emerged in the 1980s, where several subversive groups used explosive devices to carry out attacks in different parts of Peru [1].

In recent years, EOD robots have been designed with specific characteristics depending on the use that will be given to them, such as the degrees of freedom to access small spaces, the range of the actuator, the type of locomotion to adapt to the type of terrain, the light weight to be transported quickly, and so forth [2]–[4].

For the evaluation of robots, standardized tests established

by both the NIST (National Institute of Standards and Technology) and the ASTM (American Society for Testing and Materials) are used. NIST promotes innovation and industry competition through advances in metrology, standards, and technology [5], while ASTM is an international standards organization that develops and publishes voluntary technical standards agreements for a wide range of materials, products, systems, and services [6].

The proposed tests evaluate mobility, manipulation, human interaction, energy, sensors, radio communication, safety, and durability [7]. And they are used by world-class competitors like the Robocup, where robots are tested in their different configurations and the capabilities of the participants can be objectively compared. In this work, only some representative tests will be taken, since the prototype still has deficiencies that prevent complex tests.

In 2021, UDEX-Arequipa built an EOD robot prototype (Fig. 1), which has a chassis that resembles that of a tank and a robotic arm that has 4 degrees of freedom and a 3-finger claw. This robot was built empirically, so it has many shortcomings when performing operations. In the study carried out at UDEX-Arequipa as part of the project "Design and development of the prototype of a bomb disposal robot equipped with a tele-operated precision actuator" the data of the incidents presented with explosive devices in a period from 2013 to 2020 were analyzed. and the main requirements for the design of an EOD robot in Arequipa were compiled [1]. Currently, a series of advances are being made to obtain an improvement in the user experience regarding the control of



Fig. 1. Robot JVC 0.1



Fig. 3. Robot JVC 0.1

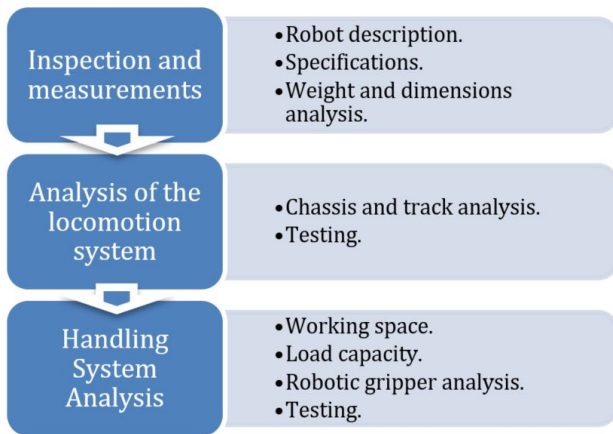


Fig. 2. Methodology

the robot [8] [9] [10] [11] .

II. METHODOLOGY

The objective of this study is to analyze the JVC 0.1 considering some international standards of NIST and ASTM, and specific requirements of UDEX-Arequipa through field tests and digital simulation.

The analysis and evaluation have been divided into 3 parts (Fig 2): First, we will proceed with the general inspection and measurement taking for the 3D modeling by means of SolidWorks computer-aided design. Then, the analysis of the locomotion system will be carried out and later the handling system. The manipulator tests and the locomotion tests will be guided considering the ASTM and the NIST, to see the effectiveness of the robot. Finally, the design parameters will be defined by using the FMEA matrix [12].

III. ANALYSIS OF THE ROBOT JVC 0.1

A. Inspections and measurements of the robot JVC 0.1

The JVC 0.1 robot (Fig. 3) is a hybrid robot, it means that it consists of a mobile platform and a manipulator arm. It has the following characteristics:

- 1) It has a system of caterpillars with independent motors, to move in unstructured terrain, this system allows turning on its own axis and climbing stairs with an inclination of up to 45 degrees.
- 2) It has DC servomotors with reductive gearboxes to control the speed, provide high torque, and maintain the position of the arm, even without power.
- 3) Anthropomorphic manipulator arm, with 4 degrees of freedom, with a three-finger gripper for handling and grasping objects.

When measuring the total mass of the JVC 0.1, a value of 155.20 kg was obtained. According to the ASTM E-2592-07 packaging weight and volume evaluation standard for emergency robots [13], the weight of a robot in emergency scenarios must be up to 90 kg so that two people can move it; however, in this case the robot exceeds this limit by 73%, requiring a minimum of four people to transport it.

TABLE I
WEIGHT DISTRIBUTION OF THE JVC 0.1.

Components of robot JVC 0.1	Weight(Kg) / Percentage
Motors	32.92 / 21%
Battery	18.624 / 12%
Chassis	68.288 / 44%
Robotic arm	35.696 / 23%

From table I, we deduce that the construction material of the chassis as well as the robotic arm can be changed for one of lower density, considerably reducing the weight dramatically. Nevertheless, motors and the battery are elements that correspond to an important part of the total weight, because even if the weight were reduced due to a possible change of the construction material, these elements will continue to represent an important weight. In the case of the battery, by changing it for a lighter lithium alloy one to reduce weight and space by taking advantage of its high energy density, and in the case of motors, by using lower power motors.

Regarding dimensions, the Allen Vanguard MK2 [14] was considered as a reference to compare it with the JVC 0.1,

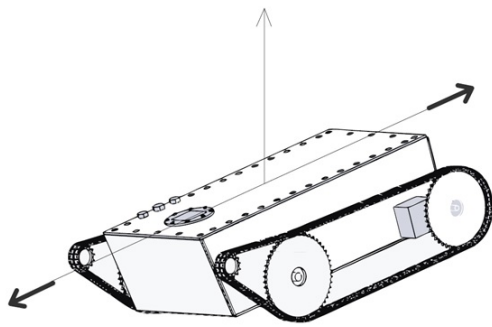


Fig. 4. Chassis JVC 0.1

because it is a model that can be easily moved by the agents in the UDEX-Arequipa patrols, it can go through doors and be transported in small spaces without many complications. Table 2 shows a comparison between the dimensions of the MK2 robot and the JVC 0.1.

The general dimensions required for a packaging of the robots are considered; however, in the case of the height of the JVC 0.1, two height measurements were taken to appreciate to what extent the robotic arm influences the volume, which in the MK2 is negligible because it has a folding arm, which is much more convenient. by requiring less space moved and lowering the center of gravity to achieve greater stability. From table 2 we can deduce that the cubic volume required to transport the JVC 0.1 is more than 4 times that of the MK2, so it is convenient to reduce dimensions and have a folding arm in the next version.

TABLE II
DIMENSIONS MK2 VS. JVC 0.1

	Length (m)	Width (m)	height (m)	Cubic volume (m ³)
MK2	0.915	0.435	0.405 with arm	0.161
JVC 0.1	1.00	0.70	0.97 with arm 0.30 without arm	0.679

B. Analysis of the Locomotion System

The JVC 0.1 chassis (Fig. 4) has a box-shaped architecture, is made of 6mm thick steel sheet, and houses the electronic control systems, batteries and serves as a support for the caterpillar-type traction system.

It consists of two motors with 12V DC reductive gearboxes wound in series with an output of 1.2KW each for a total of 2.4 KW, with independent movement. Each motor is coupled to a gearbox and later to an output shaft which is coupled to the tracks for movement, this allows it to support up to 400 Kg of additional weight on the chassis and the towing capacity of a vehicle of up to 1361 Kg on flat surface (Fig. 3) Its weight of 155.20 kg, allows greater load capacity and stability; however, at UDEX-Arequipa such capacities are not



Fig. 5. JVC 0.1 robot towing a UDEX vehicle

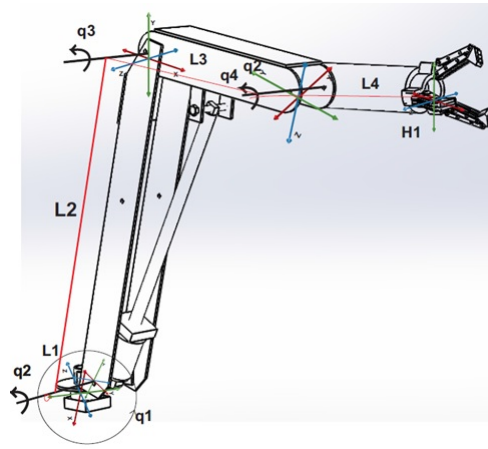


Fig. 6. Assignment of parameters following the DH algorithm

required and reducing the power of these motors is convenient. The traction system allows it to travel at a speed of 0.70 km/h, providing a low degree of maneuverability in closed spaces.

C. Handling System Analysis

The dimensions of the robot were considered to carry out the study of the position. Analysis was performed using the Denavit-Hartenberg (DH) algorithm [15]. In (Fig. 6) the assignment of the DH parameters is shown and their values in Table 3.

TABLE III
PARAMETROS DH

Articulation	θ	D	A	a
1	θ_1	d1	0	90
2	θ_2	0	a2	0
3	θ_3	0	a3	0
4	θ_4	0	a4	90

The use of DH parameters allows us to find the orientation and position of each joint in space with respect to a defined coordinate system. From Table 3, by means of transformation and translation matrices, we found the working space of the manipulator. The results obtained are shown in (Fig. 7) The workspace of the JVC 0.1 compared to commercial robots like the MK2 is noticeably smaller. In addition, considering that the dimensions of its links are larger, it is not able to take advantage of its linear drive system and its joints are limited.

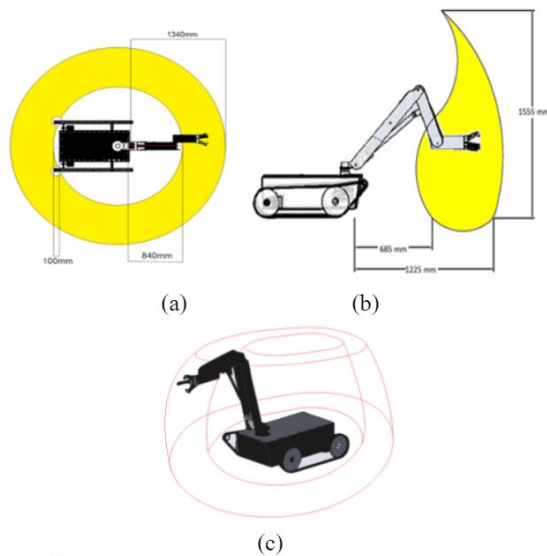


Fig. 7. Work area of the manipulator. (a) Projection on the floor of the Work Area. (b) Side view workspace projection. (c) Workload

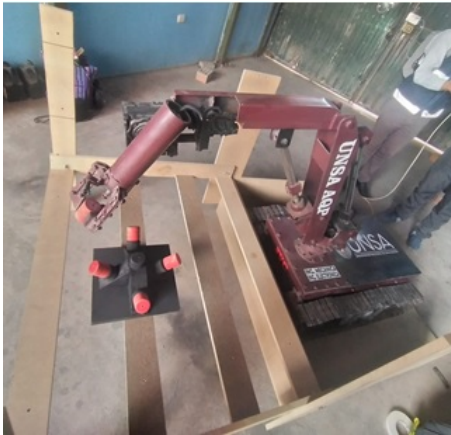


Fig. 8. Robot JVC 0.1 at test bench

At ground level (Fig 7a) the work area is not optimal, since it has difficulties with the actuator when approaching the chassis. So as the arm is raised in the side view (Fig 7b) the space is reduced. In (Fig 7c) the size of the robot and its volume of work can be seen, which barely becomes a ring around the robot.

D. Test result

Tests were performed according to ASTM and NIST standards by performing the test bench with the given specifications [13]. In (Fig. 8) it can be seen how the robot picks up the targets of the omnidirectional device.

The results obtained from these tests led us to understand that the robot is too slow to perform these tests, because the actuators limit the movement of the arm joints, obtaining average speeds of 0.12 rpm at the shoulder and 0.28 rpm at the elbow.

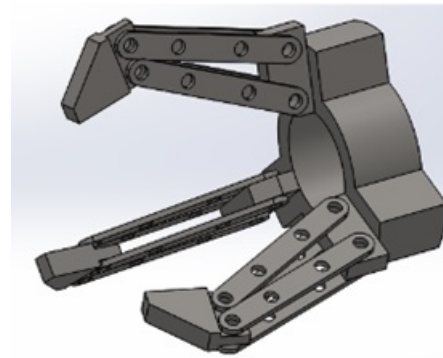


Fig. 9. Robotic Gripper of the JVC 0.1

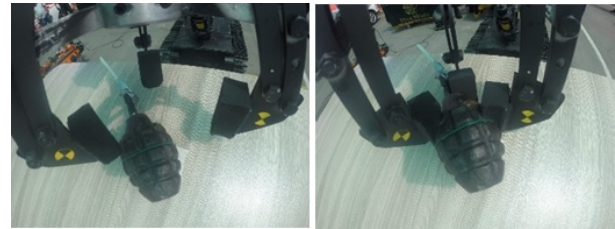


Fig. 10. Gripper holding a pineapple-type grenade

E. Robotic Gripper

The robotic gripper (Fig. 9) belonging to the JVC 0.1 presents three fingers separated by 120° that are activated simultaneously by a power screw mechanism located in the central axis of the gripper.

The manipulator can lift different shapes of objects. In tests carried out at UDEX-Arequipa, it was found that due to their geometry and with proper positioning, the grippers are capable of lifting objects such as grenades, dynamite, among others.

For a weight of 15Kg (maximum load) a Clamping Force of 163N is obtained, being more than necessary (147 N) to hold the object firmly. The main problem is the lack of rotation of the gripper, since the JVC 0.1 is forced to perform a large number of maneuvers to pick up or deposit an object to align the jaws accordingly the shape of the object.

Therefore, the last link in the arm requires an additional degree of freedom. Being a 3-finger gripper, it has a greater grip when grasping an object, the main shapes it could capture are spherical (Fig. 10), in the case of cylinders and prisms, it can easily grasp them only if they are vertical. A 2-finger gripper could easier grasp a dynamite (cylindrical object) and in the same way prismatic objects, it would also adapt to spherical objects if the fingers were aligned to the center of gravity of the object; thus, it is concluded to use a 2-finger gripper in the redesign.

F. Robotic Arm Load Capacity

When gathering the requirements of the UDEX-Arequipa, (2013 - 2020), they indicate that a load of 10Kg is enough to cover 97% of all incidents with explosive devices [1]. So, we



Fig. 11. Load capacity test

would have a 50% margin that would be wasted by increasing weight and power consumption unnecessarily. For the JVC 0.1 loading test, 3 explosive devices were selected at UDEX-Arequipa as shown in (Fig.11), grenade, mortar and tank shell. The targets were picked up from the ground and maneuvered in extended arm position. Sequences of 10 operations were performed for each object and the results of effectiveness were obtained in the following table.

TABLE IV
ROBOT LOADING EFFECTIVENESS WITH RESPECT TO DIFFERENT EXPLOSIVES.

Explosive Device	Mass (kg)	Effective percentage
Grenade (0.6kg)	0.6	100
Mortar (13.5kg)	13.5	100
Tank bullet (15.3kg)	15.2	80

According to the recommendations of the ASTM for carrying out these tests, we must reach a success rate of over 80% [6]. The results in Table 4 reflect that the JVC 0.1 had no inconveniences with the first two artifacts, but with the tank bullet it barely achieved 80%.

The tank bullet load shows enough capacity in actuators, but at this point the chassis begins to lose stability. Despite the high weight of the design, the incorrect distribution of mass in the chassis decreases the load levels, making it unstable in the position of extended arm. For this reason, 15.2 kg was considered as the maximum load.

IV. FMEA MATRIX

The FMEA matrix is a method that is used to analyze a product or process in its design phase, in this case we will analyze the JVC 0.1 bomb disposal robot. The main objective with the FMEA matrix will be to highlight the critical points to eliminate them or establish actions to be taken to avoid their appearance in the next version of the robot [12].

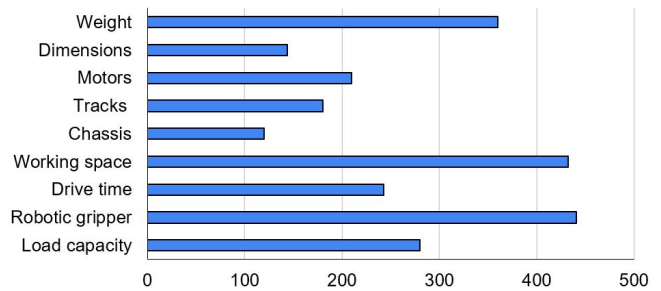


Fig. 12. RPI matrix FMEA

As seen in table 5, the failure modes were divided into the three stages mentioned above, their causes and effects are described as well as the actions to be taken that would become our redesign parameters for a latter version.

In (Fig. 12) the Risk Priority Index (RPI) found for each analyzed aspect is plotted. An RPI of less than 100, an adequate one, which does not require any intervention; however, we have higher values, that is, the JVC 0.1 prototype is far from optimal for UDEX-Arequipa. The most critical aspects will be regarding the robotic gripper and the workspace, where the increase of one degree of freedom for the gripper rotation, the redesign of the two-finger gripper and the change of the drive system will be extremely essential. With a little less criticality, we have to reduce the weight of the robot, adjust the load capacity without oversizing motors and structure, as well as increase the drive speed.

V. CONCLUSIONS

In this work, the JVC 0.1 explosive deactivation robot prototype was analyzed through performance tests and digital simulation, where it was found that this design was not the most suitable for the tasks at UDEX-Arequipa, mainly because it was very difficult to transport, it had a reduced workspace, the actuators were oversized and it had low speed in the robotic arm drive. According to the studies carried out on the JVC 0.1 robot, its redesign with the following parameters is of vital importance:

- 1) Aluminum main construction material.
- 2) The last link in the robotic arm requires one degree of freedom to rotate the robotic gripper.
- 3) Increase speeds, to improve the user experience of a conventional EOD robot.
- 4) Reducing the size of the robot
- 5) Use a two-finger robotic gripper.
- 6) Change the linear drive system to worm-crown gear-boxes.

The JVC 0.1 prototype, in addition to having the difficulty of being the first of the EOD robots manufactured in the city of Arequipa, had a short time for its development. This factor did not allow the complete definition of client traits; therefore, the specifications of JVC 0.1 were not the best. The new JVC 0.2 version will be designed based on the mentioned parameters. First, modeling, simulation and analysis will be

TABLE V
FMEA MATRIX

Function	N°	Failure Mode	Effect	Causes	G	F	D	RPI	Actions to take
Inspections and Measurements	1	Weight	Hard to move	Excessive material. Acid-lead battery. Motors	9	8	5	360	Switch construction materials to aluminum. Use lithium batteries. Reduce motor power.
Inspections and Measurements	2	Dimensions	Hard to move	Requirements missing	6	8	3	144	Reduce JVC 01 size closer to MK2. Design arm in a rest position.
Analysis of the Locomotion system	3	Motors	Unnecessary power, energy waste	Oversized motors	5	7	6	210	Reduce motors power
Analysis of the Locomotion system	4	Caterpillar	Stranded robot	Link turnbuckles missing	6	6	5	180	Design caterpillar system with turnbuckles.
Analysis of the Locomotion system	5	Chassis	Unnecessary material	Too heavy structure	4	6	5	120	Reduce material. Distribute weight for extended arm load.
Analysis of Manipulation System	6	Workspace	Reduced workspace. Objects near to the chassis cannot be gripped	Limited joints due to motors adapted to linear.	9	8	6	432	Switch to a worm gear reducers system.
Analysis of Manipulation System	7	Drive time	Too long response time	Motors adapted to linear.	9	9	3	243	Increment motors speed.
Analysis of Manipulation System	8	Robotic gripper	Objects cannot be obtained easily.	Badly-designed gripper. Gripper twist missing.	9	7	7	441	Redesign of 2-fingered gripper. Augment the freedom degree for gripper twist.
Analysis of Manipulation System	9	Load capacity	Wasted energy. Unnecessary material	Oversized motors and structure	7	8	5	280	Lighten up linkers structure. Dimension motors adequately up to 10 kg extended arm

carried out in software. Then we will move on to construction and evaluation. The redesign will be carried out in two tasks, one focused on the robotic arm and the other on the chassis, where standardized tests according to NIST and ASTM will take on greater relevance [16]–[19].

VI. ACKNOWLEDGMENTS

Our thanks to the Universidad Nacional de San Agustín de Arequipa for the development of the project through the information collected and for their invaluable guidelines to explore the facets of this work "Design and development of the prototype of a bomb disposal robot equipped with a teleoperated precision actuator" with the contract N. IBA-IB-27-2020-UNSA and UDEX-AQP; therefore our participation in this initial stage. Also, to the Explosives Deactivation Unit of Arequipa for giving us access to its facilities, the required information, and the time to carry out the tests.

REFERENCES

[1] D. V. G. E. S. C. E. S. Y. S. V. J. G. Mamani, P. Pari, "Compilation and analysis of requirements for the design of an explosive ordnance disposal robot prototype applied in udex-arequipa," in *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–8.

[2] S. Deng, H. Cai, K. Li, Y. Cheng, Y. Ni, and Y. Wang, "The design and analysis of a light explosive ordnance disposal manipulator," in *2018 2nd International Conference on Robotics and Automation Sciences (ICRAS)*, pp. 1–5, 2018.

[3] J. Yuan, W. Zhang, and J. Tao, "Development of big danger disposal manipulator - proposal and mechatronic system design -," in *2008 IEEE International Conference on Robotics and Biomimetics*, pp. 1415–1420, 2009.

[4] "O. olwan, a. matan, m. abdullah, and j. abu-khalaf, "the design and analysis of a six-degree of freedom robotic arm for explosive ordnance disposal applications," 2015.

[5] "National institute of standards and technology," in *NIST". NIST U.S. Department of Commerce*, 2009.

[6] "Standard test methods for response robots astm international standards committee on homeland security applications; operational equipment," in *Robots*, 2009.

[7] A. H.-M. H. Jacoff, "Emergency response robot evaluation exercise," in *Natl. Institute Standards and Techno*, pp. 737–742, 2010.

[8] D. Vilcapaza Goyzueta, J. Guevara Mamani, E. Sulla Espinoza, E. Supo Colquehuanca, Y. Silva Vidal, and P. P. Pinto, "Evaluation of a nui interface for an explosives deactivator robotic arm to improve the user experience," in *International Conference on Human-Computer Interaction*, pp. 288–293, Springer, 2021.

[9] M. A. Andres, L. Pari, and S. C. Elvis, "Design of a user interface to estimate distance of moving explosive devices with stereo cameras," in *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, pp. 362–366, 2021.

[10] M. Postigo-Malaga, E. Supo-Colquehuanca, J. Matta-Hernandez, L. Pari, and E. Mayhua-López, "Vehicle location system and monitoring as a tool for citizen safety using wireless sensor network," in *2016 IEEE ANDESCON*, pp. 1–4, 2016.

[11] Y. L. S. P. L. E. S. C. Juan Montoya A., Erasmo Sulla E., "Analysis of a user interface based on multimodal interaction to control a robotic arm for eod applications,," *MDPI, Electronics*, vol. (in press).

[12] I. N. de Seguridad y Salud en el Trabajo in *INSST. NTP 679: Análisis modal de fallos y efectos. AMFE.*, 2009.

[13] "Standard practice for evaluating cache packaged weight and volume of robots for urban search and rescue 1," in *S. Practice*, pp. 1–4, 2007.

[14] Allen-Vanguard in *MK2*, 2004.

[15] R. H. Denavit, "na notación cinemática para mecanismos de pares inferiores basada en matrices," in *Tran. de la ASME. Revista de Mecánica Aplicada*, no. 23, pp. 215–221, 1955.

[16] P. L. B. C. A. R. Barrientos, A., "Fundamentos de robótica, 2nd edition, mcgraw-hill," No. 33, pp. 7–20, 2007.

[17] y. M. MerletJ.-P., GosselinC., "Espacios de trabajo de manipuladores planos paralelos," in *Teoría de Mecanismos y Máquinas*, pp. 7–20, 1998.

[18] L. Z. . S. S. Murray, R. M. in *A mathematical introduction to robotic manipulation. Florida CRC Press: Boca Raton*, 1994.

[19] A. . B. X. . A. A. Berisha, Jakup Pajaziti in *De-Mining Techniques of Improvised Explosive Materials by the usage of Mobile Robots.*, 2008.

A Model of Classification of Consumers on the Retail Electricity Market

Aleksandr Belov
Department of the Applied
Mathematics

National Research University "Higher
School of Economics
Moscow, Russia
0000-0001-7193-0633

Maria Monina
Department of the Applied
Mathematics

National Research University "Higher
School of Economics
Moscow, Russia
m.monina@hse.ru

Zhenisgul Rakhmetullina
Faculty of Basic Engineering Training
D. Serikbayev East Kazakhstan State
Technical University
Ust-Kamenogorsk city, Kazakhstan
ZhRakhmetullina@ektu.kz

Abstract In this paper, we study the problem of forecasting the energy consumption of real estate objects. The purpose of the work is to increase the efficiency of distribution companies by improving the mechanism for determining planned consumption volumes. The task of the work is to construct a mathematical model based on existing data and further forecast on its basis. To achieve the goal of the work, 2 approaches are used: forecasting energy consumption based on time series (regression task) and forecasting the energy consumption profile for each object (classification task). Various methods of machine learning, as well as clustering, were used in the work. We compared machine learning methods that solve different problems with each other, and selected the best ones. As a result, the ARIMA and Support Vector Machine methods are compared for predicting energy consumption. The choice of a particular algorithm will depend on the goals of the companies that produce and distribute energy.

Keywords— Energy consumption, mathematical model, machine learning, ARIMA, SVM method

I. INTRODUCTION

Currently, the European Union has the Treaty on the European Union and the Treaty on the Establishment of the European Community of December 13, 2007 (entered into force on December 1, 2009), which refers to the following main goals in the development of energy:

- Ensuring the functioning of the energy market;
- Ensuring energy security;
- Promoting energy efficiency and energy conservation, including the use of renewable energy sources (RES) and the development of new types of energy;
- Facilitate interconnection of energy networks.

To implement these goals, acts are adopted that are binding on all countries that are members of the European Union. One of these documents are the decisions and directives of the European Parliament and the Council of the European Union. Until recently, Directive 2009/72/EC of the European Parliament and of the Council of July 13, 2009 on general rules for the internal electricity market and the repeal of Directive 2003/54/EC [1], which establishes the basic rules for the production of , transmission, distribution and sale of electricity, taking into account the protection of consumer rights, as well as the principles of organization and functioning of the electricity industry to increase competitiveness and integrate electricity markets between the countries of the European Union. In particular:

- Separation of natural monopoly (transfer, transportation) activities from competitive ones (production, sales);
- Increasing the share of renewable energy use from 20% to at least 27% by the end of the 2030s;

- Limiting the monopoly of electricity suppliers that block the supply of electricity from competing energy companies through ownership of delivery networks to end consumers (power lines).

One of the main regulators in the electric power industry is the Directive of the European Parliament and of the Council No. 2019/944 of 06/05/2019 on the general rules of the internal electricity market and amending Directive 2012/27/EU [2].

The main provisions of the fourth energy package:

- Separation of natural monopoly (transfer, transportation) activities from competitive ones (production, sales). Despite the fact that this was also indicated under the third energy package, in many countries of the European Union there is still a share of regulated electricity supplies;
- Decentralization of electricity consumption and generation;
- Increasing the share of renewable energy use from 20% to at least 32% by the end of the 2030s [3].

The fourth energy package is aimed at improving the energy efficiency of the European economy, providing maximum freedom to consumers, and achieving the global leadership of the European Union in renewable energy.

In the European Union, much attention is paid to RES, in the field of energy a new set of rules has been developed, which is called the Clean Energy for all Europeans package, adopted in May 2019 [8]. In this package the Commission considered five main aspects of development:

- Energy efficiency - the use of energy-saving technologies in industrial and commercial buildings that consume a large amount of energy.
- The use of renewable energy sources - reaching the mark of 32% of the share in the use of renewable energy sources from the total number of energy sources used by 2030.
- Better governance in the Energy Union – each member state of the European Union is required to develop National Energy and Climate Plans (National Energy and Climate Plans) for 2021-2030, which will describe exactly how the state plans to achieve the goals set by the Energy Union, in particular in energy efficiency and the use of renewable energy sources. These projects must be reviewed and approved by the European Commission.
- More rights for consumers - individuals will be able to produce, store and sell their own energy, as well as keep track of generated electricity bills and be

able to change the distribution company at their own discretion.

A more efficient electricity market - the new energy package will improve the security of supply, integrate renewable energy into the power grid, and effectively organize cross-border cooperation [5].

Thus, we can conclude that Europe is aimed at changing the structure of the energy market and making the transition to unregulated management of the electricity market, and in the European Union, much attention is paid to the environmental situation and the use of renewable energy [6].

With the advent of a competitive unregulated electricity market, there is a separation of energy suppliers from network operators and the creation of Distribution System Operator (DSO) as a separate operating entity, as well as a separation between supply and network operations, both in terms of operations, and in terms of access to operational data and ownership. This means that the energy supplier will continue to use its system for end-user billing, while the PDM will need to implement a comprehensive meter data management (MDM) system to collect and store metering and customer data for all consumers (and manufacturers associated with distribution) of a country or region, while suppliers will only have access to their own customer data

Thus, for PDM there is a need (within the framework of the third and fourth energy packages) to forecast the consumption of electricity for each of the sales companies so that the PDM can effectively distribute the volume of generated electricity between the sales companies according to their needs. Due to the fact that the number of consumers for each supplier will be quite large, the consumption of consumers will be different, and it is labor-intensive and unproductive to build forecasts for each consumer separately, it makes sense to divide all consumers into classes with certain similar characteristics, and the nature of subscriber consumption in class should be approximately the same for the same reporting period. At the same time, it is also necessary to take into account the historical consumption of subscribers, since seasonal factors can also affect the nature of consumption.

At present, an important research task is to predict the magnitude of energy consumption. It is relevant for such areas as the economy, industrial production, environmental protection, etc. But for companies producing and distributing energy, this task plays an important role, since accurate forecasting provides more profit and minimizes the risks associated with incorrect calculation energy consumed by objects, as well as the optimal distribution of energy [7]. However, due to the large number of types of consumers and factors affecting the consumption profile, forecasting is a difficult research task.

II. PROBLEM STATEMENT

The purpose of the work is to increase the efficiency of energy distribution companies by improving the mechanism for determining planned consumption volumes.

To achieve the goal of the work, namely to improve the efficiency of distribution companies, it is necessary to develop

an algorithm that solves the problem of consumption forecasting. There are many methods that solve the forecasting problem, in this paper supervised machine learning methods are used, which are divided into two large types depending on the task: solving the regression problem and solving the classification problem. One of the tasks of the work is to determine how the type of supervised MMO is best suited to achieve the goal.

Thus, the paper considers two ways to solve the problem of predicting the energy consumption of real estate objects: based on historical data on energy consumption and based on the belonging of a real estate object to a certain consumption profile. The consumption profile of each object is determined using clustering.

Main tasks have been:

- Analysis of energy consumption data;
 - Building a mathematical model based on consumer parameters;
 - Clustering of real estate objects by consumption profile;
- Solution of the energy consumption forecasting problem in two ways: through the regression task using historical data, and through the classification task using the results of clustering.

III. METODOLOGY

A. Data Preprocessing

Data for analysis was exported from the Oracle MDMS database using Oracle SQL. Oracle MDMS is a consumption calculation tool that uses special calculation algorithms and takes into account the metering device type of appliance and the client's metering point. Since the data was extracted from different databases, the data was then combined into one dataset and processed to extract empty and anomalous values. One of the important advantages of MDMS is the support of the meter-to-cash process - the provision of calculated values at the request of an external system [8, 9]. The external system is CC&B, which queries MDMS for meter readings at the time of billing. Figure 1 shows a graph of daily electricity consumption, built on the basis of data from MDMS.

Fig. 1. CHART OF THE DAILY ELECTRICITY CONSUMPTION



To conduct cluster analysis, it is necessary to prepare the initial data. To determine which data and characteristics influence consumption volumes, it is necessary to put forward

a number of hypotheses. A hypothesis is an assumption about the influence of certain features on the formation of clusters.

It is necessary to analyze and display all the signs, and then evaluate them. Thus create a feature space. Feature selection is based on available data. Therefore, to determine the features, the existing data sample should be examined and cleared of anomalous values. After that, it is necessary to check the signs for correlation. If a correlation is found, regularization should be performed.

In the process of data research and analysis of current trends in the development of the electricity market, the following hypotheses were put forward:

1) The type of device affects the algorithm for calculating the consumed energy.

2) Metering Equipment type affects the algorithm for calculating the consumed energy

3) The process of electricity consumption is characterized by cyclicity and seasonality, which makes it possible to single out such features as seasonality and peak values of consumption for each consumer during the month. Just as important is the overall level of consumption.

The data includes information on the energy consumption of 240 commercial properties on a monthly basis from 2018 to 2021 in kW.

Each object is characterized by several parameters:

- SP Type (service point of consumption type), for example, WPUMP – Water pumping,
- Premise Type, for example, Building
- Region, for example, Nicosia
- Meter Type, for example, Low voltage single.

We consider properties belonging to the following types of industries (according to SIC- Standard Industrial Classification):

- Residential Rural
- Growing of crops
- Dairy products
- Bakery
- Beverage
- Tobacco products
- Wear and accessories
- Souvenirs
- Fishing and aquaculture

All consumers of electricity should be divided into classes according to similar consumer characteristics (for example, individuals, industrial, commercial consumers, etc.). It is assumed that consumers of the same class have a similar consumption pattern, taking into account the influence of seasonal factors. In the current context, a consumer is understood as a service metering point from which readings are taken. The nature of consumption at these points may differ depending on the technical characteristics of the connection, the connected devices, the seasonality of the use of the connected devices (for example, a warehouse heater), the nature of the subscriber's agreement with the power supply company, the tariffs used, the region of consumption, and

even proximity / remoteness from the central regions, which is determined by the postal code.

Thus, in order to separate consumers by trend regions, it is necessary to identify significant criteria (factors) that affect the volume of consumption.

In order to divide objects into clusters according to the consumption profile, clustering must be carried out according to criteria directly related to energy consumption. Therefore, the average values of energy consumption of real estate objects for the 2nd, 3rd and 4th quarters of 2021 were taken as criteria:

- o April 2021 – June 2021
- o July 2021 – September 2021
- o October 2021 – December 2021

Thus, for clustering the studied data, the k-means method is used with the join rule for clusters "single connection" and the parameters for clustering - the last 9 months of 2021 - 2,3,4 quarters with an average value of energy consumption.

B. Machine Learning Methods

As part of the problem of forecasting energy consumption, the regression problem is interesting in that a mathematical model is built for each property based on historical data on the energy consumption of this object, and for each object its own approximating function is built, taking into account the behavior of the object as an energy consumer. Accordingly, the better the regressor works, the better the amount of energy that will be spent in the future is predicted.

As part of the energy consumption forecasting problem, the classification problem is interesting in that it can be used to divide real estate objects into classes based on their electricity consumption profile, and predict for future periods not the amount of electricity that will be spent, but the consumption profile for a particular object. Such a task may be of interest to the company, since it is not always necessary to predict the total amount of electricity; more often, more detailed information about the behavior of real estate as consumers of electricity is of interest.

IV. COMPUTER SIMULATION RESULTS

To train the models, the data under study were divided into training and test sets. 240 real estate objects were divided in the ratio of objects in the samples - 80% and 20%, respectively. For each object, its time series is investigated - the values of energy consumption in the period from 2018 to 2021. Objects are serial numbers of the months, answers are energy consumption values.

The following methods were used to solve the regression problem [10, 11, 12]:

- Linear regression
- Support Vector Machine (SVM)
- ARIMA
- SARIMA

Python and libraries: scikit-learn, Statsmodels were used for computer simulation.

For each model, a graph of the dependence of the object number and the error percentage for this object was presented (Fig. 2 - Fig. 4).

Fig. 2. GRAPH FOR LINEAR REGRESSION

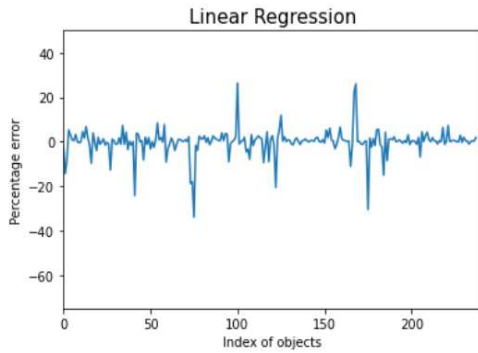


Fig. 3. GRAPH FOR ARIMA

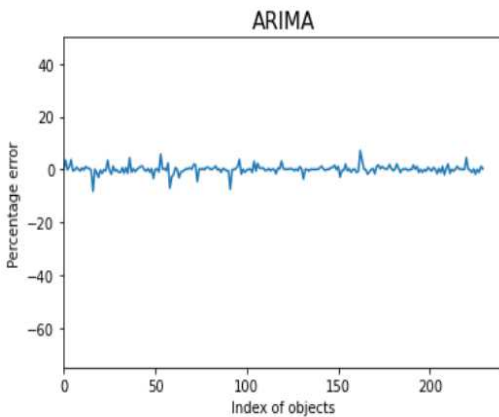
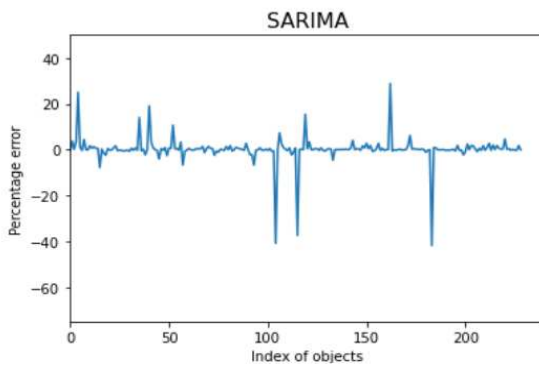


Fig. 4. GRAF FOR SARIMA



To assess the quality of forecasting, the MAPE (mean absolute percentage error) metric was used - the average absolute error in percent. It is given by the formula

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{A_t - F_t}{A_t} \right|, (1)$$

where n is the number of objects, A_t is the actual responses, F_t is the predicted responses. The results of assessing the quality of forecasting are presented in Table 1.

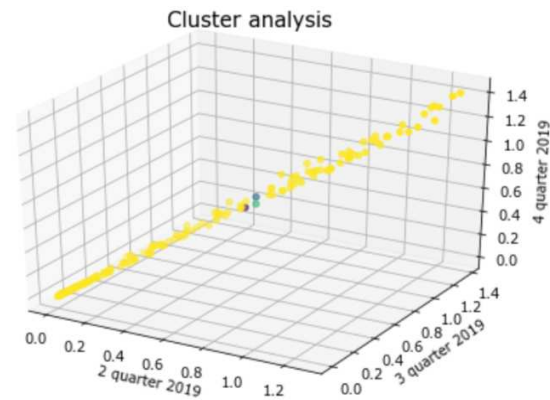
TABLE I. MAPE METRIC FOR MODELS SOLVING THE REGRESSION

Method	MAPE
Linear Regression	3.0106
ARIMA	0.9023
SARIMA	1.8186

For the clustering of 240 properties, the k-means method was used with the metric "single link" combined based on historical data for the last 3 quarters of 2019 - average values of energy consumption of properties for 2, 3 and 4 quarters of 2019.

For clustering, a similarity metric was chosen - the "distance" metric - the Euclidean distance, since when using the "squared Euclidean distance" and "Manhattan distance" metrics, the composition of the clusters practically did not change, and when using the "Chebyshev distance", the objects were practically not divided into clusters (Fig.5).

Fig. 5. REAL ESTATE CLUSTERING WHEN USING THE "CHEBYSHEV DISTANCE" METRIC



The standard k-means algorithm takes the parameter k - the number of clusters, that is, the number of final clusters is not determined automatically, and additional analysis is required to determine the optimal value of the parameter k for the data under study [11, 12].

Dividing the objects under study into 2 clusters will not give good results in predicting, therefore, you need to start analyzing possible options for the parameter k, you need to start with 2. Figure 6 shows the clustering of the objects under study with the parameter k=3, that is, with the number of final clusters 3.

Fig. 6. REAL ESTATE CLUSTERING WHEN USING "EUCLIDEAN DISTANCE" METRIC WITH PARAMETER K=3

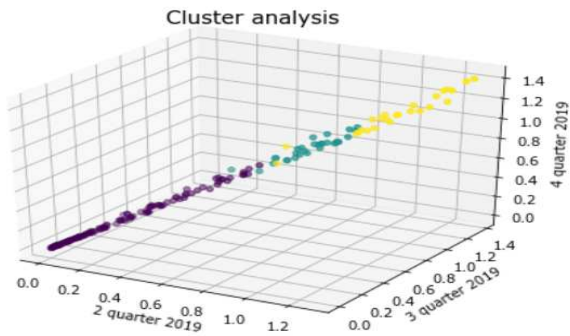


Fig. 7 Fig. 8 shows the clustering of the studied objects with parameters k=4 and k=5.

Fig. 7. REAL ESTATE CLUSTERING WHEN USING “EUCLIDEAN DISTANCE” METRIC WITH PARAMETER K=4

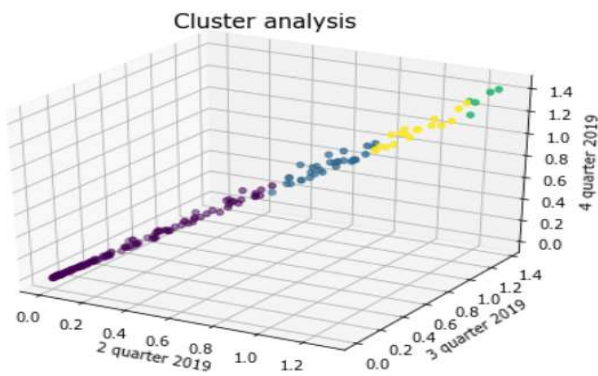
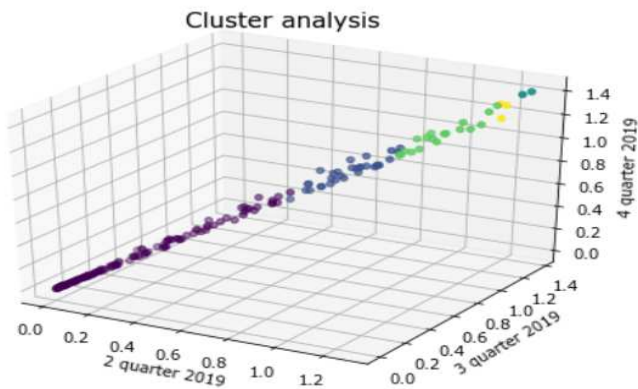


Fig. 8. REAL ESTATE CLUSTERING WHEN USING “EUCLIDEAN DISTANCE” METRIC WITH PARAMETER K=5



As can be seen from the graphs, with an increase in the number of clusters, the composition of clusters does not change much, large clusters are detailed, and the number of objects in clusters becomes disproportionate. For a given set of objects, namely 240 objects, it is sufficient to use k=3.

An example of the Python code for the clustering algorithm is shown in Fig. 9

Fig. 9. PYTHON CODE FOR THE CLUSTERING ALGORITHM

```
def clustering(d, k):
    q = queue.PriorityQueue()
    sets = [set([i]) for i in range(len(d))]
    nb_sets = len(d)
    for i in range(len(d)):
        for j in range(i+1, len(d)):
            q.put((distance((d[i][1],d[i][2],d[i][3]), (d[j][1], d[j][2], d[j][3])), (i, j)))
    dis = 0
    #print(q.get())
    while not q.empty() and nb_sets >= k:
        e = q.get()
        u = e[1][0]
        v = e[1][1]
        dis = e[0]
        set_uv = [s for s in sets if u in s or v in s]
        if len(set_uv) > 1 and set_uv[0].isdisjoint(set_uv[1]):
            unionset = set_uv[0].union(set_uv[1])
            sets_c = copy.copy(sets)
            sets.remove(set_uv[0])
            sets.remove(set_uv[1])
            sets.append(unionset)
            nb_sets -= 1
    return sets_c
```

```
d = {}
for i in range(len(data_for_clust)):
    x = x_list[i]
    y = y_list[i]
    z = z_list[i]
    d[i] = (i, x, y, z)
```

```
k = 3
cl = clustering(d, k)
```

```
color_list = np.zeros(len(d))
```

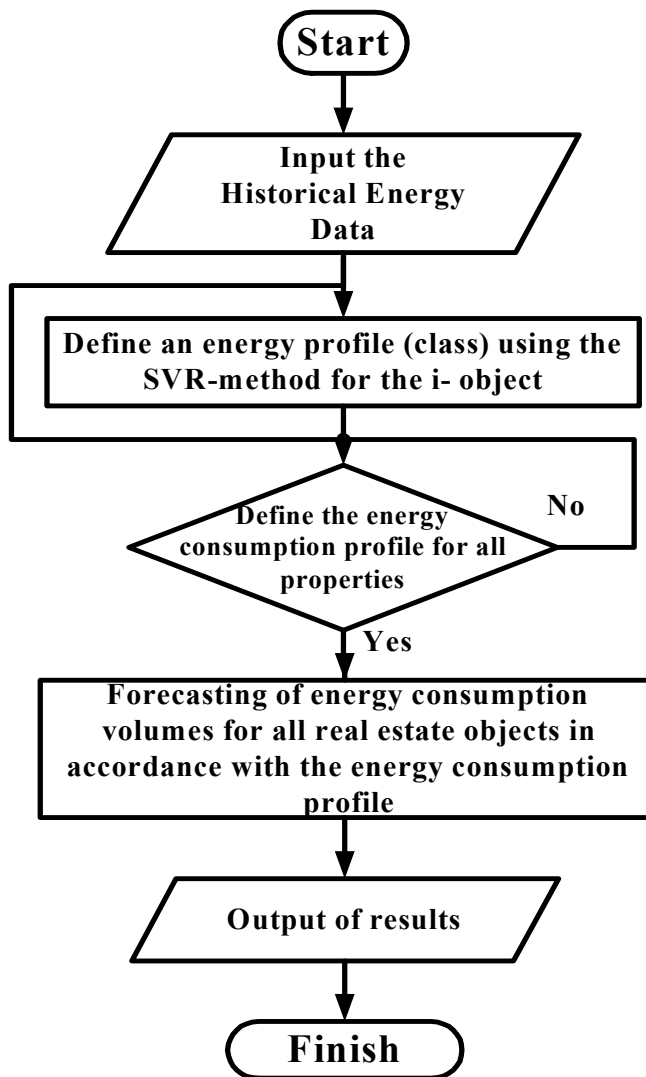
```
for i in range(len(cl)):
    for item in cl[i]:
        color_list[item] = i
```

```
d = {}
for i in range(len(data_for_clust)):
    x = x_list[i]
    y = y_list[i]
    z = z_list[i]
    d[i] = (i, x, y, z)
```

The proposed algorithms for clustering and forecasting energy consumption will be used to implement the energy management system for real estate objects.

Fig. 10 shows the profiling algorithm for electricity consumers, which will be used in the energy management system.

Fig. 10. BLOCK DIAGRAM OF THE ENERGY PROFILING SYSTEM



V. CONCLUSION

The paper analyzes the data provided by the energy company of one of the European countries. The following are applied to the data: clustering, which determines the object's consumption profile; mathematical models suitable for describing the data under study; machine learning methods that solve regression and classification problems. In addition, the evaluation of the work of applied machine learning algorithms with the data under study was carried out.

Using the Python programming language, machine learning methods and clustering methods were implemented. Methods for assessing the quality of algorithms are also implemented [13, 14, 15].

To assess the quality of methods that solve the regression problem, the MAPE metric was chosen. According to the results of such an assessment, ARIMA best describes the data.

The F-measure was chosen to assess the quality of methods that solve the classification problem. According to the results of such an assessment, the support vector machine with kernel = linear best describes the data.

REFERENCES

- [1] European Commission. Energy strategy. Energy Union, Available at <https://ec.europa.eu/energy/en/topics/energy-strategy/energy-union-0> (accessed 02/04/2022)
- [2] Renewable Energy Markets. Available at <https://www.energyweb.org/solutions/renewable-energy-markets/>, (accessed 24 Feb 2020)
- [3] European Commission. Fourth Report on the State of the Energy Union, Available at https://ec.europa.eu/commission/publications/4th-state-energy-union_en (accessed 13.01.2022)
- [4] Clean Energy for all Europeans Available at <https://op.europa.eu/en/publication-detail/-/publication/b4e46873-7528-11e9-9f05-01aa75ed71a1> (accessed: 20.11.2021)
- [5] Clean energy for all Europeans package completed: good for consumers, good for growth and jobs, and good for the planet Available at https://ec.europa.eu/info/news/clean-energy-all-europeans-package-completed-good-consumers-good-growth-and-jobs-and-good-planet-2019-may-22_en (accessed: 20.11.2021)
- [6] Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions. A policy framework for climate and energy in the period from 2020 to 2030 Available at <https://ec.europa.eu/transparency/regdoc/rep/1/2014/EN/1-2014-15-EN-F1-1.pdf> (accessed: 20.03.2022)
- [7] LO3 Energy, Innovations empowering communities through localized energy solutions, Available at <https://lo3energy.com/innovations/>, (accessed 1 Feb 2020)
- [8] Oracle Utilities C2M. Available at <http://www.oracle.com/us/industries/utilities/oracle-utilities-c2m-brief-4072957.pdf> (accessed: 20.03.2022)
- [9] Oracle Utilities Meter Data Management Documentation Available at https://docs.oracle.com/cd/F21810_01/index.htm (accessed: 20.03.2022)
- [10] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed.: Springer-Verlag, 2009. P. 746.
- [11] Larose T., Daniel T. Discovering knowledge in data: an introduction to data mining. N.Y.: John Wiley & Sons Limited. 2005
- [12] Y.Chakhchoukh, P.Panclatici, L.Mili.Electric Load Forecasting Based on Statistical Robust Methods. IEEE Transactions on Power Systems, vol 26, no. 3, 2011, pp.982-991
- [13] Marino DL, Amarasinghe K, Manic M (2016) Building energy load forecasting using deep neural networks. Ptoceedings 42nd Ann Conf IEEE Ind Electron Soc, IECON 2016, 2016,pp.7046–7051
- [14] Luo L, Hong T, Yue M (2018) Real-time anomaly detection for very short-term load forecasting. J Mod Power Syst Clean Energy, no. 6, 2018,pp. 235–243
- [15] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollar'. Learning to refine object segments. In European Conference on Computer Vision., Springer, 2016, pp. 75–91

Ontology Construction of City Hotline Service for Urban Grassroots Governance

Zuohai Chen
School of Computer Science and
Technology
Qilu University of Technology
Jinan, China
10431200678@stu.qlu.edu.cn

Heng Qian, Yongchao Gao
Shandong Computer Science
Center
Qilu University of Technology
Jinan, China
qianheng@seiti.cn
gaoyc@sieti.cn

Hongli Lyu
Department of Electrical and
Computer Engineering
Lakehead University
Thunder Bay, Canada
hlyu1@lakeheadu.ca

Qiuyue Wang
Shandong Standard Institute of
Emerging Technologies and
Innovations Ltd
Qilu University of Technology
Jinan, China
wangqy@sieti.cn

Abstract—In order to satisfy the current needs of urban grassroots governance, a novel city hotline service ontology is constructed to improve the efficiency and effectiveness of city hotline service in this paper. Firstly, the TOVE method is applied to construct government hotline service ontology for urban grassroots governance based on the cases of city hotline service. Secondly, the necessary concepts are defined, including the hierarchical relationships, attributes, and constraints among the seven modules of the city, personnel, organization, government, hotline appeal, hotline complaint object, and hotline processing results. Finally, the city hotline service ontology is built with the protégé ontology development tool and then tested in hotline scenarios of different urban grassroots governance. The testing results verify the validity and practicability of the proposed ontology so that the needs of hotline service are met in urban grassroots governance.

Keywords—city hotline service, urban grassroots governance, TOVE method, ontology, protégé

I. INTRODUCTION

Ontology plays an essential role in the standardized description and sharing of domain knowledge with the wide application of ontology. It can extract and integrate massive intelligence information to build a knowledge framework in the field of urban grassroots governance and then the intelligent governance of urban grassroots is realized [1]. With the increase of city hotline service data, it is necessary for the city hotline service ontology to standardize the city hotline data. As one of the essential styles of urban grassroots governance, the urban hotline service can be applied to accurately understand the needs of the public [2]. The problems that existed in the operation of the city can be fully mined and the hidden value behind its massive consultation, complaint data and other information data are able to improve the level of the urban grassroots governance and the processing efficiency of city hotlines [2].

Currently, many urban government performance management systems have been developed in a variety of cities all over the world, such as CitiStat, CompStat, and GMAP based on the 311 system [3]. Meanwhile, the 311 system has become the primary basis for the assessment and accountability of city government departments [3]. One of the main goals of the 311 system is to increase the accessibility of urban services. At the same time, the efficiency of the city's response to public inquiries is improved. The hotline information in the city has been summarized and the needs and problems of the public are organized in the 311 system.

In China, the “12345” citizen service hotline system is mainly used in the performance management system of the city government. In the Beijing “12345” citizen service hotline system, “public complaints have been processed without delay” has played a critical role as one working mechanism in the government’s performance appraisal [4]. Additionally, the “12345” citizen service hotline system can efficiently solve residents’ timely worries, and the improvement of governmental service performance is realized continuously [4]. There are a series of advantages for the “12345” citizen service hotline such as fast acceptance, quick transfer, and efficient settlement. Meanwhile, the “12345” citizen service hotline reduces the administrative cost of government departments and also satisfies the demands of the masses [5].

In order to strengthen the construction of city hotline services in China and other countries, it is necessary to simplify and speed up the communication process between governments and citizens so that many relevant problems can be solved or improved [2]. Compared to the 311 system with the “12345” citizen service hotline system, it is obvious that both are evaluations to handle efficiency of hotline service cases for urban grassroots governance. Hence, by completing the construction of the city hotline

service ontology, the processing efficiency of the city hotline can be significantly accelerated. At the same time, it can improve the performance appraisal of governments at all levels and the level of urban grassroots governance.

Currently, the government can only implement simple statistics on the city hotline data because of the complexity of the data in the city hotline system [2]. Correspondingly, the efficiency of the government in both solving citizens' problems and the level of urban grassroots governance are reduced. Hence, it is a challenging problem to conduct in-depth analysis in the information processing for the city hotline system. Based on the hotline service case of urban grassroots governance, the protégé tool is applied to construct the hotline service ontology for urban grassroots governance in this paper. A knowledge framework is proposed for efficient processing of urban hotlines for the low efficiency of existing urban hotlines. The city hotline service can provide the hotline staff with efficient processing methods so as to find out the critical information in the case, then the efficiency of handling the case will be improved, thereby the government's execution efficiency and citizens' satisfaction with the government can be increased finally.

II. PRELIMINARY KNOWLEDGE

A. Ontology Concept

Ontology is a distinct normative description and knowledge representation about a domain's hierarchical structure. And it is a system for people to understand the unified nature of concepts in a particular domain. The ontology conceptualizes the description objects and provides a uniform set for terms and relations in the domain [6]. The main target of ontology is to acquire knowledge in relevant fields, so that the general concepts are recognized and understood for the knowledge in this field and the relationship between these concepts is defined well. Finally, the semantics of concepts can be represented to clarify the interrelationships between these concepts [7].

B. Ontology Construction

The TOVE enterprise modeling method was proposed by Fox to construct a hotline service ontology for smart cities in 1994 [8] and it is firstly employed to the city hotline service systems by authors' team in this paper. The TOVE enterprise modeling method was not only constructed directly through the knowledge logic model in the form of ontology but also first used to establish the informal description [9]. Generally, the TOVE enterprise modeling procedure included the following steps: [10-11]

- Acquisition of motivational plots;
- The explicit expression of formal competency questions, the competency questions are based on

use cases requirements to test the capabilities of ontology information retrieval and question answering;

- Normalization of terms;
- Explicit description of formal competency questions;
- Formalization for the rules into axioms;
- Adjust the conditions for the solutions of the ability problem so that the ontology of knowledge tends to be complete.

Based on the above steps the modeling structure is shown in Fig. 1.

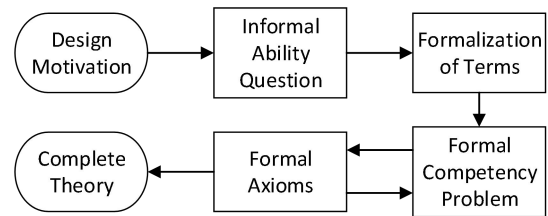


Fig. 1. TOVE enterprise modeling method flowchart.

Ontology Building Tool

Protégé is an open-source application tool based on the Java environment developed by Stanford University [6]. The ontology structure in Protégé is displayed in a tree-like hierarchy under which the classes, subclasses, attributes, instances, and relationships of the ontology can be either edited or extended, even deleted. There also exist a variety of plug-ins that can provide extensible markup languages (XML), resource description framework (RDF), web ontology language (OWL) and other storage formats [12].

III. CITY HOTLINE SERVICE CASES AND COMPETENCY QUESTIONS

The following four types are the major items accepted by the government hotline, which almost covered the main issues in urban grassroots governance. The city hotline service center will report the following problems mentioned to the relevant departments respectively, and the relevant departments will take actions to deal with them according to the follow-up investigation and then reply to the public one by one.

A. Transportation

The city hotline service center receives varieties of reports and complaints from citizens, mainly including traffic facility, traffic order, motor vehicle, and driver management, bus operation, rental management, operation

vehicle management, logistics management, and railway transportation, etc.

Competency questions:

- As of time t , how many complaints are related to transportation in city c ?
- From time a to time t , which one receives the most complaints hotline among the traffic facilities, traffic order, motor vehicle and driver management, rental management, operating vehicle management, logistics management, railway management, port and shipping and highway administration in the transportation of the city c ?
- From time a to time t , what problem does the citizen p reflect in the transportation of the city c ?
- On the moment t , what is the result of the city c accepting the citizen p 's handling of the taxi detour problem?

B. Social Security

The city hotline service center has received hotlines from citizens about urban social security, including public security management, case handling, fire management, household registration management, entry and exit management, network security, and mountain quarrying, etc.

Competency questions:

- From time a to time t , which have received the most complaints about dog raising, living noise, explosives, official seal engraving, and mass gatherings in city c ?
- On the moment t , what is the contact information of citizen p who complained about fighting and gambling in the city c ?
- As of time t , which district or county in the city c has received the most complaints about social security?
- On the moment t , what is the processing result after citizen p in the city c reports that the fire escape is blocked?

C. City Management

The city hotline service center has recently received hotlines from citizens about the urban appearance and environment, including municipal environment, landscape lighting, energy-saving insulation, and building enclosures.

Competency questions :

- On the moment t , what is the contact information of the citizen p who has built privately or arbitrarily in the cityscape environment of city c ?

- From time a to time t , which one received the most complaints of the city c 's city appearance environment, landscape lighting, energy-saving thermal insulation and building enclosure?
- As of time t , the city c receives comments from citizen p on energy conservation renovation and old residential areas. What is the result of the subsequent treatment?
- As of time t , has the relevant departments implemented the suggestion of citizen p to improve the brightness of landscape lights in the urban management of the city c ?

D. Urban Management Law Enforcement

The city hotline service center receives hotlines about the needs for urban management to enforce the law, mainly in municipal engineering, greening management, urban planning, civil air defense, and environmental protection management.

Competency questions:

- From time a to time t , which hotline has the most complaints about municipal engineering, greening management, environmental protection management, urban planning and civil air defense in urban management law enforcement in the city c ?
- As of time t , which have received the most complaints among the uncivilized road construction, unlicensed road excavation, unauthorized occupation of urban roads, and unauthorized establishment of gas storage stations in the municipal engineering of the city c ?
- On the moment t , citizen p reports illegal construction in urban planning and civil air defense in city c . Will the relevant departments handle it afterward?
- On the moment t , city q reflects the behavior of cutting down trees without authorization in city c . What is the contact information of the citizen p ?

IV. ONTOLOGY CONSTRUCTION OF CITY HOTLINE FIELD

ISO/IEC 5087 is a family of standards that define city data model. This standard defines the foundation level of the core concept model. It also addresses some of the core concept and cuts across the domain knowledge models. It will also define knowledge models for the services of citizen livelihood, urban management, and smart transportation illustrated [13-14].

The city hotline service ontology mainly adopted ISO/IEC 5087-1 and ISO/IEC 5087-2 city data models. The namespace prefixes used in the city hotline service ontology is shown in TABLE I.

TABLE I. NAMESPACE PREFIXES

Prefixes	URI
rdfs	http://www.w3.org/2000/01/rdf-schema#
geo	http://www.opengis.net/ont/geosparql#
xsd	http://www.w3.org/2001/XMLSchema#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
org	http://www.w3c.org/ns/org#
schema	http://schema.org/
activity	http://ontology.eil.utoronto.ca/5087/1/Activity/
city	http://ontology.eil.utoronto.ca/5087/2/City/
contact	http://ontology.eil.utoronto.ca/5087/2/Contact/
mereolog	http://ontology.eil.utoronto.ca/5087/1/Mereology/
y	
orgs	http://ontology.eil.utoronto.ca/5087/1/OrganizationStructure/
person	http://ontology.eil.utoronto.ca/5087/2/Person/
time	http://www.w3.org/2006/time#
spatialloc	http://ontology.eil.utoronto.ca/5087/1/SpatialLoc/
loc	http://ontology.eil.utoronto.ca/5087/1/SpatialLoc/

Pattern Introduction of City Hotline Service Ontology

A pattern is a set of concepts that are related to the topic and inter-connected by properties, thereby forming a graph. The city hotline service ontology is mainly composed of seven patterns: city, person, government, organization, complaint object, hotline appeal, and hotline processing result.

The city pattern mainly involves the name of the city, the administrative area, and the community under its jurisdiction.

The person pattern mainly covers all personnel information and geographic location information in the city.

Some of the objects that citizens call the hotline are shops and organizations, and the organization pattern systematically classifies these objects. The organization pattern mainly covers all the shops and some institutions in the city. Organizations are mainly divided into three categories, namely government organizations, for-profit organizations, and non-profit organizations. Government organizations are government units, for-profit organizations are some businesses and shops, and nonprofit organizations are mainly public welfare organizations such as the Red Cross.

The government pattern mainly contains implementing the public's complex problems and the release of some government announcements and suggestions during the epidemic management stage. The government mainly covers many functional departments, among which the city government mainly contains two parts, one is the municipal government, and the other is the district government.

The hotline appeal pattern mainly refers to whether the hotline calls made by citizens to the government about

complaints, suggestions, help, and inquiries are implemented and whether the citizens are satisfied.

The hotline complaint object pattern mainly introduces hotline complaint objects, including the person, organization, or phenomenon that the hotline complained about.

The functional departments of the government mainly apply the hotline processing result pattern to cope with the content of the hotline calls made by the citizens to make the citizens feel satisfied with the handling of the government, to avoid the problem of social security. After the city hotline service center is received from the citizens, the government department will implement specific measures, and there are also specific implementation situations. After the implementation, the city hotline service center also needs to reply to the citizens who call the hotline, give the government's solutions and then get the reply about whether they are satisfied with the implementation.

The overall pattern structure diagram of urban hotline service ontology is shown in Fig. 2 in the Appendix.

A. City Pattern

The city pattern of the city hotline service ontology mainly employs the city pattern in 5087, and the city pattern mainly contains “class city: City”, “class District” and “class Community” [13-14].

B. Person Pattern

The person pattern mainly utilizes the person pattern in 5087. The person pattern mainly contains “class person: Person”, “class contact: Address”, “class contact: AddressType”, “class contact: PhoneType” and “class person: Sex”. In order to meet the needs of hotline application scenarios, the city hotline service ontology adds *hasRoad* and *hasBridge* to the “class contact: Address” [13-14].

Address: The following properties have been added.

- *hasRoad*: The address specifies the name of the road.
- *hasBridge*: The address specifies the name of the bridge.

C. Organization Pattern

The organization pattern mainly reuses the organization pattern in 5087, and the organization pattern mainly contains the “class org: Organization”, “class org: ForProfitOrganization”, “class org: NonProfitOrganization” and “class org: GovernmentOrganization” [13-14].

D. Government Pattern

The government is a special kind of organization, and the government pattern mainly contains “class MunicipalGovernment” and “class DistrictGovernment”. The “class MunicipalGovernment” and “class

DistrictGovernment” are subclasses of the class “org: Organization”, corresponding to municipal and district governments, respectively. Since the government is the main party of the city hotline service, a separate pattern is set up [13-14].

MunicipalGovernment: The following properties have been added.

- Schema: name: The name of the municipal government.
- Spatialloc: hasLocation: The city where the municipal government unit is located.
- Mereology: hasProperPart: Districts and counties under the jurisdiction of the city government.

DistrictGovernment: The following properties have been added

- Schema: name: The name of the district and county government.
- Spatialloc: hasLocation: The district and county of the city where the district government is located.

The key classes, attributes and constraints in the government pattern of city hotline service ontology are shown in TABLE II.

TABLE II. GOVERNMENT PATTERN

Class	Property	Value Restriction
MunicipalGovernment	rdfs: subclassOf	org: Organization
	schema: name	exactly 1 xsd: string
	spatialloc: hasLocation	exactly 1 geo: Feature
	mereology: hasProperPart	some DistrictGovernment
	rdfs: subclassOf	org: Organization
DistrictGovernment	schema: name	exactly 1 xsd: string
	spatialloc: hasLocation	exactly 1 geo: Feature
	rdfs: subclassOf	org: Organization

Hotline Appeal Pattern

The hotline appeal pattern mainly contains “class HotlineEvent” and “class Complainant”. Among them, the “class HotlineEvent” describes the specific questions and difficulties encountered by the questioning masses. The “class Complainant” is a hotline appeals officer [13-14].

HotlineEvent: The following properties have been added.

- hasOccurDate: When the event occurred in the citizen dialing the hotline.
- hasCallDate: The time when the citizen called.
- hasComplainant: The citizen who called the hotline.
- hasOccurLocation: The location of the incident in the hotline that the citizen calls.
- hasHotlineEventType: The type of incident that the citizen calls the hotline to complain about.

- hasContent: The event's content is mentioned in the citizen's call to the hotline.
- asComplainedObject: The object that the citizen calls the hotline to complain about.
- associatedHotlineProcess: The associated hotline processing process.
- hasComplaintTelephone: The telephone number for citizens to call the hotline.

Complainant: The following properties have been added.

- associatedHotlineEvent: The associated hotline appeal event.
- hasComplainantName: The name of the citizen who called the hotline.
- hasComplainantSex: The gender of the citizen calling the hotline.

The hotline apple pattern structure diagram is shown in Fig. 3 in the Appendix. The key classes, attributes and constraints in the hotline appeal pattern of city hotline service ontology are shown in the following TABLE III.

TABLE III. HOTLINE APPEAL PATTERN

Class	Property	Value Restriction
HotlineEvent	rdfs: subclassOf	activity: Activity
	hasOccurDate	exactly 1 time: Instant
	hasCallDate	exactly 1 time: Instant
	hasComplainant	only Complainant
	hasOccurLocation	only loc: Feature
	HotlineEventType	exactly 1 xsd: string
	hasContent	exactly 1 xsd: string
	hasComplainedObject	some ComplainedObject
	associatedHotlineProcess	only HotlineProcess
	hasComplaintTelephone	only xsd: string
Complainant	rdfs: subclassOf	orgs: Role
	associatedHotlineEvent	only HotlineEvent
	ComplainantName	exactly 1 xsd: string
	ComplainantSex	exactly 1 Sex

Hotline Complaint Object Pattern

The hotline complaint object pattern mainly contains “class ComplainedObject”, “class ComplainedPerson”, “class ComplainedOrganization”, and “class Complained-Phenomenon”. The “class ComplainedPerson”, “class ComplainedOrganization”, and “class ComplainedPhenomenon” are subclasses of the “class ComplainedObject”. The “class ComplainedPerson” indicates that the complained object is a person; the “class Complained-Organization” denotes that the complained object is an organization; and the “class ComplainedPhenomenon” represents that the complained object is a phenomenon [13-14].

ComplainedObject: The following properties have been added.

- associatedHotlineEvent: the associated hotline appeal event.
- hasComplainedObjectAddress: The address of the complained object

ComplainedPerson: The following properties have been added.

- hasComplainedPersonName: The name of the complainant.
- hasComplainedPersonTelephone: The phone number of the complainant.
- hasComplainedPersonSex: The gender of the complainant.

ComplainedOrganization: The following properties have been added.

- hasComplainedOrganizationName: The name of the organization being complained about.
- hasComplainedOrganizationTelephone: The phone number of the complained organization.

ComplainedPhenomenon: The following properties have been added.

- hasComplainedPhenomenonName: The name of the phenomenon being complained about.

The Hotline complaint object pattern structure diagram is shown in Fig. 4 in the Appendix.

The key classes, attributes and constraints in the complaint object pattern of city hotline service ontology are shown in the following TABLE IV.

TABLE IV. COMPLAINT OBJECT PATTERN

Class	Property	Value Restriction
ComplainedObject	associatedHotlineEvent	HotlineEvent
	contact:hasAddress	only contact: Address
ComplainedPerson	rdfs: subclassOf	orgs: Role
	rdfs: subclassOf	ComplainedObject
	hasComplainedPersonName	exactly 1 xsd: string
	hasComplainedPersonTelephone	only xsd: string
	hasComplainedPersonSex	exactly 1 Sex
ComplainedOrganization	rdfs: subclassOf	org: Organization
	rdfs: subclassOf	ComplainedObject
	hasComplainedOrganizationName	exactly 1 xsd: string
	hasComplainedOrganizationTelephone	only xsd: string
ComplainedPhenomenon	rdfs: subclassOf	activity: Activity
	rdfs: subclassOf	ComplainedObject

	hasComplainedPhenomenon	exactly 1 xsd: string
--	-------------------------	-----------------------

E. Hotline Processing Result Pattern

The hotline processing result pattern mainly contains “class HotlineProcess”, “class HotlineServiceStaff”, and “class Hotline-Replier”. Among them, the “class HotlineProcess” is the process content of the relevant departments to solve the problems raised by the masses. The “class HotlineServiceStaff” is the hotline operator, and the “class HotlineReplier” is the responder after processing the hotline event [13-14].

HotlineProcess: The following properties have been added

- hasStartDate: The time the hotline event started processing.
- hasFinishDate: The event that the hotline event processing ends.
- hasWay: The way the hotline event is handled.
- hasResult: The result of the hotline event processing.
- hasFeedbackTelephone: The telephone number of the hotline processing result feedback.
- associatedHotlineEvent: The associated hotline appeal event.
- hasHotlineServiceStaff: The operator for hotline handling.
- hasHotlineReplier: The responder of the hotline processing result.
- hasEventHandler: The handler of the hotline event.

HotlineServiceStaff: The following properties have been added

- associatedProcess: The process of the associated hotline handling.
- associatedHotlineEvent: The associated hotline appeal event.
- worksAt: The organization the operator works for.

HotlineReplier: The following properties have been added

- associatedProcess: The process of the associated hotline processing.
- associatedHotlineEvent: The associated hotline appeal event.
- hasPost: The hotline responds to the person's post.
- worksAt: The organization the hotline responder works for.

The hotline complaint processing result pattern structure is shown in Fig. 5 in the Appendix. The key classes, attributes and constraints in the hotline processing result pattern of city hotline service ontology are shown in the following TABLE V under the next page.

V. VERIFICATION OF CITY HOTLINE SERVICE ONTOLOGY

The verification procedure of the city hotline service ontology focuses three major steps: instantiate the ontology in the protégé; store the instantiated data in the GraphDB database; and verify the accuracy and feasibility of the ontology through the SPARQL statement. As an example, the main functions based on the city hotline service data during the epidemic and the corresponding cases are selected as follows for verification in combination with the four directions in Section III.

Class	Property	Value Restriction
HotlineProcess	rdfs: subclassOf	activity: Activity
	hasStratDate	exactly 1 time: Instant
	hasFinishDate	exactly 1 time: Instant
	hasWay	exactly 1 xsd: string
	hasResult	exactly 1 xsd: string
	hasFeedbackTelephone	only xsd: string
	associatedHotlineEvent	only HotlineEvent
	hasHotlineServiceStaff	only HotlineServiceStaff
	hasHotlineReplier	only HotlineReplier
hasEventHandler	some (org: GovernmentOrganization or org: ForProfitOrganization or org: NonProfitOrganization)	
HotlineServiceStaff	rdfs: subclassOf	org: Employee
	associatedHotlineProcess	only HotlineProcess
	worksAt	some Organization
HotlineReplier	rdfs: subclassOf	orgs: Role
	associatedHotlineProcess	only HotlineProcess
	org: hasPost	only Post
	worksAt	some Organization

A. Follow Closely

The city hotline service center received a call from one citizen who found a Hubei-registered vehicle in the city, and requested the relevant departments to investigate the vehicle's itinerary and the owner information. The city hotline service center reported the citizen's questions and suggestions to the relevant departments for implementation and gave a personnel reply by phone.

Competency question:

- From 13th to 18th in January, how many citizens in the city Jinan reported the discovery of Hubei-registered vehicles through the hotline?

Answer:

2314.

B. Government Epidemic Prevention Consultation

The city hotline service center received complaints from citizens that the publicity of epidemic prevention and control near their communities was not adequate, and that outsiders encountered problems when they came to the city. The city hotline service center will report the above-mentioned problems to the relevant departments, and the relevant departments will deal with them according to the investigation and reply the results to the citizens.

Competency question:

- From 17th to 24th in January, how many people come to the city Jinan?

Answer:

2658.

C. Return to Work and School

The city hotline service center received a request to delay the start of work from citizens because in Jinan there are new patients increased recently. At the same time, another request is to ask students when school can start normally. In addition, there are some off-campus remedial classes to be arranged to start school early, but it is also hoped to delay the starting of the school because of the current serious epidemic. On the other hand, there is no protective measure to be provided after the company starts work, and there are many people who start work but not comply with the epidemic control policy. The city hotline service center will report these problems to the relevant departments, and the relevant departments will implement them after investigation and reply the results to the citizens.

Competency question:

- From 22nd to 26th in January, how many citizens in city Jinan want to delay the start of construction?

Answer:

1497.

D. Market Supervision

The city hotline service center received complaints from citizens about the short supply of masks. Currently, almost every pharmacy has no masks to sell while problems appeared in some private stores such as driving up the price of masks arbitrarily, and so on. At the same time, some vegetable markets and supermarkets are driving up some necessities of life. The city hotline service center will report these problems to the relevant departments, and the relevant departments will immediately investigate the problems and implement them, and then reply to the public on the implementation situation.

Competency question:

- From 23rd to 25th in January, how many people are calling to complain about the shortage of masks in city Jinan?

Answer:

4896.

VI. CONCLUSION

The modelling tools protégé and the TOVE enterprise modelling method are applied to construct a hotline service ontology for urban grassroots governance in this work, and the ontology is firstly stored in the GraphDB graph database by instantiating the urban hotline data in different scenarios. Then, the SPQRQL statements are used to verify the validity and practicality of the ontology. Finally, the urban grassroots governance level and the processing efficiency of urban hotline are improved by the hotline service body of urban grassroots governance with increasingly the processing efficiency of urban hotline data.

ACKNOWLEDGMENT

Thanks to the Shandong Foreign Experts Program (WSG2020020) for supporting this work.

REFERENCES

- [1] Q. Yang, and Y. Chen. "A survey of the ontology methodology," *Appl. Res. Comput.* Vol. 4, pp. 5-7, 2002.
- [2] C. Zhuang, "Research on government hotline construction from the perspective of service-oriented government-based on the 12345 government hotline of shuyang county," master's thesis from Nanjing University, pp. 2-13, 2018.
- [3] L. Zhu, "The Enlightenment of American Urban Management Informatization Construction to China-Take CompStat, CitiStat and GMAP as examples (in Chinese)," *E-Gov.* vol. 9, pp. 120-125, 2008.
- [4] C. Wang, and M. Liang, "How does performance feedback impact performance improvement-evidence" from the case of "Sue for Action" in Beijing. *Chn. Pub. Admin.* Vol. 11, pp. 117-125, 2020.
- [5] L. Ma, "Data-driven and People-oriented Government Performance Management: Case Study Based on Beijing's "Action on Request"," *Exp. Hz.* Vol. 2, pp. 50-55, 2021.
- [6] Z. Huang, "Use protégé to construct the new media domain ontology," master's thesis from Wuhan University of Technology, pp. 4-10, 2013.
- [7] Y. Shi. "Research on domain-oriented semantic search," master's thesis from Guangxi Normal University, pp. 14-19, 2010.
- [8] K. D. Tham, M. S. Fox, and M. Gruninger, "A cost ontology forenterprise modeling department of industrial engineering," *Infrastructure Collaborative Enterprises, WV-Workshop on Enabling Tech. USA*, pp. 111-117, *Proceedings 3rd Workshop on Enabling Tech. Morgantown*, 1994.
- [9] L. Yue, and W. Liu. "A comparative study of the construction method of domain ontology at home and abroad," *Intel. Theor. Prac.* Vol. 39, no. 8, pp. 119-125, 2016.
- [10] X. Fan, and C. Shi, "A survey on constructing ontology," *Ship. Elec. Engr.* Vol. 6, no. 53, pp. 15-18, 2011.
- [11] Y. Liu, "Research of approaches and development tools in constructing ontology," *Mod. Intel.* Vol.9, pp. 17-24, 2009.
- [12] A. Boustil, R. Maamri, and Z. Sahnoun, "A semantic selection approach for composite Web services using OWL-DL and rules," *Serv. Oriented Comput. Appl.* Vol. 8, no. 3, pp. 221-238, 2014.
- [13] ISO/IEC CD 5087-1. *Information technology – City Data Model – Part 1: Foundation Level Concepts*, 2022.
- [14] ISO/IEC CD 5087-2. *Information technology – City Data Model – Part 2: City Level Concepts*, 2022.

APPENDIX

In the following figures Fig. 2-5, the meaning represented by the shape in the figures, the square represents the entity class in the ontology; the ellipse represents the attribute of the entity; \longrightarrow Represents the relationship between entity classes and entity classes or between entity classes and attributes; \dashrightarrow Represents the affiliation between entity classes and entity classes.

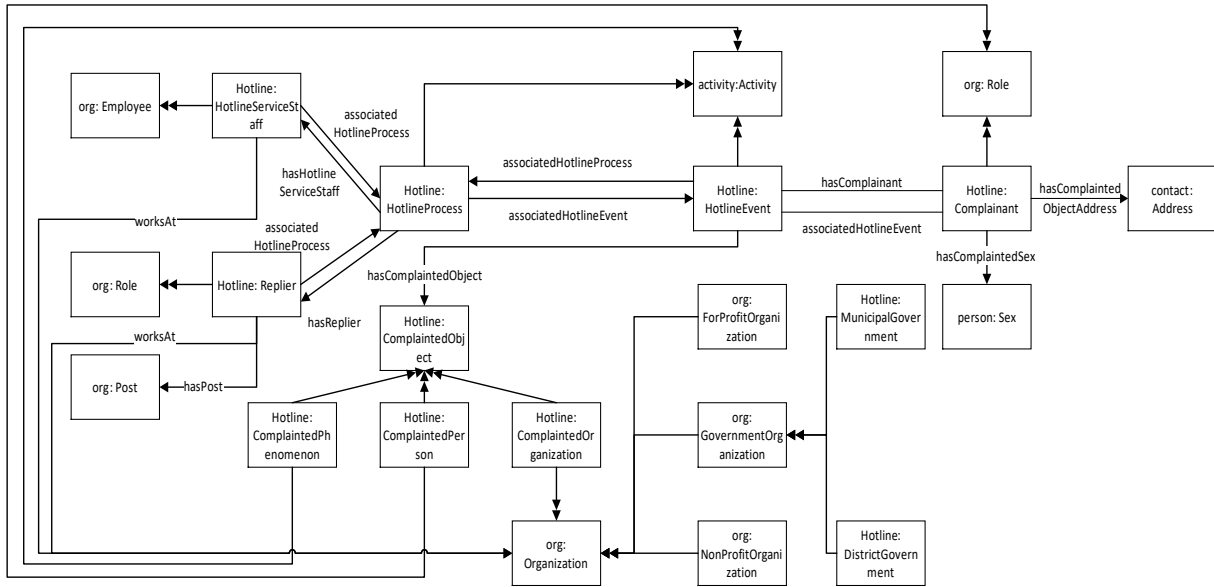


Fig. 2. Overall pattern structure diagram of city hotline service ontology.

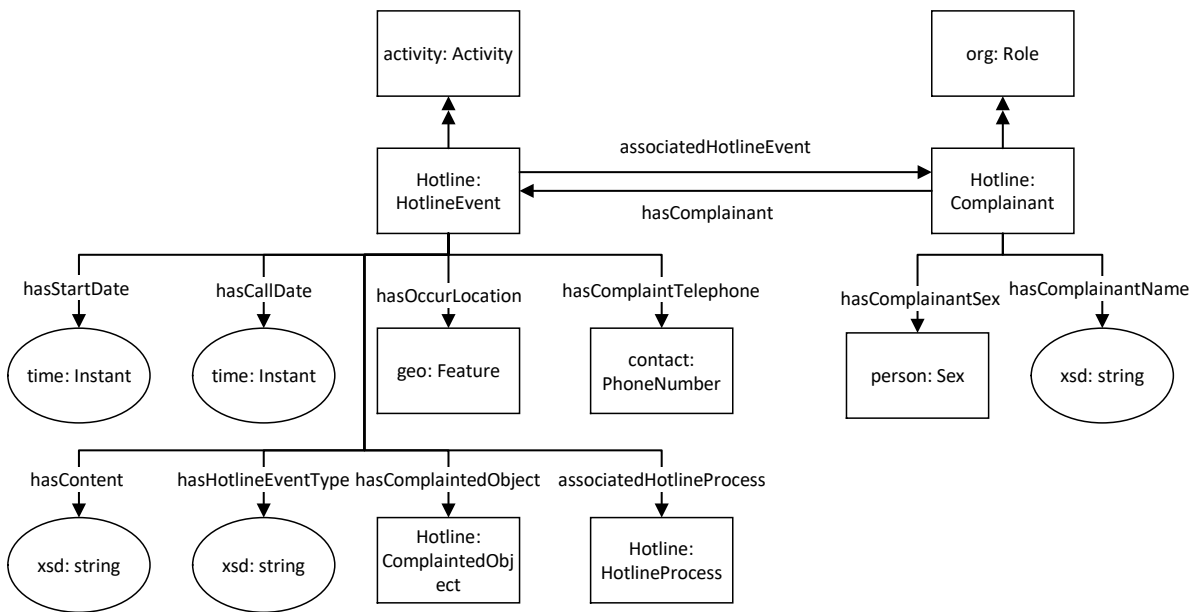


Fig. 3. Hotline appeal pattern structure diagram.

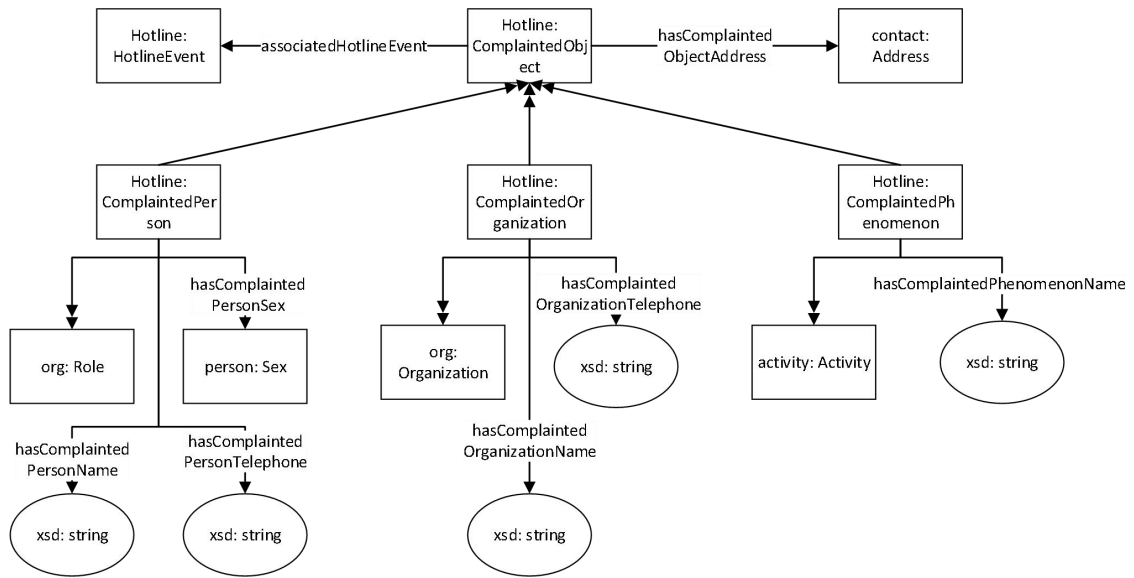


Fig. 4. Hotline complaint object pattern structure diagram.

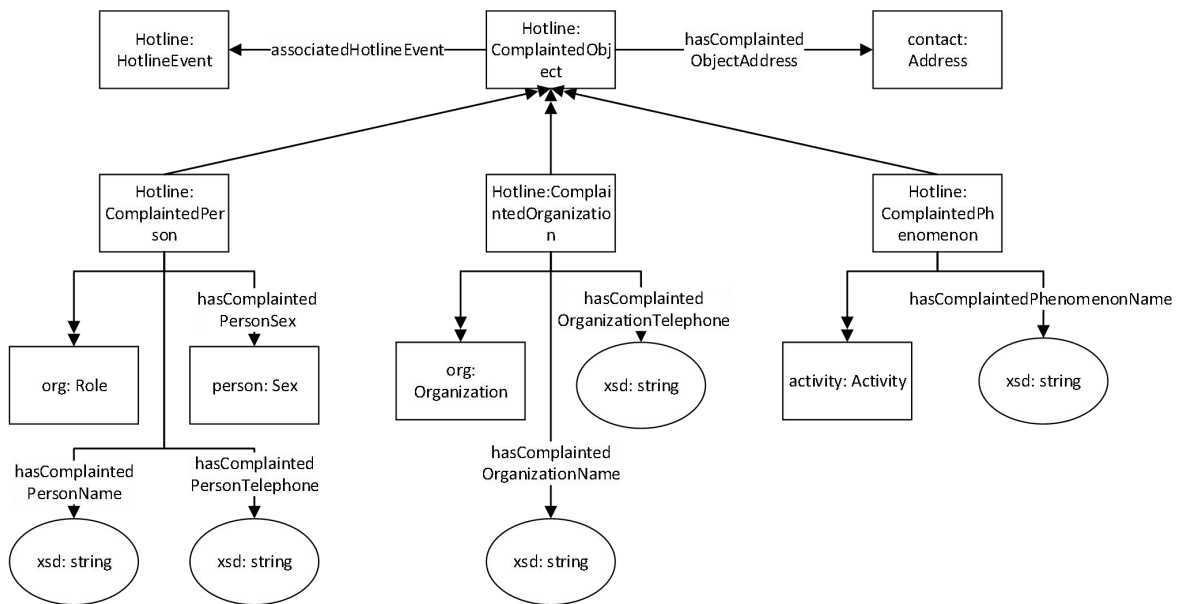


Fig. 5. Hotline processing result pattern structure diagram.

Leg Geometry Optimization of Thermoelectric Cooler to Maximize COP Through Gaussian Process Modelling

Ethan Robyn V. Ebuena
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
ethanrobyn.ebuena@ust.edu.ph

Jowen Louis Francisco
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
jowenlouis.francisco.eng@ust.edu.ph

La Vern Ramir Certeza
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
ldcerteza@ust.edu.ph

Carl Vincent Villanueva
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
carlvincent.villanueva.eng@ust.edu.ph

Johannes Kurt Tecson
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
johanneskurt.tecson.eng@ust.edu.ph

Jomar Lord Cauton
Department of Mechanical Engineering
University of Santo Tomas
Manila, Philippines
jomarlord.cauton.eng@ust.edu.ph

Abstract— Various researchers have studied how to increase the performance of thermoelectric cooling devices by optimizing their module design. In the present study, the researchers investigate whether a truncated square pyramid leg geometry results to a higher COP than a symmetrical rectangular leg by varying the middle and contact areas. The values of the input parameters for the thermal-electric analysis were generated using a sphere packing design. Then, the different leg geometries were modelled in FUSION 360 and simulated in ANSYS to determine the COP from the resulting hot and cold side temperatures. Afterward, a mathematical model was generated using Gaussian Process modelling to model COP as a function of the middle and contact areas. By maximizing the desirability function of the model, results show that for the experimental design space considered for this study (contact area: $0.108 \text{ mm}^2 < AC < 1.000 \text{ mm}^2$; middle area: $1.150 \text{ mm}^2 < AM < 2.560 \text{ mm}^2$), the highest COP was attained at $AC = 1.000 \text{ mm}^2$ and $AM = 1.150 \text{ mm}^2$. The researchers therefore conclude that an asymmetrical truncated square pyramid leg geometry yields a higher COP compared to a symmetrical rectangular leg geometry, which is consistent with the results of past studies.

Keywords— *thermoelectric cooling, space-filling design, Gaussian process modelling, optimization*

I. INTRODUCTION

Significant consensus has already been established within the scientific community regarding the degree of impact that anthropogenic carbon emissions have on global warming and climate change. A significant step towards decarbonization was the adoption of the Montreal Protocol in 1987 in which signatory nations pledge to phase out the manufacture of chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HCFCs) and to phase down the use of hydrofluorocarbons (HFCs). These chemicals are widely used in refrigeration and air-conditioning systems. Although HFCs have a lower ozone depleting potential compared to CFCs and HCFCs, they still unfortunately have a considerably high global warming potential. Therefore, efforts must be made to make cooling systems that are free from emissions which harm the environment and the ecosystem.

Thermoelectric cooling (TEC) systems are a viable substitute to cooling systems that run on vapor compression cycles. TEC devices operate on the principle of Peltier effect. This refers to the phenomenon in which a temperature differential is induced between two ends of a thermocouple wire, consisting of two different semiconductor materials, when electrical current is allowed to pass through it [1].

Benefits of using TEC systems over the conventional cooling systems include its compactness, high reliability, and the absence of moving parts and working fluids. In addition, these can also be powered by direct current sources and can easily be switched from cooling to heating modes. However, its major drawbacks are its high cost and low energy conversion efficiency. These limitations restrict the use of TEC devices in aerospace, waste heat recovery, and electronic cooling applications [2], where reliable and quiet operation is more important than affordability and high efficiency [3].

Various studies have been published in the recent decades with the objective of improving the performance of TEC devices. Attempts to increase the coefficient of performance of TEC devices revolve around discovering novel materials for Peltier modules, optimizing the module design and fabrication, and analyzing the overall Peltier cooling system and heat exchange efficiency [4].

Pourkiaei et al. [5] produced a comprehensive review of materials that are currently used in the manufacture of thermoelectric modules. They grouped these materials into semiconductors, ceramics, and polymers, then characterized these materials based on important thermoelectric properties such as Seebeck coefficient, electrical resistance, and thermal conductivity. This review study was able to provide a list of promising materials for thermoelectric module fabrication with their corresponding figures of merit for certain values of operating temperatures.

In addition to the TEC material, the thermoelectric module geometry was also the subject of improvement in various studies. Shittu et al.'s work [6] presents the most current review of the studies that have been published so far that are related to thermoelectric geometry and structure optimization.

Based on this comprehensive review, the four main geometrical parameters that are the subject of most optimization studies are leg length or height, cross-sectional area, number of legs, and leg shape. In addition, several optimization methods, such as three-dimensional finite optimization and multi-objective optimization, were cited as the main optimization routines used to improve TEC efficiency with respect to the module geometry.

This research intends to add to the growing body of literature that attempts to address the question on how to determine the optimum thermoelectric leg shape and geometry to attain the maximum possible COP. For context, several studies have already been published in this regard. Lamba et al. [7] used genetic algorithm to optimize the cooling capacity and coefficient of performance (COP) of a trapezoidal thermoelectric leg. The choice of trapezoidal leg geometry was anchored on the study of Sahin and Yilbas [8] which found that the COP increases with increase in the shape parameter (ratio of the cross-sectional area at the cold side to that at the hot side). The variable parameters considered in the optimization were the temperature dependent thermoelectric properties, shape parameter, and the temperature ratio (ratio of the hot side temperature to the cold side temperature). In addition, Lin and Yu's [9] study proved using numerical simulation that the trapezoid-type two-stage Peltier couple reduced the thermal resistance compared to the standard geometry of TEC, increases heat rejection on the hot side due to the larger cross-sectional area of the trapezoid-shaped thermoelectric legs, and enables Joule heat conduction to lean towards the hot side due to the asymmetrical thermal resistance. The results of these studies are consistent with the findings of Fabian-Mijangos et al.'s research [10] on the performance of thermoelectric modules with asymmetrical legs. This is one of the first studies that provided experimental validation confirming that pyramidal legs aid in decreasing the overall thermal conductance of a TEC device, thereby increasing the thermal gradient in the legs and harnessing the Thomson effect, which is generally neglected in symmetrical rectangular thermoelectric legs.

The main objective of this study is to optimize the leg geometry of a semiconductor thermoelectric cooling Peltier module with respect to the cross-sectional areas at the contacts (A_C) and at the middle (A_M) to maximize the coefficient of performance. The following are the sub-objectives of this study:

- To determine the allowable range of working current for a proposed TEC module with truncated square pyramid leg geometry using ANSYS
- To generate the values of the input parameters using sphere packing design
- To perform thermal-electric analysis of the truncated square pyramid leg geometries and conventional TEC geometry
- To determine the values of the hot and cold side temperatures for all modelled leg geometries and compute the corresponding COP

- To generate a Gaussian Process model to fit the response variable (COP) as a function of the contact area and middle area
- To confirm the goodness-of-fit of a Gaussian Process model by analyzing model diagnostics in JMP®
- To determine the optimal cross-sectional area of the middle and contact areas of the proposed truncated square pyramid geometries and the corresponding maximized COP value
- To determine whether the optimized truncated square pyramid leg geometry has an increase in COP compared to a conventional TEC-12706 module.

The optimization of the TEC module will be focused on the semiconductor leg geometry. For this study, Bismuth Telluride (Bi_2Te_3) will be used as the semiconductor material due to its applications in various thermoelectric modules [11][12]. Material properties such as the Seebeck coefficient, thermal conductivity, electrical resistivity, and current range will be set as control variables of the study to focus on the optimization of the modified leg cross-sectional areas to maximize the COP. Moreover, a single thermoelectric thermocouple will be modelled in the simulation. The truncated square pyramid leg design will be tested using thermal-electric heat analysis through ANSYS 2021 Software. The final optimized model was not fabricated during this study.

A segmented single thermocouple analysis was performed on the thermoelectric leg after the material's dimensions had been determined. The areas of the leg were assigned as A_C and A_M . The contact surfaces of the semiconductor legs to the copper plate were denoted as A_C , and the middle area of the legs was denoted as A_M . The dimensions of the contact areas of the top and bottom parts of the legs were set as equal. Furthermore, the same geometry configuration was used for both the p-type and n-type semiconductor legs.

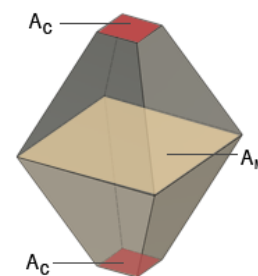


Fig. 1. Truncated square pyramid leg design with its indicated cross-sectional area assignments.

II. METHODOLOGY

The study consisted of six phases: space filling design, determination of the working current range, FUSION 360 modelling, ANSYS thermal-electric analysis, COP computation, and statistical analysis. Figure 2 represents the research procedure performed in the study.

The first step of the study was establishing a design space for A_C and A_M to be investigated. The working current was determined for all the samples using the minimum and maximum values of the design space through limit testing of simulated TEC working temperatures. A range of working current was established to ensure that the simulation results were consistent and ensure that the parameters do not exceed the suitable operating conditions of the thermoelectric leg. Simultaneously, Sphere Packing was used as the space filling design method using JMP software to obtain the working sample dimensions. The generated dimensions from the Sphere Packing were modeled in the Autodesk FUSION 360 software, which was used as the model for the simulation. The thermal and electrical properties of the thermoelectric pair together with its physical model were simulated in ANSYS using the finite element approach. A thermal-electric simulation system was used to simulate the thermoelectric conditions for the different samples. After the simulation, the highest COP gathered from the different samples was used for Gaussian Process modeling in the JMP software to define the COP characteristics of the design space with respect to the change in cross-sectional areas of the modelled legs and obtain the optimized truncated square pyramid leg geometry. Only one Gaussian process model was generated during the research. Afterward, the TEC-12706 was simulated with the same current range, and its highest COP will be compared to the maximized COP of the modified TEC.

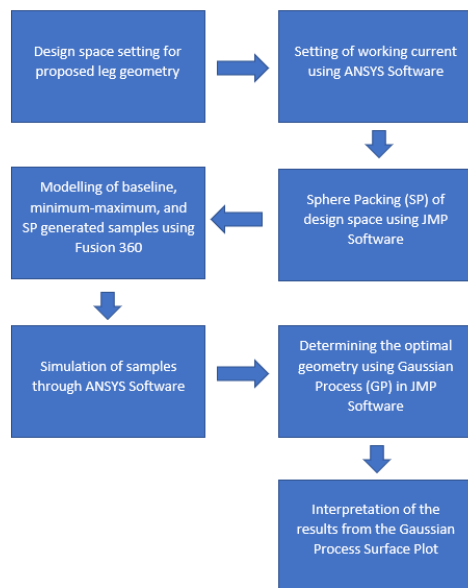


Fig. 2. Research procedure employed during the study.

A. Current range

The allowable working current range was determined by testing the minimum and maximum contact and middle areas through ANSYS thermal-electric analysis. The allowable working current range for all runs was based on the current range that would not result to an operating temperature of 250 °C. This was set as the allowable operating temperature

because bismuth telluride undergoes phase change upon reaching temperatures of 250 to 300°C [13].

The assumed material properties used for the simulation are presented in Table 1. Current was induced in the n-type copper plate, and a ground voltage was set in the p-type copper plate. The assumed values of the other simulation parameters are given in Table 2.

TABLE I. MATERIAL PROPERTIES FOR ANSYS SIMULATION

	Seebeck Coefficient $\frac{V}{K}$	Thermal Conductivity $\left(\frac{W}{m \cdot K}\right)$	Resistivity (Ωm)
Copper alloy	-	401	ANSYS table values
n-type Bi ₂ Te ₃	-1.65x10 ⁻⁴	1.285	1.05x10 ⁻⁵
p-type Bi ₂ Te ₃	2.1x10 ⁻⁴	1.285	9.8x10 ⁻⁶
Ceramic plate	-	13.74	3.162x10 ¹³

TABLE II. VALUES OF PARAMETERS FOR ANSYS SIMULATION

Convection film coefficient	1x10 ⁻⁸ $\frac{W}{m^2 \cdot ^\circ C}$
Ambient temperature	32°C
Cold side temperature	20°C
Current	0A (ramped)
Ground voltage	0V

The working current range was then ramped from 0A to the highest allowable current in the ANSYS software.

B. Sphere packing space filling

Space filling designs are often used in computer-generated experiments due to its ability to spread design points evenly across the design space [14]. For this study, a sphere packing space filling design using JMP® statistical analysis software was used to determine the values of the variable input parameters. Sphere packing maximizes the minimum distance between pairs of design points in the experimental region. Sphere packing also favors areas near the edge of the design space, then plots points within the middle to maximize the minimum distance between points.

TABLE III. INPUT PARAMETERS OF A_C AND A_M IN SPHERE PACKING.

	A_C (mm ²)	A_M (mm ²)
Minimum	0.108	1.150
Maximum	1.000	2.560

The minimum and maximum values of the contact cross-sectional area (A_C) and the middle cross-sectional area (A_M) represents the design space used in the study. These values were used for the sphere packing model, for a total of 10 runs. The values inputted in the JMP software for sphere packing design was summarized in Table 3.

C. FUSION 360 model

FUSION 360 was used as the 3D-modelling software in the design of the thermoelectric leg segments. The dimensions of the model were based on the actual TEC-12706 thermoelectric module and were equally segmented to show a per pair measurement. Table 4 shows the measured dimensions for the segmented thermoelectric pair.

TABLE IV. DIMENSIONS OF THE SEGMENTED THERMOELECTRIC PAIR.

	Dimensions (L × W × H)
Ceramic Plate	40.1 mm × 40.02 mm × 1mm
Copper Plate (Larger)	4 mm × 1.6 mm × 1mm
Copper Plate (Smaller)	2mm × 1.6mm × 0.33mm
Leg Length	1.7mm

D. ANSYS thermal-electric analysis

The modelled thermoelectric leg segments from FUSION 360 were exported to ANSYS for thermal-electric analysis. Similarly, the material properties and simulation parameters in Tables 1 and 2 were used in the simulation. Each thermoelectric leg was tested separately, where the hot side and cold side temperatures across the working current were used for the computation of the COP.

E. COP calculation

The COP of the segmented thermoelectric device is derived from its cooling load, Q_c , and the electrical power required, P , which was determined by using the COP equation 1:

$$COP = \frac{Q_c}{P} = \frac{(\alpha_p - \alpha_n)T_c I - 0.5I^2(R_p + R_n) - (K_p + K_n)(T_h - T_c)}{(\alpha_p - \alpha_n)I(T_h - T_c) + I^2(R_p + R_n)} \quad (1)$$

where, α is the Seebeck coefficient (α_p for p-type, α_n for n-type), T is the temperature of the thermoelectric pair (T_c for cold-side temperature, T_h for hot-side temperature), R is the resistivity of the material (R_p for p-type, R_n for n-type), and K is the conductance of the material (K_p for p-type, K_n for n-type).

The resistivity was computed using the equation 2:

$$R = \rho \frac{L}{A_{ave}} \quad (2)$$

and the conductance was computed using the equation 3:

$$K = \sigma \frac{A_{ave}}{L} \quad (3)$$

The COP for the TEC-12706 was determined through equation 4.

$$COP_{TEC12706} = \frac{ZT_c^2(T_h - T_c)}{ZT_h T_c} \quad (4)$$

Z in equation 4 can be found using equation 5.

$$Z = \frac{(\alpha_p - \alpha_n)^2}{\{(K_p + K_n)(R_p + R_n)\}} \quad (5)$$

COP formulas used in the study were gathered from [1].

F. Statistical Analysis

JMP® statistical analysis software was used for the data analysis of the study. JMP is a statistical tool used to analyze data tables in design of experiments (DOE). A specialized model through Gaussian process modelling was performed to model the results and was used for optimization.

1) Model validation

An actual by predicted plot was used for regression analysis to determine the fit of the modelled data. The actual by predicted plot shows the data points of the actual response COP in the x-axis to the jackknife predicted values of the y-axis. Points lying closer to the 45° regression line indicate better fit model.

2) Gaussian process modelling

Gaussian process models are commonly used to model the response of computer-generated experiments [14]. Moreover, the Gaussian process is a simple and effective method to observe the behavior of different parameters while effectively measure the prediction uncertainty of the model [15].

The highest COP for each run was used to model the results through the Gaussian process model. A marginal model plot was generated to visualize the relationship of each factor to the response. Each factor has a marginal model plot that visualizes the predicted values [16][17]. Furthermore, a model report table that describes the main effect and interaction between each variable was generated. Lastly, a surface plot for the relationship of the response COP to the changes in A_C and A_M was generated.

3) Desirability factor

The optimal cross-sectional areas A_C and A_M that yield the maximum COP were determined by maximizing the desirability function of the response COP in the prediction profiler of the JMP software. Desirability is the estimated response value that lies between 0 to 1, where values closer to one indicate the optimal performance of factors within the design space [18].

4) COP comparison

The determined COP through the desirability function was compared to the COP of the TEC-12706 simulated model. The leg geometry parameters A_C and A_M were also compared to each other.

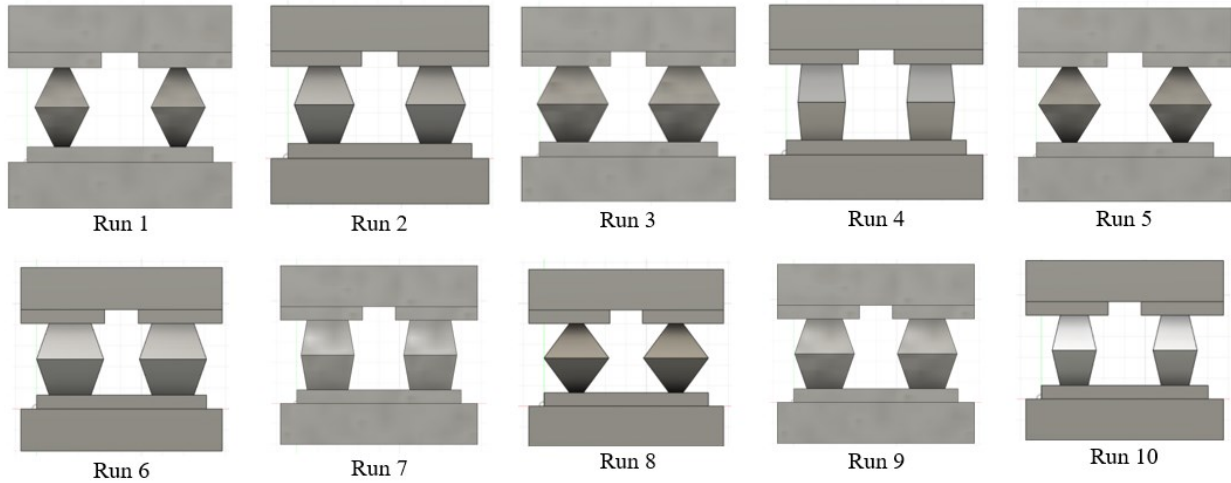


Fig. 3. Leg geometry designs of the 10 sphere packing runs.

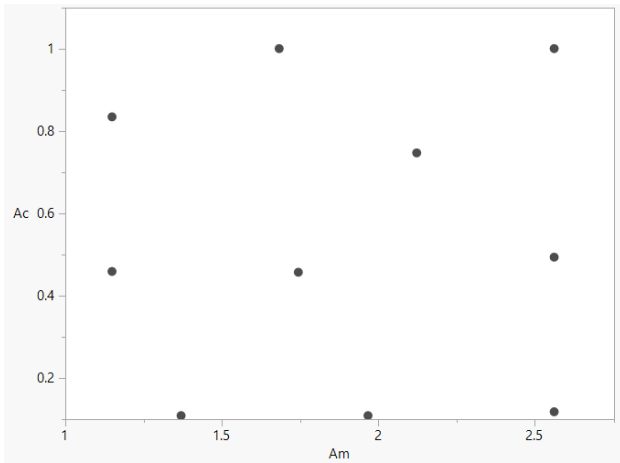


Fig. 4. Sphere packing space filling A_C vs A_M bubble plot.

7	1.000	1.683
8	0.117	2.560
9	0.747	2.121
10	0.458	1.150

Data gathered from the sphere packing design was summarized in Table 5. Figure 3 shows the visualization of the data points across the design space generated from the sphere packing design. As mentioned in section 2.2, sphere packing favors values lying at the edges of the design space before plotting points at the middle.

III. RESULTS AND DISCUSSION

A. Sphere packing runs

TABLE V. A_C AND A_M VALUES FROM SPHERE PACKING.

	A_C (mm ²)	A_M (mm ²)
1	0.108	1.370
2	0.456	1.744
3	0.493	2.560
4	0.834	1.150
5	0.108	1.966
6	1.000	2.560

B. Current range

TABLE VI. DETERMINATION OF ALLOWABLE CURRENT RANGE.

Current (A)	Highest Temperature (°C)	Valid or Invalid
0	20	VALID
0.1	45.778	VALID
0.2	78.93	VALID
0.3	121.67	VALID
0.4	177.05	VALID
0.5	250.17	VALID
0.6	348.83	INVALID

Initial tests performed on the modelled thermoelectric leg using the minimum and maximum A_C and A_M values showed that the operating temperature reached 250.17°C when the current was set at 0.5A. Therefore, the researchers used 0.5A

as the maximum operating current during the runs since the operating temperature was still within 250 to 300°C range.

C. FUSION 360 models

The values of A_c and A_m in the sphere packing runs and the measured dimensions from the TEC-12706 module were used in modelling the segmented thermoelectric pairs. Using FUSION 360, the segmented thermoelectric pairs were modelled using the loft feature to create a loft between the contact cross-sectional area and the middle cross-sectional area. A total of 10 models were produced in the software and was exported in IGES format to enable cross-software compatibility between FUSION 360 and ANSYS. The models are shown in Figure 4.

D. Temperature profile

The corresponding temperature profiles for each run during ANSYS thermal-electric analysis are shown in Figure 5. The hot side and cold side temperatures across the current range were collected to determine the COP for each current increment per run.

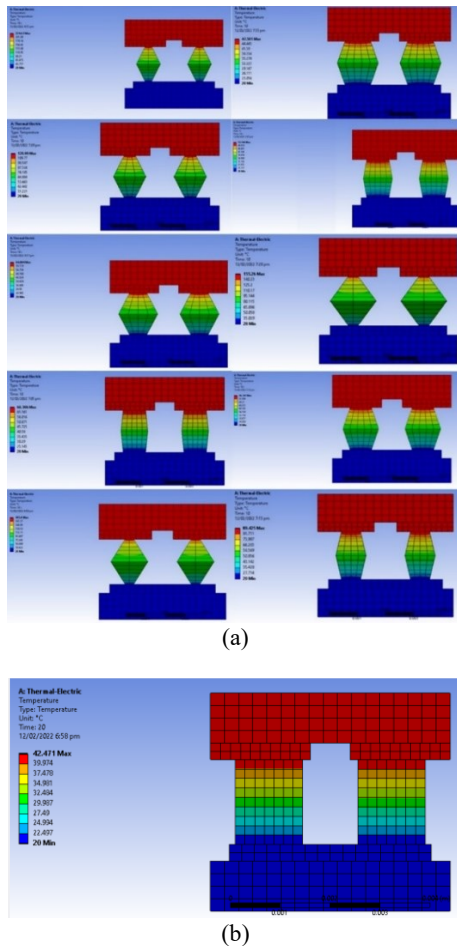


Fig. 5. Temperature profile of the leg geometry designs from ANSYS thermal-electric analysis (a) sphere packing runs (b) TEC-12706.

E. Determination of highest COP

TABLE VII. HIGHEST COP VALUES PER RUN.

Runs	Ac (mm ²)	Am (mm ²)	COP
TEC-12706	1.823	1.823	0.486
1	0.108	1.370	-2.244
2	0.456	1.744	-1.254
3	0.493	2.560	-1.911
4	0.834	1.150	0.037
5	0.108	1.966	-3.313
6	1.000	2.560	-1.028
7	1.000	1.683	0.466
8	0.117	2.560	-4.366
9	0.747	2.121	-1.070
10	0.458	1.150	-0.465

The highest COP for each run is listed in Table 7. Negative COP for runs 1, 2, 3, 5, 6, 8, 9, and 10 indicated that Peltier cooling did not overcome both heat conduction and Joule heating when current was induced in the device. Peltier cooling of thermoelectric coolers should overcome these heating mechanisms to produce a cooling effect [1]. Meanwhile, runs 4 and 7 overcame both heat conduction and Joule heating, which led to positive COP values of 0.037 and 0.466, respectively. It was observed that the positive COP from the runs was lower than the COP of TEC-12706.

F. Gaussian process model

1) Actual by predicted plot

Regression analysis using the actual by predicted plot of COP shows that the COP values were close to the regressed diagonal line, which means that the predicted COP values were close to the actual COP values. Figure 6 shows that data is scattered almost symmetrically above and below the 45° line. Therefore, the predicted values generated by the Gaussian process model were not underestimating or overestimating.

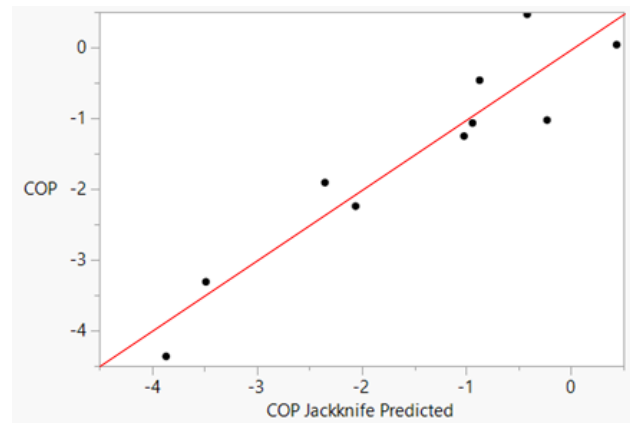


Fig. 6. Actual COP by COP Jackknife Predicted.

2) Gaussian process model report

Figure 7 presented that the COP increases as A_C increases. The relationship was seen to exhibit an increasing periodic function. Meanwhile, COP decreases as A_M increases. This relationship between A_M and the COP was observed to be linear.

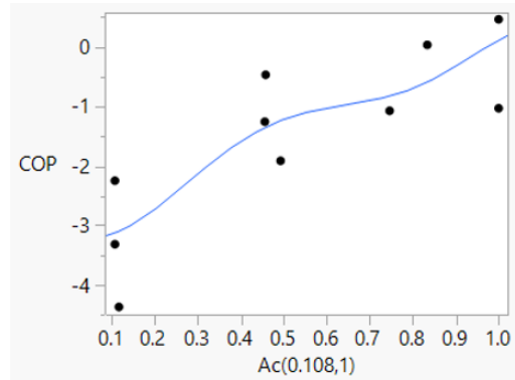
The main effect is the integrated total variation by the individual factor to the response, which is computed by the ratio of the functional effect and the total variation. Similarly, the functional interaction effects describe the effects of the factors on the response. Lastly, the total sensitivity measures the influence of each factor and its two-way interaction on the response variable [19].

TABLE VIII. MODEL REPORT FROM THE GAUSSIAN PROCESS MODEL.

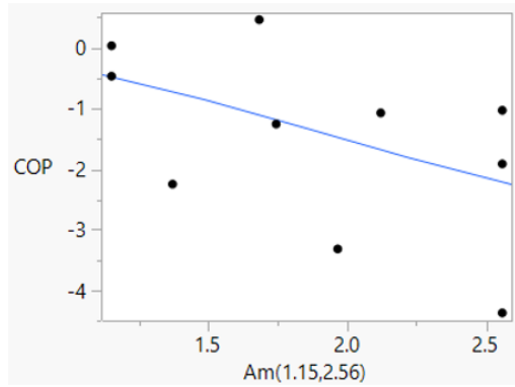
Column	Theta	Total Sensitivity	Main Effect	Ac Interaction	Am Interaction
Ac	1.098	0.742	0.720	-	0.0212
Am	0.191	0.280	0.258	0.0212	-
μ	σ^2	$-2*\text{LogLikelihood}$			
-1.568	3.076	24.544			

From the model report in Table 8, A_C showed a higher main effect value of 0.720 compared to A_M at 0.258. This means that A_C has a larger effect compared to A_M in the behavior of the response COP. Meanwhile, A_C and A_M interactions had a value of 0.0212. This value describes the changes of the response COP to the effect of A_M depending on A_C , and vice versa. Lastly, the total sensitivity of A_C and A_M variable were 0.742 and 0.280, respectively. The $-2*\text{LogLikelihood}$ is the negative of the log-likelihood function. The log-likelihood describes the deviance of the model. A higher log-likelihood indicates better parameter estimate. In contrast, smaller $-2LL$ values indicate better model fits. This value is usually used to compare different models to determine the better fit model [20]. Since the study did not compare two different models, the $-2LL$ was not used for data interpretation. However, the $-2LL$ may serve for future model comparisons.

Meanwhile, the Gaussian process surface plot in Figure 8 shows that the COP was influenced by the change in A_C and A_M . It was observed that as the value of A_C increases, the value of the COP increases. Inversely, as A_M increases, the COP decreases. The Gaussian process modelled surface plot for both factors can be seen in the figure, which represents the combined influence of A_C and A_M to the COP.



(a)



(b)

Fig. 7. Marginal model plot of COP with respect to (a) A_C and (b) A_M .

3) Prediction profiler

Figure 9 shows the prediction profiler generated through JMP software. The maximized COP within the experimental design space was determined to be 0.881, with a desirability of 0.996. The COP was achieved when the cross-sectional areas lie on the edges of the design space, where A_C was at maximum with a value of 1.000mm², and A_M was at minimum, with a value of 1.150mm².

TABLE IX. VALUES OF A_C , A_M , AND COP WITH THE HIGHEST DESIRABILITY FUNCTION

Desirability	A_C (mm ²)	A_M (mm ²)	COP
0.996	1.000	1.150	0.881

A desirability value close to 1 was expected due to only having one response variable in the study. Furthermore, the behavior of the factors to the response as presented in Figures 7 and 8 explain why the maximum A_C and the minimum A_M were the critical points of the COP response. Figure 7 shows that A_C and A_M have opposite effects to the COP, where the COP increases in relation to an increase in A_C and a decrease in A_M .

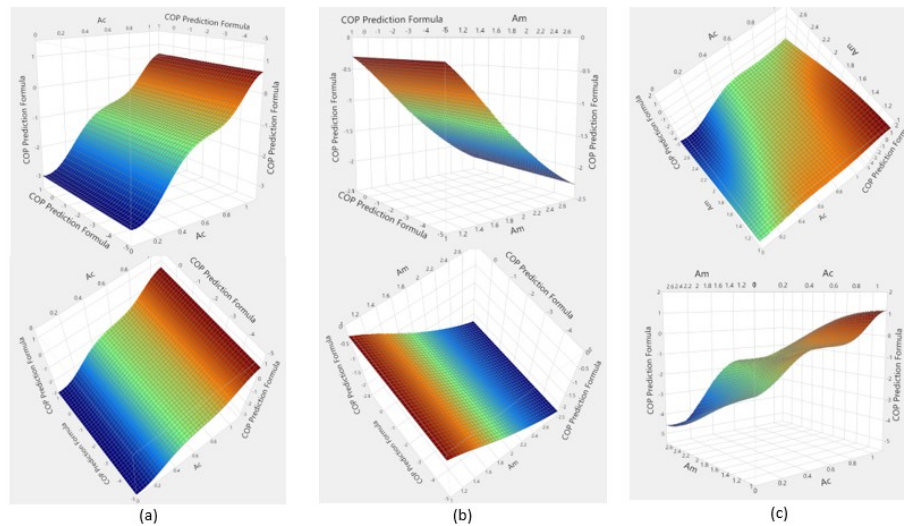


Fig. 8. Gaussian process surface plot of the COP in response to A_C and A_M (a) A_C and COP Predicted (b) A_M and COP Predicted (c) A_C , A_M , and COP.

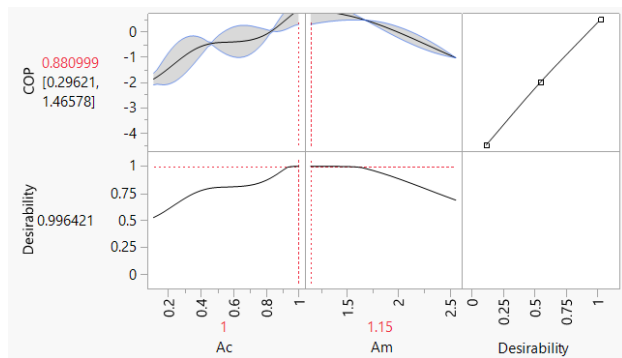


Fig. 9. Prediction profiler of the response COP.

4) COP comparison

TABLE X. COMPARISON OF RESULTS FOR MAXIMUM COP.

Parameter	Optimized Model	TEC-12706
A_C (mm ²)	1.000	1.823
A_M (mm ²)	1.150	1.823
COP	0.881	0.486

The optimized leg geometry from the Gaussian process model has a higher COP than the TEC-12706. Table 10 summarizes the COP and cross-sectional areas of each model. The design parameter A_C was set to be lower than that of the TEC-12706 model to achieve a truncated square leg geometry, which explains the lower value of this cross-sectional area compared to that of the baseline model.

IV. CONCLUSION

In this study, the research was able to optimize the leg geometry of the semiconductor thermoelectric cooler A_C and A_M through Gaussian process modelling. Results from the experiment showed that through the Gaussian process model, the COP of the optimized TEC was greater than the

conventional TEC-12706. During the study, it was determined that the allowable working current for the established design space was 0 to 0.5A. Through the Gaussian model, the researchers were able to determine the relationship of the middle and contact cross-sectional areas to the COP. The Gaussian model showed that A_C had a greater effect to the COP than A_M .

Moreover, the optimal cross-sectional areas for A_C and A_M were within the edges of the design space, with A_C equal to the maximum contact cross-sectional area and A_M equal to the minimum middle cross-sectional area. Following this study, future research could investigate different values of the contact and middle cross-sectional areas of the TEC leg. The behavior of the COP in response to the changes in these cross-sectional areas indicates that decreasing the middle area may result in higher COP values. Thus, it is recommended that future research investigate the effects of the leg geometry with a large middle cross-sectional area than the contact area. Moreover, optimizing other properties of the thermoelectric cooler such as the Seebeck coefficient, thermal conductivity, and electrical resistivity using the Gaussian process model could also be investigated. Lastly, fabricating an actual model with the optimized dimensions and running physical experiments utilizing statistical design of experiment (DOE) principles could be performed to validate the results of this study.

REFERENCES

- [1] H.J. Goldsmid. Introduction to Thermoelectricity (Springer Series in Material Science, 121, 2nd ed. 2016)
- [2] L. Chen, F. Meng, F. Sun, Thermodynamic analyses and optimization for thermoelectric devices: The state of the arts. Sci. China. Tech. Serv. 59, 442-455 (2016) <https://doi.org/10.1007/s11431-015-5970-5>
- [3] D. Zhao, and G. Tan. A review of thermoelectric cooling: Materials, modeling and applications, Appl. Therm. Eng. 66. 15-24 (2014) <http://dx.doi.org/10.1016/j.applthermaleng.2014.01.074>
- [4] S.B. Riffat and Xiaoli Ma, Improving the coefficient of performance of thermoelectric cooling systems, Int. J. of E. Res. 28(9). 753-768 (2004) <https://doi.org/10.1002/er.991>

- [5] S.M. Pourkiaei, M.H. Ahmadi, M. Sadeghzadeh, S. Moosavi, F. Pourfayaz, L. Chen, M.A.P. Yazdi, and R. Kumar, Thermoelectric cooler and thermoelectric generator devices: A review of present and potential applications, modeling and materials. *Energy*. 186. 115849 (2019) <https://doi.org/10.1016/j.energy.2019.07.179>
- [6] S. Shittu, G. Li, X. Zhao, and X. Ma, Review of thermoelectric geometry and structure optimization for performance enhancement, *Appl. Energy*. 268. 115075 (2020) <https://doi.org/10.1016/j.apenergy.2020.115075>
- [7] R. Lamba, S. Manikandan, S.C. Kaushik, and S.K. Tyagi. Thermodynamic modelling and performance optimization of trapezoidal thermoelectric cooler using genetic algorithm. *Therm. Sci. and Eng. Progress*. 6, 236-250 (2018) <https://doi.org/10.1016/j.tsep.2018.04.010>
- [8] A. Sahin, and B. Yilbas. The thermoelement as thermoelectric power generator: Effect of leg geometry on the efficiency and power generation. *Energy. Conv. and Mgmt.* 65, 26-32 (2013) <http://dx.doi.org/10.1016/j.enconman.2012.07.020>
- [9] S. Lin and J. Yu, Optimization of a trapezoid-type two-stage Peltier couples for thermoelectric cooling applications. *Int. J. of Refrig.* 65, 103-110 (2016). <http://dx.doi.org/10.1016/j.ijrefrig.2015.12.007>
- [10] A. Fabián-Mijangos, G. Min, and J. Alvarez-Quintana, Enhanced performance thermoelectric module having asymmetrical legs. *Energy Conv. and Mgmt.* 148. 1372-1381 (2017) <http://dx.doi.org/10.1016/j.enconman.2017.06.087>
- [11] F.F. Jaldurgam, Z. Ahmad, F. Touati, A.A. Ashraf, A. Shakoor, J. Bhadra, N.J. Al-Thani, D.S. Han, and T. Altahtamouni, Optimum sintering method and temperature for cold compact Bismuth Telluride pellets for thermoelectric applications. *J. of Aly. and Comp.* 877, 160256 (2021) <https://doi.org/10.1016/j.jallcom.2021.160256>
- [12] M. -W. Tian, F. Aldaw, H. Moria, H. Dizaji, and M. Wae-hayee. Cost-effective and performance analysis of thermoelectricity as a building cooling system; experimental case study based on a single TEC-12706 commercial module. *Case Studies in Therm. Eng.* 27, 101366 (2021) <https://doi.org/10.1016/j.csite.2021.101366>
- [13] I.T. Witting, T.C. Chasapis, F. Ricci, M. Peters, N.A. Heinz, G. Hautier, and G.J. Snyder. The Thermoelectric Properties of Bismuth Telluride. *Adv. Electronics Mat.*, 5(6), 1800904 (2019) <https://doi.org/10.1002/aelm.201800904>
- [14] C. Natoli, and S. Burke, *Computer Experiments: Space Filling Design and Gaussian Process Modeling* (Report No. STAT COE 7-2018). Air Force Institute of Technology (2018)
- [15] T. Wang, C. Zhang, H. Snoussi, and G. Zhang, Machine Learning Approaches for Thermoelectric Materials Research *Adv. Functional Mat.* 30(5), 1906041 (2019) <https://doi.org/10.1002/adfm.201906041>
- [16] J. Fox, and S. Weisberg, *Multivariate linear models in R. An R Companion to Applied Regression*. Los Angeles: Thousand Oaks. (2011)
- [17] R.D. Cook, and S. Weisberg, Graphics for assessing the adequacy of regression models, *J. of the American. Stat. A.* 92(438), 490-499 (1997) <https://doi.org/10.1080/01621459.1997.10474002>
- [18] S. Indumathi, V.D. Reddy, and G. Krishnaiah, Optimization of machining parameters using desirability function analysis and anova for thermo-mechanical form drilling. *Int. J. of Ind. Eng. and Tech.* 4(1), 19-26 (2014)
- [19] JMP Statistical Discovery, *Model Report. Predictive and Specialized Modeling*, (2021, March 28)
- [20] JMP Statistical Discovery, *Likelihood, AICc, and BIC. Fitting Linear Models*, (2021, March 28)

A Sliding Mode based Finite-Time Consensus Protocol for Heterogeneous Multi Agent UAS

Madhumita Pal

Institute of Engineering & Management
madhumita.pal@iemcal.com

Abstract—This paper presents a finite-time consensus for a connected network of heterogeneous unmanned aerial vehicles subjected to wind disturbances. A sliding mode control and graph algebraic theory based strategy is developed for consensus while independent tracking controller ensures robust trajectory following of an agent. Stability of the agents are decoupled from distributed consensus algorithm. Finite time convergence of the measurement vectors of heterogeneous agents to the agreement value are proved. Comparison results with the recent work show the superior performance of the proposed algorithm.

Index Terms—Consensus, sliding mode control, heterogeneous, multi agent systems

I. INTRODUCTION

Various distributed consensus protocols for multi-agent systems have proved to be useful in many applications scenarios, such as formation control of ground and aerial vehicles, consensus control of quad rotors for payload transportation, etc[1].

A typical consensus problem demands some or all of the states of the cooperating agents, which can be considered as general dynamical systems, reach to a common agreement based on their local information exchange. Such information exchange between these agents can be represented using a directed or an un-directed graph and various graph theoretic methods [2], [3], have been used to treat the consensus problem. Traditional control theory knowledge such as the Lyapunov direct method [4] [5], or the frequency domain analysis method [6], [7], etc, have also been used to solve the consensus problems of MASs. For the first-order, second-order, and high-order MASs, a lot of consensus protocols have been proposed. However, these algorithms are mostly proposed for homogeneous multi-agent systems and in particular systems having simple and double integrator dynamics, see [8]; [9]. However, in practical scenarios, the dynamics of various MASs are far from such conservative assumptions.

The consensus problems of the heterogeneous MASs which consist of agents with different dynamical properties have not been studied enough in past. This paper considers finite time consensus for heterogeneous multi-agent systems. Only a few papers consider heterogeneous cases of consensus problem. In particular, among many [10], [11] solved

the output consensus problem with a non-linear approach. Recently, [12] consider output consensus for heterogeneous multi-agent systems with linear dynamics. In other works, [13], [14] the problem of heterogeneous consensus is addressed where agents are considered as linear dynamical systems. Consensus for heterogeneous multi-agent systems composed of simple and double integrator are presented in [13] while [15] focused into a formation control problem. However, most consensus algorithms based on this technique assure consensus amongst the agents only asymptotically.

In most literature, consensus for MASs are achieved asymptotically. However, it is usually necessary to reach consensus in a finite time. The systems in such cases, show better disturbance rejection properties under a finite-time control. In recent years, focus is on obtaining the state consensus in homogeneous networks, that is, networks with identical agent dynamics. Many results of homogeneous finite-time consensus based on sliding mode are reported in the literature [16], [17], [18], [19]. The major advantage of sliding mode control is in its robustness against matched uncertainties, that is, when uncertainties are manifested through the input channel.

Because the dynamics of many mechanical systems can be modeled as double integrator, the MAS with double-integrator dynamics have gained great attention in recent years [16], [17]. Recently, MAS with general high-dimensional linear dynamics have attracted much attention of researchers in the control field [20], [21]. However, designing consensus algorithms using sliding mode for higher order MAS remains challenging.

This article proposes a sliding mode based control strategy for consensus on measurement variable of heterogeneous agents. The heterogeneous agents outputs are directed to follow reference trajectories generated by same number of ideal agents based on the idea of model reference control protocol. Ideal agents have simple first order dynamics and are connected with same directed graph as the original agents are. Reference trajectories are designed based on sliding mode based finite time consensus protocol. A finite time sliding mode based model reference tracking control is then designed for the practical heterogeneous agents to meet the

reference trajectories. The stability of the close loop for each agent is assured and decoupled from the outer loop sliding mode based tracking controller.

II. PROBLEM STATEMENT

In this section, we are going to study finite time consensus algorithms for a set of heterogeneous linear systems. Consensus control for heterogeneous agents with finite time convergence are not much explored in recent past. Although some results for such problem have been proposed in literature, most of them presents major drawbacks such as asymptotic convergence, computational effort, complexity or accuracy of the solution. Therefore, in this article a simple control design is proposed addressing the above drawbacks.

A. System Description

Consider the multi-agent linear systems

$$\begin{aligned} \dot{x}_i &= \bar{A}_i x_i + B_i u_i \\ y_i &= x_i \quad \forall i \in \mathcal{N} = 1, \dots, N \end{aligned} \quad (1)$$

$$z_i = C_i x_i \quad (2)$$

where $x_i \in \mathcal{R}^{n_i}$, $y_i \in \mathcal{R}^{n_i}$, $z_i \in \mathcal{R}^m$ and $u_i \in \mathcal{R}^m$ are the states, outputs, measurements and input vectors respectively. For all $i \in \mathcal{N}$, the matrices $\bar{A}_i \in \mathbb{R}^{n_i \times n_i}$, $B_i \in \mathbb{R}^{n_i \times m}$ and $C_i \in \mathbb{R}^{m \times n_i}$ with $m < \min n_i$ and $n_i > m$ are assumed to be constant and known.

The objective is to design an inner-loop controller which ensures stability for each subsystem and an outer-loop controller which achieves consensus on measurement vectors of each agent.

Assumption 1. *The N systems are considered to be heterogeneous, i.e., the matrices defining the systems may differ from one agent to another one and the vectors x_i may have different dimensions.*

Assumption 2. *The measurement vectors z_i represents the same quantity of interests for all agents. Therefore, the measurement vectors have the same dimension, $z_i \in \mathbb{R}^m \forall i \in \mathcal{N}$, where $m < \min\{n_i\}$.*

Assumption 3. *$\forall i \in \mathcal{N}$, the condition $\text{rank}(C_i B_i) = m$ holds. Therefore, the input vectors directly affects the measurement vector*

Assumption 4. *To design the controller for stability of each agent, each pair (A_i, B_i) is assumed to be controllable.*

It is natural to model information exchange between agents in a cooperative team by directed/ undirected graphs.

Assumption 5. *In this paper, it is assumed that communication graph is represented by directed spanning tree.*

Denoting for any agent $i \in \mathcal{N}$, \mathcal{N}_i as the subset of \mathcal{N} including all neighbors of agent i , i.e., all nodes that agent

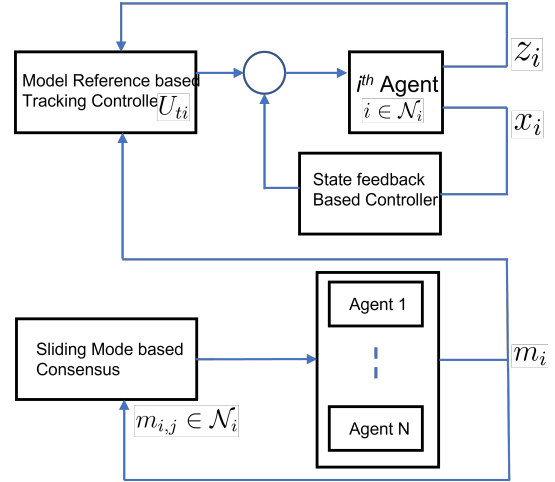


Fig. 1: Control Architecture

i can communicate. For the graph \mathcal{G} with N vertices and an edge set given by $E = \{(i, j) : j \in \mathcal{N}_i\}$ the adjacency matrix $A(\mathcal{G}) = (a_{ij})$ is a $N \times N$ matrix given by $a_{ij} = 1$, if $(i, j) \in E$ and $a_{ij} = 0$, otherwise. If there is an edge joining vertices i, j , i.e., $(i, j) \in E$, then i, j are adjacent. The degree d_i of a vertex i is defined as the number of its neighboring vertices, i.e., $d_i = \#j : (i, j) \in E$. Denote also $d_{max} = \max d_i$. Let Δ be the $N \times N$ diagonal matrix of d_i 's. The Laplacian of \mathcal{G} is the matrix $L(\mathcal{G}) = \Delta - A$.

III. CONTROLLER ARCHITECTURE

In order to achieve finite time consensus for MASs, a controller composing two parts, one for local stability and other for tracking an ideal trajectory leading to consensus is proposed, as shown in Fig.1 . Therefore, the control law for each agent is represented by

$$u_i(t) = u_{s_i}(t) + u_{t_i}(t), \quad i \in \mathcal{N} \quad (3)$$

where u_{s_i} is the local controller to assure stability for each agent and u_{t_i} is the tracking controller leading to consensus.

The ideal trajectory is generated by sliding mode based consensus algorithm. The tracking controller for each agent is also designed based on sliding mode. In the following subsections, methods are described for the design of stability and tracking controller.

A. Stability Design

According to the Assumption 4, for each agent, there exists a local state feedback Hurwitz controller assuring stability of each system.

$$u_{s_i} = -K_i x_i \quad (4)$$

This ensures that the matrix A_i is Hurwitz which is represented as $A_i = \bar{A}_i - B_i K_i$. Therefore, the system dynamics

in (1) can be represented in closed loop form as:

$$\begin{aligned} \dot{x}_i &= Ax_i + B_i u_i \\ y_i &= x_i \quad \forall i \in \mathcal{N} = 1, \dots, N \end{aligned} \quad (5)$$

$$z_i = C_i x_i \quad (6)$$

The objective is to design a finite time consensus algorithm which guarantees that the measurement vectors z_i s reach an agreement. In the literature, classical distributed consensus algorithms have been intensively studied. Inherently, their stability and performance properties are well documented see, for instance, [9], [22], etc.

B. Distributed Consensus Algorithm

A closed loop Model-Reference-Tracking-Controller which tries to compare the measurement output z_i of an agent with an ideal reference trajectory for consensus is proposed. The reference trajectories are generated by introducing N -number of simple first order homogeneous agents. First a sliding mode based finite time consensus controller is designed for these N simple agents. Then these N ideal trajectories are followed by N -number of heterogeneous agents with sliding mode based finite time tracking controller. Therefore, in this paper SMC and graph theories are used for the development of consensus algorithms for a group of heterogeneous agents connected with fully connected graph.

1) *Reference Trajectory Generation:* To generate reference trajectories N -number of m th order (same as the measurement variable order of the heterogeneous agents) agents are considered as follows:

$$\dot{m}_i = v_i \quad i = 1, \dots, N \quad (7)$$

where $m_i \in \mathcal{R}^m$ represent the agent's state and $v_i \in \mathcal{R}^m$ is its input. The agents are said to reach consensus when

$$m_i = m_j, \quad i, j = 1, \dots, N \quad (8)$$

The agents are assumed to constitute directed and connected communication graph same as heterogeneous agents. The inputs v_i are designed based on SMC to satisfy the consensus condition. Therefore, switching surfaces are designed so that if the sliding mode occurs then the consensus condition (8) holds.

2) *Preliminaries:* In this section, a few basic results of SMC and algebraic graph theories pertinent to the consensus problem will be discussed.

Sliding Modes: Consider the design of an SM controller for the system of the form

$$\dot{x} = f(x) + b(x)u, \quad x \in \mathcal{R}^n, u \in \mathcal{R}^m \quad (9)$$

The state $x \rightarrow 0$ involves the selection of switching functions $s(x) \in \mathcal{R}^m$ and controls u acting discontinuously on $s(x)$

in the form

$$u_i = \begin{cases} u_i^+ & \text{if } s_i(x) > 0, \\ u_i^- & \text{if } s_i(x) < 0 \end{cases} \quad \forall i = 1, \dots, m \quad (10)$$

With the appropriate choice of control magnitudes, the states of the system reach $s(x)$ in finite time and then continue to remain on it. The system motion on the surface $s(x) = 0$ is known as sliding mode and is of reduced order (by m). Details can be found in [23]. For 1st order system $\dot{s} = u$,

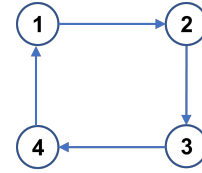


Fig. 2: Fully Connected Directed Graph

the discontinuous controller $u = -M \text{sign}(s)$, $M > 0$ will enforce sliding mode on the surface $s = 0$ at the time instant

$$T_s \leq \frac{s_0}{M}, \quad s_0 = s(t = 0). \quad (11)$$

This is used to tune the consensus time. Sliding mode consensus algorithm assume occurrence of ideal sliding mode on the surface $s(t)$ after a finite-time interval T_s , so that $s(t \geq T_s) = 0$.

Algebraic Graph: It is natural to model information exchange between agents in a cooperative team by directed/undirected graphs (e.g. [24]). A digraph (directed graph) consists of a pair (\mathcal{N}, E) , where \mathcal{N} is a finite nonempty set of nodes and $E \in \mathcal{N}^2$ is a set of ordered pairs of nodes, called edges. A directed path is a sequence of ordered edges of the form $(v_{i1}, v_{i2}), (v_{i2}, v_{i3}), \dots$, where $v_{ij} \in \mathcal{N}$, in a digraph. An undirected path in an undirected graph is defined analogously, where (v_{ij}, v_{ik}) implies (v_{ik}, v_{ij}) . A digraph is called strongly connected if there is a directed path from every node to every other node. An undirected graph is called connected if there is a path between any distinct pair of nodes. For a directed algebraic graph \mathcal{G} with N vertices and an edge set given by $E = \{(i, j) : j \in \mathcal{N}_i\}$, the following matrices can be formulated: the $N \times N$ adjacency matrix $\mathcal{A}(\mathcal{G}) = [a_{ij}]$ with $a_{ij} = 1$, if $(i, j) \in E$ and $a_{ij} = 0$, otherwise. Therefore, if there is an edge connecting two vertices i, j , i.e. $(i, j) \in E$, then i, j are called *adjacent*. The *degree* d_i of vertex i is defined as the number of its neighboring vertices, i.e., $d_i = \#j : (i, j) \in E$. Define $d_{max} = \max\{d_i\}$. The $n \times n$ diagonal degree matrix $\Delta(\mathcal{G}) = \text{diag}\{d_i\}$ and the Laplacian matrix $L(\mathcal{G}) = \Delta(\mathcal{G}) - \mathcal{A}(\mathcal{G})$. As example, the Laplacian

matrix for the agents connected as in Fig. 2 is

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{bmatrix} \quad (12)$$

As the Laplacian matrix is at the core of the consensus algorithms, one of its important property is highlighted here. An $N \times N$ Laplacian matrix is positive semi-definite with rank $(N - 1)$. Its single zero eigenvalue corresponds to a right eigenvector made up of identical elements—it is this property that is used to provide consensus. For instance, if there exists a vector of non-zero elements $m = [m_1 \dots m_n]^T$ that satisfies the system of algebraic equations

$$L(G)\mathbf{m} = \mathbf{0} \quad (13)$$

then $m_i = m_j \forall i, j = 1 \dots N$. In fact, if the inputs v_i are selected to yield the closed-loop dynamics for the system in (7) as

$$\dot{\mathbf{m}} = -L\mathcal{G}\mathbf{m} \quad (14)$$

then the agents reach a consensus with value found in [8], [25]

$$m_c = \frac{1}{N} \sum_{i=1}^N m_{i0}, \quad m_{i0} = m_i(t=0). \quad (15)$$

This well-known algorithm, however, ensures consensus only asymptotically.

These relations gives clue at the role of sliding mode control: if the inputs v_i are chosen to enforce sliding mode on $s = L(G)\mathbf{m} = 0$, then the agents reach a consensus. Below we states a theorem [26]

Theorem 1 ([26]). *Considering N - fully connected agents with at least 2 agents and the following properties: 1) The dynamics of each agent is given by $\dot{m}_i = v_i$; and 2) Its graph, G , generates the $N \times N$ Laplacian matrix $L(G)$ If the inputs $\mathbf{v} = [v_1 \dots v_N]^T$ are selected to be discontinuous, component-wise, on the surfaces*

$$\mathbf{s} = [s_1 \dots s_N]^T = L(G)[m_1 \dots m_N]^T,$$

in the form $v_i = -M \text{sign}(s_i)$, $M \geq 0$, so that sliding mode occurs on each s_i , then the agents reach a consensus within a finite time interval. Moreover, the consensus value attained by the agents is a constant.

Theorem 2 ([26]). *Consider N agents having same properties as those mentioned in Theorem 1. In addition, if there are agents m_p and m_q whose initial states $m_{p,q0}$ satisfy*

$$m_{p0} \geq 0, m_{p0} \geq m_{k0} \geq m_{q0}, \quad k = 1 \dots N, \quad k \neq p, q,$$

than, once sliding mode occurs on $\mathbf{s} = 0$ owing to the discontinuous inputs, the agents reach a consensus value $m_c = (m_{p0} + m_{q0})/2$ at the time instant $T_c = (m_{p0} -$

$m_{q0})/2M$.

The rest of the contribution consists in using this well known sliding mode based consensus algorithm to reach an agreement on those reference dynamics, while applying a sliding mode based model tracking controller to the original heterogeneous system. Thus, reference trajectory is generated in order to achieve

$$(m_i - m_j) = 0 \quad \forall i, j \in \mathcal{N}^2 \quad \text{in finite time} \quad (16)$$

In other words, system (7) or (14) can then be seen as a reference model for the system in (1).

3) *Model Reference Tracking Controller:* The objective of the tracking controller is fulfilled by defining

$$(m_i - z_i) = 0 \quad \forall i, j \in \mathcal{N}^2 \quad \text{in finite time} \quad (17)$$

The above objective is achieved in finite time by sliding mode based tracking controller. This ensures that the real MASs will have identical performances as the reference model. To meet the objective in (17), we define the sliding surface by

$$s_i = z_i - m_i, \quad \forall i \quad (18)$$

To ensure that sliding mode occurs on each $s_i = 0$ in finite time, the dynamics of s_i has the following form

$$\dot{s}_i = -\beta \text{sign}(s_i), \quad \beta \geq 0 \quad (19)$$

Thus, it follows:

$$\dot{z}_i - \dot{m}_i = -\beta \text{sign}(z_i - m_i) \quad (20)$$

$$C_i(A_i x_i + B_i u_{ti}) + L(G)_i \mathbf{m} = -\beta \text{sign}(z_i - m_i) \quad (21)$$

Due to the Assumption 3, $(C_i B_i)$ is invertible for all agent i , and thus the reference tracking controller is defined by

$$u_{ti} = (C_i B_i)^{-1}(\dot{m}_i - \beta \text{sign}(z_i - m_i) - C_i A_i x_i) \quad (22)$$

Therefore, the design of tracking controller is independent of the consensus algorithm applied to ideal homogeneous agents generating reference trajectories. In other words, control strategy for heterogeneous agents is decoupled from consensus algorithm.

IV. CONVERGENCE ANALYSIS

The following theorem states our main result:

Theorem 3. *If Assumptions 1-5 are satisfied, then the control law (3) given by*

$$u_i = -K_i x_i + (C_i B_i)^{-1}(\dot{m}_i - \beta \text{sign}(z_i - m_i) - C_i A_i x_i) \quad (23)$$

where \dot{m}_i is given by (7), guarantees that the closed loop system of (1) is stable and reaches consensus on measurement variable in finite time.

Proof. According to Assumption 3, there exist a change in coordinates [27], in the form

$$P_i x_i = \begin{bmatrix} \chi_i \\ z_i \end{bmatrix}$$

such that system in (5) can be decomposed in the form as

$$\begin{bmatrix} \dot{\chi}_i \\ \dot{z}_i \end{bmatrix} = \begin{bmatrix} A_{11i} & A_{12i} \\ A_{21i} & A_{22i} \end{bmatrix} \begin{bmatrix} \chi_i \\ z_i \end{bmatrix} + \begin{bmatrix} 0 \\ B_{2i} \end{bmatrix} u_{ti} \quad (24)$$

where $\chi_i \in \mathcal{R}^{n_i-m}$ and

$$P_i A_i P_i^{-1} = \begin{bmatrix} A_{11i} & A_{12i} \\ A_{21i} & A_{22i} \end{bmatrix}$$

In the above representation, $A_{11i} \in \mathcal{R}^{(n_i-m) \times (n_i-m)}$ is Hurwitz, $B_{2i} \in \mathcal{R}^{m \times m}$ is full rank and invertible and satisfies $T_i B_i = [0 \ B_{2i}^T]^T$. [27] proposed a method to construct the matrix P_i for changing coordinates of a system. This type of representation is called canonical representation of a system. The system is now represented in an appropriate manner with respect to the problem of measurement variable consensus. Applying the canonical transformation to the system, the control law in (23) can be rewritten as:

$$u_i = -K_i x_i + B_{2i}^{-1} [\dot{m}_i - \beta \text{sign}(z_i - m_i) - [0 \ I_m] P_i A_i x_i] \quad (25)$$

With the above control law (25) each closed loop system in (1) becomes:

$$\dot{x}_i = (\bar{A}_i - B_i K_i) x_i + B_i (B_{2i})^{-1} [\dot{m}_i - \beta \text{sign}(z_i - m_i) - [0 \ I_m] P_i A_i x_i]$$

Recalling that $s_i = z_i - m_i$, one has

$$\begin{bmatrix} \dot{\chi}_i \\ \dot{s}_i \end{bmatrix} = \begin{bmatrix} A_{11i} & A_{12i} \\ 0 & -\beta \frac{I_m}{|s_i|} \end{bmatrix} \begin{bmatrix} \chi_i \\ s_i \end{bmatrix} + \begin{bmatrix} A_{12i} m_i \\ 0 \end{bmatrix} \quad (26)$$

The variable \mathbf{m} is obtained by solving a sliding mode based consensus problem $\dot{\mathbf{m}} = -\beta L(\mathcal{G}) \text{sign}(\mathbf{m})$ which guarantees consensus in finite time interval. Moreover the consensus value attained by the first order homogeneous agents is a constant [26]. Consequently, $(m_i - m_j) = 0 \ \forall i, j \in \mathcal{N}^2$ is achieved in finite time. Finally, according to the separation principle, conclusion on the stability of the multi-agent system in (26) for all $i \in \mathcal{N}$ can be made. Thus, for all $i \in \mathcal{N}$, $m_i - z_i = 0$ is achieved in finite time [23]. \square

Remark 1. Note that, control strategy for heterogeneous agents is decoupled from consensus algorithm. In other words, analysis of each agent is separated from the analysis of distributed consensus algorithm. It has been shown that (1) will achieve consensus on measurement variables z where the final value of z_i depends on the initial conditions of the consensus variables m_i . Thus, no matter what the initial state of the system is, (1) will always achieve consensus on the agreement value of m (see for instance Figure 5).

V. SIMULATION RESULTS

In this section, simulation results regarding the efficiency of the approach introduced in the previous sections is presented. Therefore, consider a set of $N = 4$ heterogeneous Unmanned aerial systems as considered in recent paper [28]. The matrices \bar{A}_i , B_i and C_i in equation (1) for the MAS are considered as

$$\begin{aligned} \bar{A}_1 &= \begin{bmatrix} -1 & 0.5 \\ 0.05 & -1 \end{bmatrix}; & B_1 &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}; & C_1 &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \\ \bar{A}_2 &= \begin{bmatrix} -2 & 1 \\ 0 & -0.9 \end{bmatrix}; & B_2 &= \begin{bmatrix} -1 \\ -1 \end{bmatrix}; & C_2 &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \\ \bar{A}_3 &= \begin{bmatrix} 0 & 1 & 1 \\ -2 & 0.1 & 2 \\ 1 & 2 & 3 \end{bmatrix}; & B_3 &= \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}; & C_3 &= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \\ \bar{A}_4 &= \begin{bmatrix} 0 & 1 & 1 & 0 \\ -2 & 0.1 & 2 & 2 \\ 0 & 1 & 2 & 3 \\ 0.2 & 0.4 & 0.1 & 0.3 \end{bmatrix}; & B_4 &= \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}; & C_4 &= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}^T \end{aligned}$$

Clearly the considered agents are heterogeneous as they have different dimensions and stability properties. Also, the Assumptions considered in Section II are all fulfilled for all the systems. Therefore, a pole placement controller gain K_i 's can be found such that $A_i = \bar{A}_i - B_i K_i$ are Hurwitz for all agents.

The above considered four agents are connected through a graph expressed by the following Laplacian matrix satisfying

$$\text{Assumption 5, } L = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

The unmanned aerial agents achieve a consensus on the common height by applying a control strategy described in Section III. To generate ideal reference trajectory, four first order homogeneous agents are considered. They meet consensus in finite time by applying sliding mode control as detailed in Section III-B1. Heterogeneous agents track the reference trajectory in finite time by applying the control as derived in (22). Simulation results are shown on the effect of identical and non-identical initial conditions between the ideal and practical agents.

Figure 3 compares asymptotic approach and sliding mode based approach for achieving altitude consensus for the ideal agents. The initial conditions for the agents m_i s are considered same as $z_i(0)$ s. $m_1(0) = z_1(0) = 10$, $m_2(0) = z_2(0) = 20.6$, $m_3(0) = z_3(0) = -3$, $m_4(0) = z_4(0) = 14$. Thus, consensus is achieved on the average value of $m_i(0)$ s according to (15). These trajectories will be considered as the reference trajectories for the practical heterogeneous agents to follow. Clearly in Figure 4, SM based consensus approach reaches agreement value in finite time compared

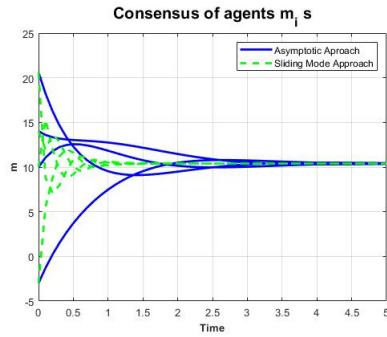


Fig. 3: Evolution of the variable m_i 's without and with sliding mode approach

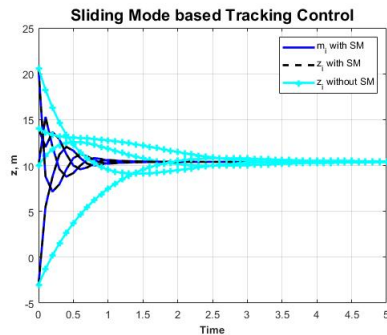


Fig. 4: Evolution of the variable m_i 's and measurement z_i 's when $m_i(0) = z_i(0)$

to the asymptotic approach adopted in [28]. In Figure 4, the measurement variables z_i s in MASs (1) accurately follow the final consensus value achieved by MASs in (7), where initial condition on both the variables z_i s and m_i s are considered same. Thus, system (1) achieves consensus converging to the agreement value of the ideal agents m_c . Figure 4 and 5 show a comparison results among SM based and Asymptotic based approach on how the practical agents track the reference trajectories in the case when the practical agents have identical and non-identical initial conditions. In 5, $m_i(0)$ s are considered as $m_1(0) = 30$, $m_2(0) = 25$, $m_3(0) = -2$, and $m_4(0) = -13$ which are different than $z_i(0)$ s.

Figure 6 shows how the asymptotic based approach adopted in [28] failed to meet consensus on the measurement variables z_i s in presence of matched disturbances. Whereas with the SM based approach, z_i s successfully track the final agreement value m_c in presence of disturbances. Disturbance applied through the input channel of heterogeneous agents is given in Figure 7.

VI. CONCLUSIONS

In this paper, an approach for heterogeneous multi-agent systems to reach measurement variable consensus in finite-time has been presented. The main advantage with this approach remains in the separation of the stability of each

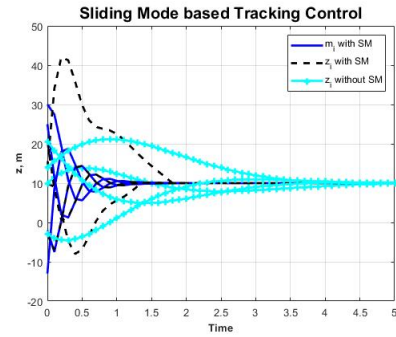


Fig. 5: Evolution of the variable m_i 's and measurement z_i 's when $m_i(0) \neq z_i(0)$

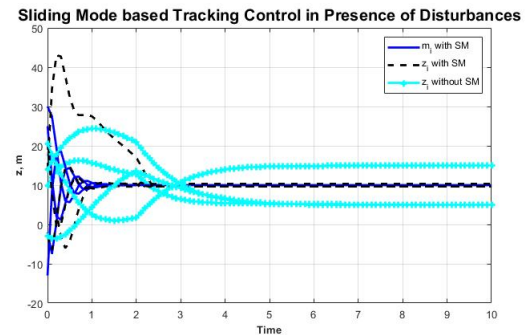


Fig. 6: Evolution of the variable m_i 's and measurement z_i 's when $m_i(0) \neq z_i(0)$ in presence of disturbances

agent and the distributed consensus algorithm. Sliding mode based distributed consensus algorithm for simple integrator dynamics is used in order to derive an appropriate tracking control law for a MASs composed of non-identical heterogeneous agents. Both the distributed consensus control for first order agents and tracking control for heterogeneous MASs are designed based on sliding mode approach. It is shown that the performance of both the algorithms improves in terms of convergence time w.r.t the literature.

REFERENCES

- [1] W. Ren, R. Beard, and E. Atkins, "A survey of consensus problems in multi-agent coordination," *American Control Conference, 2005. Proceedings of the 2005*, pp. 1859–1864 vol. 3, June 2005.
- [2] P. Lin and Y. Jia, "Consensus of second-order discrete-time multi-agent systems with nonuniform time-delays and dynamically changing topologies," *Automatica*, vol. 45, pp. 2154–2158, 09 2009.
- [3] F. Xiao and L. Wang, "State consensus for multi-agent systems with switching topologies and time-varying delays," *International Journal of Control*, vol. 79, pp. 1277–1284, 11 2006.
- [4] W. Ni and D. Cheng, "Leader-following consensus of multi-agent systems under fixed and switching topolo-

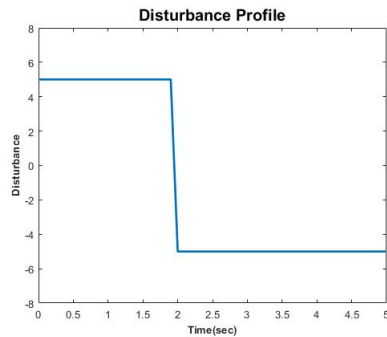


Fig. 7: Disturbance Applied through the Input Channel of Heterogeneous Agents

- gies,” *Systems and Control Letters*, vol. 59, pp. 209–217, 03 2010.
- [5] Y. Hong, L. Gao, D. Cheng, and J. Hu, “Lyapunov-based approach to multiagent systems with switching jointly connected interconnection,” *Automatic Control, IEEE Transactions on*, vol. 52, pp. 943 – 948, 06 2007.
- [6] Y.-P. Tian and C.-L. Liu, “Consensus of multi-agent systems with diverse input and communication delays,” *Automatic Control, IEEE Transactions on*, vol. 53, pp. 2122 – 2128, 11 2008.
- [7] G. X. Yan J and L. X., “Consensus pursuit of heterogeneous multi-agent systems under a directed acyclic graph,” *Chinese Physics B*, vol. 20, p. 048901, 04 2011.
- [8] R. Olfati-Saber and R. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.
- [9] W. Ren and R. W. Beard, *Distributed Consensus in Multi-Vehicle Cooperative Control*, ser. Communications and Control Engineering. Springer Publishing Company, Incorporated, 2007.
- [10] Z. Qu, J. Chunyu, and J. Wang, “Nonlinear cooperative control for consensus of nonlinear and heterogeneous systems,” in *2007 46th IEEE Conference on Decision and Control*, 2007, pp. 2301–2308.
- [11] H. Du, G. Wen, D. Wu, Y. Cheng, and J. Lü, “Distributed fixed-time consensus for nonlinear heterogeneous multi-agent systems,” *Automatica*, vol. 113, p. 108797, 2020.
- [12] N. Danaeefard and V. J. Majd, “Consensus of heterogeneous multi-agent systems using output feedback,” in *2015 AI Robotics (IRANOPEN)*, 2015, pp. 1–7.
- [13] Y. Zheng, Y. Zhu, and L. Wang, “Consensus of heterogeneous multi-agent systems,” *Control Theory & Applications, IET*, vol. 5, pp. 1881 – 1888, 12 2011.
- [14] H. Kim, H. Shim, and J. H. Seo, “Output consensus of heterogeneous uncertain linear multi-agent systems,” *IEEE Transactions on Automatic Control*, vol. 56, no. 1, pp. 200–206, 2011.
- [15] U. T. Jönsson and C.-Y. Kao, “Consensus of heterogeneous linear agents applied to a formation control problem,” in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 6902–6907.
- [16] S. Khoo, L. Xie, and Z. Man, “Robust finite-time consensus tracking algorithm for multirobot systems,” *IEEE/ASME Transactions on Mechatronics*, vol. 14, no. 2, pp. 219–228, 2009.
- [17] G. Masood and G. N. Sergey, “Finite-time coordination in multiagent systems using sliding mode control approach,” in *2013 American Control Conference*, 2013, pp. 2050–2055.
- [18] P. Alessandro, F. Mauro, P. Alessandro, and U. Elio, “Recent advances in sliding-mode based consensus strategies,” in *2014 13th International Workshop on Variable Structure Systems (VSS)*, 2014, pp. 1–6.
- [19] J. Dávila, “Distributed tracking of first order systems using second-order sliding modes,” *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 1392–1397, 2014, 19th IFAC World Congress.
- [20] J. H. Seo, H. Shim, and J. Back, “Consensus of high-order linear systems using dynamic output feedback compensator: Low gain approach,” *Automatica*, vol. 45, no. 11, pp. 2659–2664, 2009.
- [21] Z. Zuo, “Nonsingular fixed-time consensus tracking for second-order multi-agent networks,” *Automatica*, vol. 54, pp. 305–309, 2015.
- [22] R. Olfati-Saber, J. A. Fax, and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” *Proceedings of the IEEE*, vol. 95, pp. 215–233, 2007.
- [23] V. I. Utkin, *Sliding Modes in Control Optimization*. Berlin: Springer-Verlag, 1992.
- [24] C. Godsil and G. F. Royle, *Algebraic Graph Theory*. Springer, 2001.
- [25] M. Ji and M. Egerstedt, “A graph-theoretic characterization of controllability for multi-agent systems,” in *2007 American Control Conference*, 2007, pp. 4588–4593.
- [26] S. Rao and D. Ghose, “Sliding mode control-based algorithms for consensus in connected swarms,” *International Journal of Control*, vol. 84, no. 9, pp. 1477–1490, 2011.
- [27] C. Edwards and S. K. Spurgeon, “Sliding mode stabilization of uncertain systems using only output information,” *International Journal of Control*, vol. 62, no. 5, pp. 1129–1144, 1995.
- [28] C. Gabriel, B. A. Lara, S. Alexandre, and N. Silviu-Iulian, “On the consensus of heterogeneous multi-agent systems: a decoupling approach,” in *2012 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems*, 2012, pp. 246–251.

Improving Accessibility of Remote Drone Control with a Streamlined Computer Vision Approach

Evan Lowhorn

Rocio Alba-Flores

Department of Electrical and Computer Engineering

Georgia Southern University

Statesboro, USA

evan_w_lowhorn@georgiasouthern.edu

ralba@georgiasouthern.edu

Abstract—The purpose of this work is to develop a method for classifying hand signals and using the prediction output in a drone control algorithm. To achieve this, methods based on Convolutional Neural Networks (CNNs) were applied. The hand signals chosen were the numerical hand signs for one through five for two-dimensional movement with a separate idle signal, and a fist for land. A script was created to automate one minute of training image capture for each class. Transfer learning with PyTorch (Python) was performed using a pre-trained 18-layer residual learning network (ResNet-18). The training process completed in three minutes and 43 seconds with five epochs and a final overall validation accuracy of over 99%. Implemented with the drone control, the classification performed as desired at approximately 60 predictions per second on desktop and 20 predictions per second on a Nvidia Jetson Nano.

Index Terms—Convolutional Neural Network, Drones, Human-Machine Interaction

I. INTRODUCTION

Deep learning is a subset of artificial intelligence (AI), and computer vision (CV) can be thought of as a subset of deep learning. CV utilizes deep learning while focusing on models that can interact with and interpret digital imaging. A fundamental component of CV is the convolutional neural network (CNN). CNNs receive inputs with grid-like topologies such as digital images and are trained to perform various classifications and other AI applications using the pixel data [1]. Traditional neural networks do have the capability of receiving pixel data in the form of an array, and can be trained to perform basic classifications based on this data. However, a traditional neural network does not infer image data with the same methods as human brain. For example, a human brain can see an image of the number three and determine it is in fact the number three based on its edges which combine to

form shapes and features. A traditional neural network does not find these edges and is only observing the encoded pixel color values themselves while trying to relate and detect patterns with each pixel. This method can be successful to a degree when trained adequately, but it fails in some circumstances. Hypothetically, if the neural network were to receive a random noise image that activated the correct nodes, the network would be tricked into a false positive with complete confidence. With CNNs, the convolutional layers can extract edges, features, etc. from the image before the data is put into the network. Fig. 1 shows this basic architecture, which allows CNNs to process and train on images in a way that more closely aligns with the human brain.

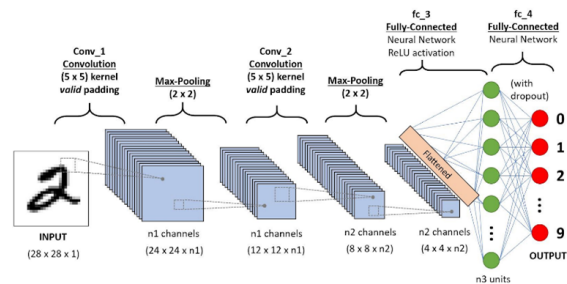


Fig. 1. Convolutional Neural Network layer diagram [2]

Consumer drones typically come with a dedicated remote controller or mobile software application that allows for multiple dimensions of movement. These controllers and the mobile devices running the control applications often require both hands to operate an assortment of joysticks and buttons. With the popularity of drones increasing within many age groups, these controls are sometimes considered too complex or even completely inaccessible to some persons with disabilities. In recent years, alternative control options that are more intuitive and accessible have been developed.

These include examples of hand gesture control utilizing CV concepts [3] [4]. Most current CV hand-based control models utilize the first-person view (FPV) camera of the drone. It is a simple and reliable method due to the camera being included on nearly every consumer drone. By using the FPV camera, the pilot is able to move with the drone and ensure its safe operation. However, this method has two primary disadvantages. The first is that with the constantly changing background of the FPV feed, more complex CV models must be utilized to isolate the pilot for effective and reliable flight control. Secondly, FPV camera CV applications require the pilot to always be in the camera view, often without a reference feed that shows the current drone FPV camera image. If the pilot mistakenly exited the frame, it would cause an unexpected idle, premature landing or potentially a collision depending on the control algorithm. More importantly, there are a significant number of people who would be unable to follow the drone due to physical limitations i.e., age and disabilities. Therefore, it becomes necessary to combine the accessibility of CV hand control with a stationary form of remote operation. With a stationary CV application, a more simplified model can be trained to the specific user given that setup time remains limited.

II. BASIC DRONE CONTROL

A. Drone Selection

This work is primarily focused on the high-level control aspect of the drone system i.e., movement commands. Therefore, emphasis was placed on having a complete pre-manufactured system with open compatibility for control methods via desktop applications. Fortunately, the Ryze Robotics Tello drone provides an inexpensive and fully developed platform with all required specifications for this and future research. Shown in Fig. 2, the Tello is a small quad-copter drone designed for indoor use. It satisfies both requirements of being a pre-built system while also having open connectivity for control through its 2.4 GHz WiFi connection. When connected, the drone is capable of sending and receiving data using set user datagram protocol (UDP) ports and sockets. The UDP ports are outlined in the Tello documentation [5].

B. Drone Control

After the drone was selected, a desktop software for controlling the drone had to be chosen. With the desktop software development kit (SDK) of the Tello, essentially all programming languages and software that can send encoded string messages using sockets and UDP ports are compatible. These include C++, MATLAB, and Python. C++ is a fast programming language capable of both controlling the drone and running AI models. However, this language contained the least amount of immediately available official documentation for the Tello,



Fig. 2. Ryze Robotics Tello drone

and is less intuitive for initial development. MATLAB offers an add-on package specifically for Tello drone control [6] and toolboxes that include many popular computer vision models. The Tello package simplifies WiFi connection and control messaging but omits useful commands such as the RC control command. This command allows a movement state to be sent to the drone that contains speed values for three dimensions of movement and a fourth value for yaw rotation speed [5]. This state command can be updated and sent with minimal delay and is therefore the best method for real-time control when compared with set time and distance movements.

This leaves Python, which was chosen for this work. It is the main language used in official documentation and example programs while providing full access to all SDK commands [5] [7]. Python also has optimized AI libraries which will allow the computer vision model and the control algorithm to run in the same script.

III. CNN DEVELOPMENT

A. Dataset Creation

The processes outlining the development of the CNN and its real-time application are shown in Fig. 3. The first step in performing transfer learning with a CNN model is collecting data for the training process. These photos must be taken in the same style as the resulting real-time classification, so the equipment setup for dataset creation must be identical to the final application. Fig. 4 shows the webcam, monitor and desktop used to create the stationary remote operation. The desktop system contains an Intel i7-6700k with a Nvidia 980Ti. The webcam (left) pointing downwards at the table allows for a constant background and a comfortable position for left-handed control. The sheet of paper taped below the camera eliminates glare from the camera light and provides a reference area to calibrate the camera position.

To create the image dataset, a Python script utilizing the OpenCV library was used to capture 1800 images at the webcam frame rate of 30 frames per second (FPS). This script was executed six times, once for

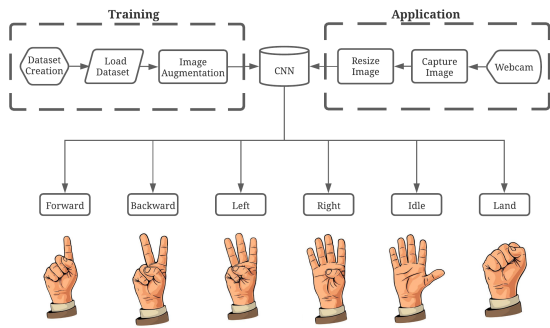


Fig. 3. CNN training and application process diagram



Fig. 4. Development workbench

each class. This dataset creation process generated a total of 10,800 images with a total run-time of approximately six minutes excluding breaks between program executions. A common flaw that can arise when training image classifiers is a concept referred to as overfitting. Overfitting occurs when the training data contains little variation and is not representative of the real-world data that will be given to the model. This results in a model that is “overfitted” to the training data, where the model appears to perform well until it is applied with this real-world data. This method of drone control does contain a level of overfitting by nature, it is certainly overfitted to the user’s hand and the table background. This is acceptable on the basis that this model is purposely trained to recognize only the user’s hand in this specific stationary position. However, further overfitting was reduced by moving the hand around the reference area during the image capturing process.

B. Transfer Learning with ResNet-18

Once the dataset has been captured, a pre-trained model must be chosen for transfer learning. Transfer learning is the process of slightly modifying an existing CV architecture and retraining it with a custom dataset to classify a new set of classes. The model chosen for this work was ResNet, specifically ResNet-18. ResNet is a residual learning network, which contains the core architecture of a series network with added skip con-

nections to reduce the effect of the vanishing gradient. Simply put, this negative effect on the gradient occurs when the depth of series architectures becomes so vast that accuracy lowers and training error increases. The skip connections in ResNet work to counter these effects and allow for more accurate and deeper CNNs [8].

ResNet networks perform accurately in a wide variety of configurations. These configurations include ResNet-18, ResNet-50, and ResNet-101. There are additional configurations, each with an appended numerical value. This value corresponds to the number of residual blocks in the network. Therefore, ResNet-150 is deeper than ResNet-101, ResNet-101 is deeper than ResNet-50, and so on. As the network increases in depth, the number of parameters increases. These parameters contribute to the complex calculations occurring during training and application, so increasing the number of parameters increases computing power required. Since the dataset being used contains only a few classes against a constant background, ResNet-18 was chosen for improved computational speed. With ResNet-18, a rapid training time and responsive real-time classification can be achieved.

Fortunately, the ResNet-18 model is included with the PyTorch torchvision library. This library is capable of importing a ResNet model that is pre-trained from popular large datasets with many classes [8]. Before the model can begin training, the dataset must be split between training and validation images. Of the complete dataset, 70% were set for training and 30% were separated for validation for each class. This selection was randomized and performed before any image augmentation took place. Image augmentation was used on the training images to further reduce overfitting while also adding robustness to the lightweight network.

These augmentations on the training images include:

- Resize to fit ResNet input layer
- Randomized horizontal flip to give appearance of right hand images
- Perspective shifting to account for inconsistent camera positioning
- Randomized “color jitter” (hue and brightness shifting) to account for inconsistent lighting and ignore skin tone

For the validation set, images were only augmented to what will be seen in application. Therefore, the only augmentation for the validation set was image resizing that was required for all images to convert the native webcam resolution to the 224x224 resolution of the standard ResNet input layer. A set of test images passed through this augmentation algorithm is shown in Fig. 5. Once the augmentations have been defined, the remaining training parameters can be set. In this training procedure, the cross entropy loss function was implemented. The optimizer was set with a learning rate of 0.001 and a momentum of 0.9. This learning rate

was decayed by a factor of 0.1 every two epochs. The training process was set to complete five epochs in total.



Fig. 5. Sample of augmented images
Note: This also contains images of a volunteer and is not a part of the official dataset

C. Real-time Inference and Control

After training was complete, the model outputs a set of class probabilities when given an input image. A function was used to extract the highest probability among all the classes which is ultimately the model’s main prediction for that image. The output classes were all labeled generically by their movement, which is not compatible with the Tello SDK. However, the extracted prediction label is a variable that can be implemented in a conditional statement algorithm for controlling the drone based on the prediction value. Within each conditional statement, the algorithm can send the corresponding Tello command in the correct syntax. OpenCV was used to show the drone camera feed which is required since the pilot will not be following the drone during flight. This was achieved in a similar manner as example programs utilizing the Tello SDK [7]. To enhance the user experience and interaction with the CV model, OpenCV was also used to display the webcam feed along with the current prediction. This allowed the user to adjust if necessary and ensure that the AI and the pilot are always in understanding with each other.

After a model trained, its new parameters were saved as a PyTorch accessible file. This saved file was then loaded into the control script previously mentioned. In the setup portion of the control script, the base ResNet-18 model from the PyTorch library is loaded into the algorithm. The parameter file is then loaded and used to overwrite the base ResNet parameters. This creates a replica of the model that was just trained for hand signal detection. This process is worth noting as it was repeated to load the same hand control ResNet-18 model on a Nvidia Jetson which represents edge devices such as modern laptops.

IV. RESULTS

As mentioned previously, the training process was set to complete a total of five epochs. For both the training and validation sets, the combined overall loss and accuracy of all classes were printed to the terminal after each epoch. After the five epochs were completed, the

```
Epoch 0/4
-----
train Loss: 0.3489 Acc: 0.9055
val Loss: 0.0169 Acc: 0.9996

Epoch 1/4
-----
train Loss: 0.0222 Acc: 0.9987
val Loss: 0.0055 Acc: 0.9996

Epoch 2/4
-----
train Loss: 0.0134 Acc: 0.9995
val Loss: 0.0049 Acc: 0.9996

Epoch 3/4
-----
train Loss: 0.0142 Acc: 0.9988
val Loss: 0.0042 Acc: 0.9996

Epoch 4/4
-----
train Loss: 0.0123 Acc: 0.9989
val Loss: 0.0044 Acc: 0.9996

Training complete in 3m 43s
Best val Acc: 0.999567
```

Fig. 6. Terminal output of training process

final training time and validation accuracy are displayed. Fig. 6 shows the terminal results for the final training procedure.

The final validation accuracy is nearly equal to one which is expected for a small dataset that is highly overfitted to a small list of classes. This accuracy value is artificial when compared to real-time classification. However, if it was exactly equal to one it would be indicative of saturation and the network learning could have ceased entirely. The training time completed in under four minutes, which is crucial to achieve a rapid setup time and allow for a unique model for the user.

While in use, the control algorithm displays the FPV camera feed from the drone along with the webcam feed and current prediction. The drone FPV camera feed is irrelevant in terms of displaying CNN performance, so shown in Fig. 7 is only a real-time classification instance of each class with the desktop that was used for training the network. The title of each window corresponds to the current prediction, and the current predictions per second (PPS) value is also included to show the speed at which the model is performing.

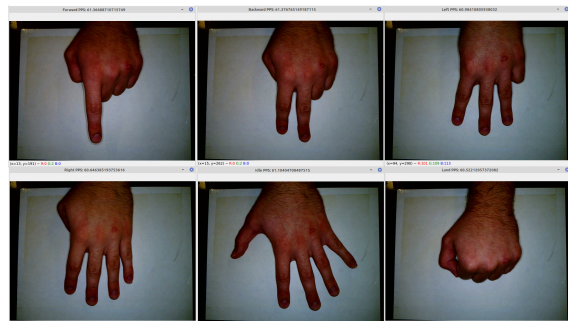


Fig. 7. Real-time output for each class with predictions per second

Each title shows an accurate prediction along with an approximate PPS of 60. This value doubles the frame rate of the webcam and provides a smooth and

responsive control algorithm. For the Nvidia Jetson, the webcam was moved from the previous workbench shown in Fig. 4 to a temporary station. The same real-time classification script and model was executed by the Jetson. Fig. 8 shows a classification instance in the same format as Fig. 7. The model and inference script performs accurately on the Nvidia Jetson and is shown detecting the "land" signal. However, the limited power of the Jetson results in 20 PPS. This is slower than the desktop by a factor of three, yet still provides a smooth and consistent experience.



Fig. 8. Jetson Nano real-time prediction

V. CONCLUSION

CNNs allow for digital image classification, which can detect objects in images based on edges and features without any additional sensing required. This form of classification can detect signals given by humans in an intuitive way that is completely non-invasive. When implemented in drone control, a stationary form of remote operation allows the pilot to operate the drone without having to remain in view of its FPV camera. This can provide more accessibility to those that want to pilot drones but are unable to use its included control device or physically follow the drone to utilize existing CV methods. This stationary model was capable of being mapped to the user with a ResNet-18 model by removing variables such as changing backgrounds and camera positions. By using this lightweight model, a total program execution time of under ten minutes was achieved during the user mapping process. This included six minutes for executing dataset collection, followed by less than four minutes of ResNet-18 transfer learning. With its low setup time and performance capabilities on edge devices, this hand signal model presents a viable option for intuitive drone control.

REFERENCES

- [1] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, Massachusetts: The MIT Press. <https://www.deeplearningbook.org/>
- [2] S. Saha, *A comprehensive guide to convolutional neural networks*, last accessed 2022/2/12 <https://towardsdatascience.com>
- [3] K. Natarajan, T. -H. D. Nguyen and M. Mete, "Hand Gesture Controlled Drones: An Open Source Library," *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, 2018, pp. 168-175, doi: 10.1109/ICDIS.2018.00035.
- [4] *Tello Gesture Control GitHub*, <https://github.com/kinivi/tello-gesture-control>, last accessed 2022/2/12.
- [5] *Ryze Tech Tello User and SDK Manual*, <https://www.ryzerobotics.com/tello/downloads>, last accessed 2022/2/12.
- [6] *Ryze Tello Hardware Support Package for Matlab*, <https://www.mathworks.com/hardware-support/tello-drone-matlab.html>, last accessed 2022/2/12
- [7] *Tello Python SDK Github*, <https://github.com/dji-sdk/Tello-Python>, last accessed 2022/2/12
- [8] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

Terminal Sliding Mode Control (TSMC) based Cooperative Load Transportation using Multiple Drones

Prof. Dr. Madhumita Pal,
Associate Professor,
Institute of Engineering &
Management,
Kolkata, India
madhumita.pal@iemcal.com

Subhojit Das,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
streamline567@gmail.com

Sagnik Banerjee,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
sagnikbanerjee886@gmail.com

Rishav Kumar,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
rishav.iem@gmail.com

Sudhanshu Kumar Shekhar,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
shekhar2119@gmail.com

Souhardya Das,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
das.souhardya2000@gmail.com

Shreyan Ghosh,
4th year student,
Institute of Engineering &
Management,
Kolkata, India
shreyan_ghosh08@gmail.com

Abstract—Objective of this paper is to design a suitable control algorithm for cooperative transportation of an object using multiple drones based on decentralized control technique. A terminal sliding mode control (TSMC) based algorithm for motion tracking of a system of multiple Quad-rotor type Unmanned Aerial Vehicle (UAV) is proposed. TSMC is a novel class of sliding mode control giving rise to a finite time convergence of reduced order dynamics, delivering fast response, high precision, and strong robustness. A TSMC based robust cooperative control strategy is proposed for the system with multiple drones and individual control vectors are defined for controlling the combined system in presence of wind disturbances. The finite time convergence of control laws are proved using homogeneous Lyapunov function. The simulation results demonstrate the effectiveness and superior performance over conventional sliding mode-based technique for the cooperative quad rotors to transport a common payload.

Keywords—cooperative control, drones, UAV, sliding mode control, terminal sliding mode control, load transportation using drones, finite time convergence.

I. INTRODUCTION

The word ‘drone’ is very popular now a days. They are being used in various technological sectors. Recently they are being considered as future load transporting vehicles. Various MNC’s are already testing delivery using drones. But for large scale load transportation multiple drones are required and there lies the importance of cooperative control of drones. Cooperative control between robots, robot and human are already very popular. Cooperative and decentralized controlling techniques for drones are discussed in the works [1],[7],[8],[9].

There are various methods or control techniques for controlling a drone. The dynamics of drone are highly affected due to aerial disturbances and cross coupling effects due to translational and rotational motion. For this reason a robust control system is required which can handle this non linearities and dynamic stability problem. Sliding Mode Control(SMC) is quite popular among them because it is quite robust towards uncertainties and disturbances. The importance and application of sliding mode control are well described in the works [10],[12]. But sliding mode control provides asymptotic convergence and thus has very slow response. So to overcome this sliding mode control has been modified with a terminal attractor and named as Terminal SMC(TSMC). TSMC provides finite time convergence, and thus also provides fast convergence compared to normal SMC. The nature of TSMC are documented in the works [5],[6],[11],[13].

In this work, we developed a fast and robust cooperative aerial load transport system. For this decentralized cooperative control laws for a group of drones in a formation are developed where each vehicle is controlled individually. The modelling and behavior of cooperative systems are documented in the works [1], [2], [7], [8], [9]. Moore Penrose optimization theory is applied for robust modelling of cooperative system.

In this paper we have developed a TSMC based decentralized cooperative system for a system of four quad rotor type drones and a common payload. The proposed controller can successfully handle external disturbances for the cooperative system and also provides very fast convergence in finite time. The organization of this paper I as follows. In Section II,

mathematical modelling of drone is established. In Section III, the cooperative control strategy is discussed. Section IV focusses on derivation of control vectors for individual drones. In Section V, simulation results are presented in the presence of disturbance and system uncertainties. Section VI discusses the conclusions and future scopes of this project.

II. MATHEMATICAL MODELLING OF DRONE

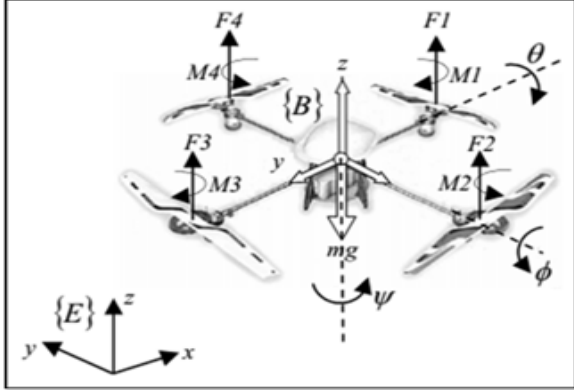


Figure 1. Configuration, Inertial and body reference frame

Quadrotor has 6 degree of freedom [x,y,z,θ,Φ,Ψ], 3 translation and 3 rotational respectively. x,y,z denote the body frame axes, Φ denotes roll angle, θ denotes pitch angle and Ψ denotes yaw angle in body frame(Drone body frame). Forces acting on body frame are to be mapped on inertial earth reference frame using Euler Rotation Matrices[3].

$$\begin{aligned}
 R_\Phi &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\Phi) & \sin(\Phi) \\ 0 & \sin(\Phi) & \cos(\Phi) \end{bmatrix} \\
 R_\theta &= \begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \\
 R_\Psi &= \begin{bmatrix} \cos(\Psi) & \sin(\Psi) & 0 \\ -\sin(\Psi) & \cos(\Psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
 R_B^E &= \begin{bmatrix} c\theta c\Psi & s\theta s\theta c\Psi - c\theta s\Psi & -c\theta s\theta c\Psi + s\theta s\Psi \\ -c\theta s\Psi & -s\theta s\theta s\Psi + c\theta c\Psi & c\theta s\theta s\Psi + s\theta c\Psi \\ s\theta & -s\theta c\theta & c\theta c\theta \end{bmatrix} \\
 \begin{bmatrix} i_E \\ j_E \\ k_E \end{bmatrix} &= R_B^E \begin{bmatrix} i_b \\ j_b \\ k_b \end{bmatrix} \quad (1)
 \end{aligned}$$

Total Lift force generated by rotors given by ,U1=F1+F2+F3+F4, where Fi are the thrust force generated by ith rotor. $F_T^b = [0 \ 0 \ U1]^T$, lift force given with respect to body frame.

Forces with respect to earth frame, $F=R_B^E * F_T^b$ (3)

Translational dynamics of quadrotor are given by following equations:

$$\begin{aligned}
 \ddot{x} &= \frac{U_1(\sin(\theta)\cos(\phi)\cos(\Psi) + \sin(\phi)\sin(\Psi))}{m} + w_x \\
 \ddot{y} &= \frac{U_1(\sin(\theta)\cos(\phi)\sin(\Psi) - \sin(\phi)\cos(\Psi))}{m} + w_y \\
 \ddot{z} &= \frac{U_1(\cos(\theta)\cos(\phi)) - mg}{m} + w_z \quad (2)
 \end{aligned}$$

$\ddot{x}, \ddot{y}, \ddot{z}$ are the acceleration of drone with respect to Earth frame due to thrust control U1. w_x, w_y, w_z are air gust disturbances, m is the mass of drone, and g is the gravitational constant (9.8 m/s²). Rotational dynamics of drone are given by following equations:

$$\begin{aligned}
 I_x \dot{p} &= (I_y - I_z)qr - J_r q\Omega + LU_2 \\
 I_y \dot{q} &= (I_z - I_x)pr - J_r p\Omega + LU_3 \\
 I_z \dot{r} &= (I_x - I_y)qp + LU_4
 \end{aligned}$$

L is the arm length of quad rotor. I_x, I_y, I_z are the moments of inertia about x, y, z axis respectively. p, q, r are the angular speeds on body frame. LU_2, LU_3, LU_4 are the moments about x, y and z axis respectively. J_r is the moment of inertia of each propeller. The inputs U_2, U_3, U_4 are related to the rotations of the quad rotor. So final rotational dynamics are given by following equations:

$$\begin{aligned}
 \dot{\Phi} &= \frac{(I_y - I_z)\theta\dot{\Psi} + LU_2 - J_r \theta\Omega}{I_x} + w_\Phi \\
 \dot{\theta} &= \frac{(I_z - I_x)\phi\dot{\Psi} + LU_3 - J_r \phi\Omega}{I_y} + w_\theta \\
 \dot{\Psi} &= \frac{(I_x - I_y)\dot{\theta}\phi + LU_4}{I_z} + w_\Psi
 \end{aligned}$$

w_Φ, w_θ, w_Ψ are the air gust disturbances. Ω is the overall residual angular speed. $\Omega = \Omega_2 + \Omega_3 - \Omega_1 - \Omega_4$, where Ω_i are the angular speed of ith rotor. Relation of Ω with control vectors are given by following,

$$\begin{bmatrix} \Omega_1^2 \\ \Omega_2^2 \\ \Omega_3^2 \\ \Omega_4^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4b} & -\frac{1}{2bl} & 0 & \frac{1}{4b} \\ \frac{1}{4b} & 0 & -\frac{1}{2bl} & -\frac{1}{4d} \\ \frac{1}{4b} & \frac{1}{2bl} & 0 & \frac{1}{4d} \\ \frac{1}{4b} & 0 & -\frac{1}{2bl} & -\frac{1}{4d} \end{bmatrix} \quad (4)$$

Here, d is thrust coefficient and b is drag coefficient. Therefore, the desired motor speed can be calculated and corresponding input can be sent to motor controllers. The drone model can be written in state space form $\dot{X} = f(X, U)$ by introducing $X = [x_1, x_2, \dots, x_{11}, x_{12}]$, where

$$[x_1 = \Phi, x_2 = \dot{\Phi}, x_3 = \theta, x_4 = \dot{\theta}, x_5 = \Psi, x_6 = \dot{\Psi}, x_7 = x, x_8 = \dot{x}, x_9 = y, x_{10} = \dot{y}, x_{11} = z, x_{12} = \dot{z}]$$

Assumption 1: It is assumed that roll, pitch and yaw angle satisfy the conditions, $\Phi \leq \pi/2, \theta \leq \pi/2, \Psi \leq \pi$ for $t > 0$.

III. MATHEMATICAL MODELLING OF COUPLED SYSTEM

Figure 2 shows a team of four quad rotor drones manipulate a cross configuration payload in three dimensions. The coordinate systems include the world frame W, and body frame B. The body frame axes are considered as the principle axes of the entire system and each quad rotor has an individual body frame Qi, where $i = \{1, 2, 3, \dots, n\}$ and n is the number of drones in the system. Here x-y-z Euler angles are used to define roll, pitch and yaw angles.

The rotation matrix B to W is given by R_B . Center of mass of ith drone is considered (x_{Qi}, y_{Qi}, z_{Qi}) .

Assumptions for quad rotor drones:

- 1)The payload is rigid and completely homogeneous, has a cross configuration structure and the center of mass of the object is intended as the center of mass of the coupled system.
- 2)The mass of payload is sufficiently small such that four drones are able to lift the object.
- 3)The payload does not flip during manipulation.

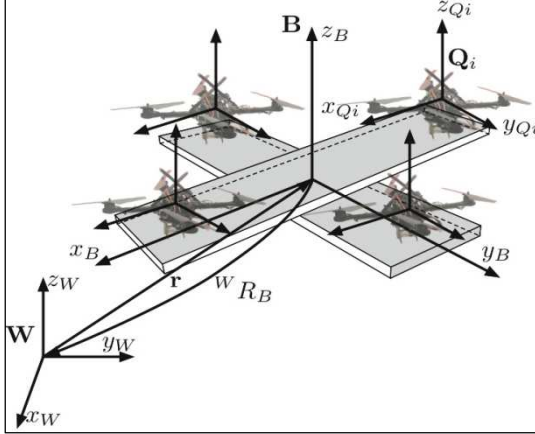


Figure 2. Scheme of payload transportation using 4 drones

The motion equation of each quad rotor is given by,

$$m_i z_{Qi}'' = -m_i g + F_{Bi} \quad (5)$$

Here m_i is the mass of i th quad rotor, g is gravity and F_{Bi} is the lift force generated by each quad rotor. Also by Newton's 2nd law

$$m_L z_{Qi}'' = -m_L g \quad (6)$$

Where m_L is the mass of the payload. Now as all the forces are used for lifting and moving toward positive Z direction, the equation of motion for the coupled system can be given as,

$$m \ddot{z} = -mg + F_B \quad (7)$$

F_B is the total lift force of all quad rotors and m is the sum of drones and object's mass, $m = m_1 + m_2 + m_3 + m_4$. The acceleration of center of mass is given by, $m \ddot{r} = -mg + R_B F_B$. r is the position vector to the center of mass of i th drone.

By considering forces and moments generated by each UAV, a relationship between the behaviour of system and the agents has to be developed. Finally the equation of motion of coupled system is obtained as,

$$\begin{bmatrix} F_B \\ M_{XB} \\ M_{YB} \\ M_{ZB} \end{bmatrix} = \sum_i \begin{bmatrix} 1 & 0 & 0 & 0 \\ y_i & \cos(\psi_i) & -\sin(\psi_i) & 0 \\ -x_i & \sin(\psi_i) & \cos(\psi_i) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_{Qi} \\ M_{xQi} \\ M_{yQi} \\ M_{zQi} \end{bmatrix}$$

Where M_{XB} , M_{YB} , M_{ZB} are the body moments in Z_{Qi} direction and F_{Qi} is the total force of each quadrotor, and M_{xQi} , M_{yQi} , M_{zQi} are the moments around each quad rotor body frame axis.

IV. COOPERATIVE CONTROL STRATEGY

This control algorithm is a decentralized control algorithm and includes three parts. In the first part of this the control fundamental vectors for n quad rotors that define a pseudo inverse matrix of the mathematical theory of optimal control using Moore - Penrose are determined. The second part consists of determining individual control vectors by TSMC. Finally in the last part the cooperative control algorithm can be established by combining the individual control vectors and fundamental control vectors.

A. Control Vectors of Cooperative Control Law

The derivation of these control vectors are well described in the works [1],[2],[7]. From there the cooperative strategy is established. By using all force vectors of the coupled system in linear system (X), four equations for four quad rotor drones can be described as,

$$[F_B, M_{XB}, M_{YB}, M_{ZB}]^T = Xu \quad (8)$$

Where $X \in R^{4 \times 4n}$, is a constant vector containing the states of translational and rotational subsystems, and $u \in R^{4n}$ includes four control input vectors for n quad rotors,

$$u = [F_{Bq1}, M_{XBq1}, M_{YBq1}, M_{ZBq1}, \dots, F_{Bqn}, M_{XBqn}, M_{YBqn}, M_{ZBqn}]^T \quad (9)$$

The cost function to minimize control inputs is given by,

$$J = \sum_i (\Lambda_{Fi} F_{qi}^2 + \Lambda_{Mxi} F_{xqi}^2 + \Lambda_{Myi} F_{yqi}^2 + \Lambda_{Mzi} M_{zqi}^2) \quad (10)$$

in which, Λ_{Fi} , Λ_{Mxi} , Λ_{Myi} , Λ_{Mzi} are the weight of each of the control inputs. By considering the features of theory of Moore - Penrose inverse, pointwise minimization of J will be done and optimal control vector can be written as,

$$u^* = \Gamma^{-2} X^T (X \Gamma^{-2} X^T)^{-1} [F_B^{des}, M_{XB}^{des}, M_{YB}^{des}, M_{ZB}^{des}]^T \quad (11)$$

$$\Gamma = [\sqrt{\Lambda_{F1}}, \sqrt{\Lambda_{Mx1}}, \sqrt{\Lambda_{My1}}, \sqrt{\Lambda_{Mz1}}, \dots, \sqrt{\Lambda_{Fn}}, \sqrt{\Lambda_{Mxn}}, \sqrt{\Lambda_{Myn}}, \sqrt{\Lambda_{Mzn}}]$$

By assuming the weight of basic control vectors,

$$\Lambda_{Fi} = \Lambda_F, \Lambda_{Mxi} = \Lambda_{Myi} = \Lambda_{Mxy} \quad (12)$$

$$\Gamma^{-2} X^T (X \Gamma^{-2} X^T)^{-1} = [U_{Fqn}, U_{Mzqn}, U_{Mxqn}, U_{Myqn}]$$

Assumption 2: While forming the control law we have assumed the X and Y coordinates of the centroid of the object and centroid of quad rotors coincide according to Figure 2. The following assumptions were made, $\sum x_i = \sum y_i = 0$, $\sum x_i y_i = 0$.

Also, since all the quad rotor drones have a common role in lifting the object off the ground and equal partnership, the body force F and yaw produced by each of them is almost equal. Thus the total force and yaw moment, roll moment and pitch moment of the coupled system are obtained as follows,

$$U_{Fqn} = \frac{1}{n} [1, 0, 0, 0, \dots, 1, 0, 0, 0]^T$$

$$U_{Mzqn} = \frac{1}{n} [0, 0, 0, 1, \dots, 0, 0, 0, 1]^T \quad (13)$$

$$U_{Mxqn} = \frac{1}{\Lambda_F \sum y_i^2 + n} \left[\frac{\Lambda_{Mxy}}{\Lambda_F} y_{q1}, C\psi_{q1}, S\psi_{q1}, 0, \dots, \frac{\Lambda_{Mxy}}{\Lambda_F} y_{qn}, C\psi_{qn}, S\psi_{qn}, 0 \right]^T$$

$$U_{Myqn} = \frac{1}{\frac{\Lambda_{Mxy}}{\Lambda_F} \sum x_i^2 + n} \left[-\frac{\Lambda_{Mxy}}{\Lambda_F} x_{q1}, -S\Psi_{q1}, C\Psi_{q1}, 0, \dots, -\frac{\Lambda_{Mxy}}{\Lambda_F} x_{qn}, -S\Psi_{qn}, C\Psi_{qn}, 0 \right]^T$$

It is important to notice to the increment of the force produced by each n quad rotor drones will lead to decrement of the individual body moments generated by them, according to the factor $\frac{\Lambda_{Mxy}}{\Lambda_F}$. In next step individual control vectors $[F_B^{des}, M_{xB}^{des}, M_{yB}^{des}, M_{zB}^{des}]^T$ are to be found and the cooperative control vectors for n collaborative drones is obtained as follows,

$$U_{coop|Qn} = U_{Fqn} F_B^{des} + U_{Mxqn} M_{xB}^{des} + U_{Myqn} M_{yB}^{des} + U_{Mzqn} M_{zB}^{des} \quad (14)$$

B. Derivation of Individual Control Vectors Based on TSMC

The In this section we will derive individual control vectors based on terminal sliding mode control. The control law consists of two parts, one continuous part which is to be derived from sliding equation and second part is discontinuous switching function which reduces chattering phenomena.

$$U_{\tilde{x}} = U_{\tilde{x}con} + U_{\tilde{x}dis} \quad (15)$$

Where $\tilde{X}=[x, y, z, \Phi, \theta, \Psi]$. The discontinuous part is given by $U_{dis} = k_D \frac{s}{|s|+\delta}$, where k_D and δ are tuning parameters [14].

Now we first derive the continuous part of control law. The sliding surface equations are given by,

$$\begin{aligned} S_x &= \dot{e}_x + \lambda e_x + b * \text{sign}(e_x)^{\frac{p}{q}} \\ S_y &= \dot{e}_y + \lambda e_y + b * \text{sign}(e_y)^{\frac{p}{q}} \\ S_z &= \dot{e}_z + \lambda e_z + b * \text{sign}(e_z)^{\frac{p}{q}} \\ S_\Phi &= \dot{e}_\Phi + \lambda e_\Phi + b * \text{sign}(e_\Phi)^{\frac{p}{q}} \\ S_\theta &= \dot{e}_\theta + \lambda e_\theta + b * \text{sign}(e_\theta)^{\frac{p}{q}} \\ S_\Psi &= \dot{e}_\Psi + \lambda e_\Psi + b * \text{sign}(e_\Psi)^{\frac{p}{q}} \end{aligned} \quad (16)$$

$$e_{\tilde{x}} = \tilde{x}_D - \tilde{x}$$

$$\text{sign}(e_{\tilde{x}})^{\frac{p}{q}} = \text{abs}(e_{\tilde{x}})^{\frac{p}{q}} * \text{sign}(e_{\tilde{x}}) \quad (17)$$

Here λ, b are positive constants and p and q are positive integers, where $p < q$. The time derivative of sliding surface is given by,

$$\dot{S}_{\tilde{x}} = \ddot{e}_{\tilde{x}} + \lambda \dot{e}_{\tilde{x}} + b * \frac{p}{q} * \text{sign}(e_{\tilde{x}}) * \text{abs}(e_{\tilde{x}})^{\frac{p}{q}-1} * \dot{e}_{\tilde{x}} \quad (18)$$

By generating sliding mode reaching law [4],

$$\dot{S}_{\tilde{x}} = -\xi_{\tilde{x}} S_{\tilde{x}} - \eta_{\tilde{x}} \text{sign}(S_{\tilde{x}}) \quad (19)$$

And replacing the acceleration equations, the control inputs are obtained as follows,

$$U_{1con} = \left[g + \left(\lambda + b * \frac{p}{q} * \text{abs}(e_z)^{\frac{p}{q}-1} * \text{sign}(e_z) \right) \dot{e}_z + \ddot{z}_D + \xi_z S_z + \eta_z \text{sign}(S_z) - w_z \right] \frac{m}{\mu_3}$$

$$U_{2con} = \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_\Phi)^{\frac{p}{q}-1} * \text{sign}(e_\Phi) \right) (\dot{\Phi}_D - \dot{\Phi}) + \ddot{\Phi}_D - \dot{\theta} \dot{\Psi} \left(\frac{I_y - I_z}{I_x} \right) + \xi_\Phi S_\Phi + \eta_\Phi \text{sign}(S_\Phi) - w_\Phi \right] \frac{I_x}{I}$$

$$U_{3con} = \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_\theta)^{\frac{p}{q}-1} * \text{sign}(e_\theta) \right) (\dot{\theta}_D - \dot{\theta}) + \ddot{\theta}_D - \Phi \dot{\Psi} \left(\frac{I_z - I_x}{I_y} \right) + \xi_\theta S_\theta + \eta_\theta \text{sign}(S_\theta) - w_\theta \right] \frac{I_y}{I}$$

$$U_{4con} = \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_\Psi)^{\frac{p}{q}-1} * \text{sign}(e_\Psi) \right) (\dot{\Psi}_D - \dot{\Psi}) + \ddot{\Psi}_D - \dot{\theta} \dot{\Phi} \left(\frac{I_x - I_y}{I_z} \right) + \xi_\Psi S_\Psi + \eta_\Psi \text{sign}(S_\Psi) - w_\Psi \right] \frac{I_z}{I}$$

$$U_{Xcon} = \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_x)^{\frac{p}{q}-1} * \text{sign}(e_x) \right) \dot{e}_x + \ddot{x}_D + \xi_x S_x + \eta_x \text{sign}(S_x) - w_x \right] \frac{m}{\mu_1}$$

$$U_{Ycon} = \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_y)^{\frac{p}{q}-1} * \text{sign}(e_y) \right) \dot{e}_y + \ddot{y}_D + \xi_y S_y + \eta_y \text{sign}(S_y) - w_y \right] \frac{m}{\mu_2} \quad (20)$$

Where μ_1, μ_2, μ_3 are function of Euler angles and are given by,

$$\mu_1 = \cos(\Phi) \sin(\theta) \cos(\Psi) + \sin(\Phi) \sin(\Psi)$$

$$\mu_2 = \cos(\Phi) \sin(\theta) \sin(\Psi) - \sin(\Phi) \sin(\Psi)$$

$$\mu_3 = \cos(\Phi) \cos(\theta) \quad (21)$$

Remarks 1: In the above equations term containing moment of inertia of rotor (J_r) has been removed. Compared to the moment of drone body, the moment of propeller is negligible.

Now for proving stability of the control law, a Lyapunov function is defined as, $V = \frac{1}{2} S^2$. For stable system the time derivative of the function must be negative definite. Now, for altitude control, $\dot{V} = S_z \dot{S}_z$ must be $\dot{V} < 0$.

$$\begin{aligned} \dot{V} &= S_z (-\xi_z S_z - \eta_z \text{sign}(S_z)) \\ \dot{V} &= -\xi_z S_z^2 - \eta_z |S_z| \end{aligned} \quad (22)$$

ξ_z and η_z are positive tuning constants, so for all non zero S_z , $\dot{V} < 0$. Now, equivalent control law consists of continuous and discontinuous part and is given by,

$$U_{1eq} = \left[g + \left(\lambda + b * \frac{p}{q} * \text{abs}(e_z)^{\frac{p}{q}-1} * \text{sign}(e_z) \right) \dot{e}_z + \ddot{z}_D + \xi_z S_z + \eta_z \text{sign}(S_z) - w_z \right] \frac{m}{\mu_3} + k_{Dz} \frac{S_z}{|S_z| + \delta}$$

$$U_{1eq} = U_{1con} + U_{1dis} \quad (23)$$

Now for designing discontinuous signal, we take a Lyapunov function $V1 = \frac{1}{2} S_z^2$. For stable control law,

$$\begin{aligned} \dot{V}1 &= S_z \dot{S}_z \\ \dot{V}1 &= S_z (\dot{e}_z + \lambda e_z + b * \frac{p}{q} * \text{sign}(e_z) * \text{abs}(e_z)^{\frac{p}{q}-1} * \dot{e}_z) \\ \dot{V}1 &= S_z \left(\ddot{z}_D - \frac{U_{1eq} * \mu_3 - mg}{m} - w_z + \lambda e_z + b * \frac{p}{q} * \text{sign}(e_z) * \text{abs}(e_z)^{\frac{p}{q}-1} * \dot{e}_z \right) \end{aligned}$$

$$\dot{V}_1 = S_z \left(-k_{Dz} * \frac{S_z}{|S_z| + \delta} * \frac{\mu_3}{m} \right) \quad (24)$$

So, for $k_D > 0$, for all non zero S_z , $\dot{V} < 0$. The same stability can be proved for altitude control also [14]. So finally the individual control vectors are given as follows,

$$\begin{aligned} F_{Bn}^{des} &= \left[g + \left(\lambda + b * \frac{p}{q} * \text{abs}(e_{zn})^{\frac{p-1}{q}} * \text{sign}(e_{zn}) \right) \dot{e}_z + \ddot{z}_D + \xi_z S_{zn} \right. \\ &\quad \left. + \eta_z \text{sign}(S_{zn}) - w_z \right] \frac{m}{\mu_{3n}} + k_{Dz} \frac{S_{zn}}{|S_{zn}| + \delta} \\ M_{XBn}^{des} &= \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_{\phi n})^{\frac{p-1}{q}} * \text{sign}(e_{\phi n}) \right) (\dot{\Phi}_D - \dot{\Phi}_n) + \ddot{\Phi}_D \right. \\ &\quad \left. - \dot{\theta}_n \dot{\Psi}_n \left(\frac{I_y - I_z}{I_x} \right) + \xi_\phi S_{\phi n} + \eta_\phi \text{sign}(S_{\phi n}) - w_\phi \right] \frac{I_x}{I} \\ &\quad + k_{D\phi} \frac{S_{\phi n}}{|S_{\phi n}| + \delta} \\ M_{YBn}^{des} &= \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_{\theta n})^{\frac{p-1}{q}} * \text{sign}(e_{\theta n}) \right) (\dot{\theta}_D - \dot{\theta}_n) + \ddot{\theta}_D \right. \\ &\quad \left. - \dot{\Phi}_n \dot{\Psi}_n \left(\frac{I_z - I_x}{I_y} \right) + \xi_\theta S_{\theta n} + \eta_\theta \text{sign}(S_{\theta n}) - w_\theta \right] \frac{I_y}{I} \\ &\quad + k_{D\theta} \frac{S_{\theta n}}{|S_{\theta n}| + \delta} \\ M_{ZBn}^{des} &= \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_{\psi n})^{\frac{p-1}{q}} * \text{sign}(e_{\psi n}) \right) (\dot{\Psi}_D - \dot{\Psi}_n) + \ddot{\Psi}_D \right. \\ &\quad \left. - \dot{\theta}_n \dot{\Phi}_n \left(\frac{I_x - I_y}{I_z} \right) + \xi_\psi S_{\psi n} + \eta_\psi \text{sign}(S_{\psi n}) - w_\psi \right] \frac{I_z}{I} \\ &\quad + k_{D\psi} \frac{S_{\psi n}}{|S_{\psi n}| + \delta} \\ U_{Xe q} &= \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_{xn})^{\frac{p-1}{q}} * \text{sign}(e_{xn}) \right) \dot{e}_{xn} + \ddot{X}_D + \xi_x S_{xn} \right. \\ &\quad \left. + \eta_x \text{sign}(S_{xn}) - w_x \right] \frac{m}{\mu_{1n}} + k_{Dx} \frac{S_{xn}}{|S_{xn}| + \delta} \\ U_{Ye q} &= \left[\left(\lambda + b * \frac{p}{q} * \text{abs}(e_{yn})^{\frac{p-1}{q}} * \text{sign}(e_{yn}) \right) \dot{e}_{yn} + \ddot{Y}_D + \xi_y S_{yn} \right. \\ &\quad \left. + \eta_y \text{sign}(S_{yn}) - w_y \right] \frac{m}{\mu_{2n}} + k_{Dy} \frac{S_{yn}}{|S_{yn}| + \delta} \quad (25) \end{aligned}$$

Where,

$$\begin{aligned} \mu_1 &= \cos(\Phi_n) \sin(\theta_n) \cos(\Psi_n) + \sin(\Phi_n) \sin(\Psi_n) \\ \mu_2 &= \cos(\Phi_n) \sin(\theta_n) \sin(\Psi_n) - \sin(\Phi_n) \sin(\Psi_n) \\ \mu_3 &= \cos(\Phi_n) \cos(\theta_n) \end{aligned} \quad (26)$$

C. Final Cooperative Control Vectors

Now combining the control fundamental vectors in eqns (13) with individual control vectors in eqns (25), the derived cooperative control vectors are given by following,

$$\begin{aligned} &\begin{pmatrix} U_{coop|q1} \\ U_{coop|q2} \\ U_{coop|q3} \\ U_{coop|q4} \end{pmatrix} \\ &= \begin{pmatrix} U_{Fq1} F_{Bq1}^{des} + U_{MXq1} M_{XBq1}^{des} + U_{MYq1} M_{YBq1}^{des} + U_{MZq1} M_{ZBq1}^{des} \\ U_{Fq1} F_{Bq2}^{des} + U_{MXq1} M_{XBq2}^{des} + U_{MYq1} M_{YBq2}^{des} + U_{MZq1} M_{ZBq2}^{des} \\ U_{Fq1} F_{Bq3}^{des} + U_{MXq1} M_{XBq3}^{des} + U_{MYq1} M_{YBq3}^{des} + U_{MZq1} M_{ZBq3}^{des} \\ U_{Fq1} F_{Bq4}^{des} + U_{MXq1} M_{XBq4}^{des} + U_{MYq1} M_{YBq4}^{des} + U_{MZq1} M_{ZBq4}^{des} \end{pmatrix} \quad (26) \end{aligned}$$

To test the robustness of controller against external disturbances, we simulate wind effects as sinusoidal velocity function against the motion of drones. The wind velocity $w(t)$ is given as follows,

$$\begin{bmatrix} w_x(t) \\ w_y(t) \\ w_z(t) \end{bmatrix} = \begin{bmatrix} 0.5 \sin\left(\frac{\pi(t-20)}{20}\right) + 0.2 \sin\left(\frac{\pi(t-20)}{10}\right) + 0.06 \sin\left(\frac{\pi(t-20)}{5}\right) \\ 0.5 \sin\left(\frac{\pi(t-30)}{20}\right) + 0.2 \sin\left(\frac{\pi(t-30)}{10}\right) + 0.06 \sin\left(\frac{\pi(t-30)}{5}\right) \\ 0.5 \sin\left(\frac{\pi(t-40)}{20}\right) + 0.2 \sin\left(\frac{\pi(t-40)}{10}\right) + 0.06 \sin\left(\frac{\pi(t-40)}{5}\right) \end{bmatrix} \quad (27)$$

D. Finite Time Convergence of TSMC

TSMC sliding surface equation is defined as ,

$$s = \dot{e} + \lambda e + b * \text{sig}(e)^{\frac{p}{q}} \quad (28)$$

Where λ, b are positive constants and p and q are positive integers such that $p < q$.

$$\dot{e} = -\lambda e - b * \text{abs}(e)^{\frac{p}{q}} * \text{sign}(e) \quad (29)$$

Considering t_2 the time from initial non zero state error $e(t_0)$ to $e(t_2)=0$,

$$\int_{e(t_0)}^{e(t_2)} \frac{de}{-\lambda e - b * \text{abs}(e)^{\frac{p}{q}} * \text{sign}(e)} = \int_{t_0}^{t_2} dt$$

$$t_2 = \frac{1}{\lambda \left(1 - \frac{p}{q}\right)} * \ln \left(\frac{\lambda * \text{abs}(e(t_0))^{(1-\frac{p}{q})} + b}{b} \right) \quad (30)$$

Here t_2 does not tend to infinite, so convergence is proved.

E. Derivation of Individual Control Vectors Based on SMC

For comparing the results of TSMC with conventional SMC, we also define SMC control vectors. The sliding surface of SMC is given by,

$$S_{\bar{x}} = \dot{e}_{\bar{x}} + \lambda e_{\bar{x}} \quad (31)$$

The control vectors will be given by,

$$\begin{aligned} F_{Bn}^{des} &= [g + \lambda \dot{e}_z + \ddot{z}_D + \xi_z S_{zn} + \eta_z \text{sign}(S_{zn}) - w_z] \frac{m}{\mu_{3n}} + k_{Dz} \frac{S_{zn}}{|S_{zn}| + \delta} \\ M_{XBn}^{des} &= \left[\lambda (\dot{\Phi}_D - \dot{\Phi}_n) + \ddot{\Phi}_D - \dot{\theta}_n \dot{\Psi}_n \left(\frac{I_y - I_z}{I_x} \right) + \xi_\phi S_{\phi n} + \eta_\phi \text{sign}(S_{\phi n}) \right. \\ &\quad \left. - w_\phi \right] \frac{I_x}{I} + k_{D\phi} \frac{S_{\phi n}}{|S_{\phi n}| + \delta} \\ M_{YBn}^{des} &= \left[\lambda (\dot{\theta}_D - \dot{\theta}_n) + \ddot{\theta}_D - \dot{\Phi}_n \dot{\Psi}_n \left(\frac{I_z - I_x}{I_y} \right) + \xi_\theta S_{\theta n} + \eta_\theta \text{sign}(S_{\theta n}) \right. \\ &\quad \left. - w_\theta \right] \frac{I_y}{I} + k_{D\theta} \frac{S_{\theta n}}{|S_{\theta n}| + \delta} \\ M_{ZBn}^{des} &= \left[\lambda (\dot{\Psi}_D - \dot{\Psi}_n) + \ddot{\Psi}_D - \dot{\theta}_n \dot{\Phi}_n \left(\frac{I_x - I_y}{I_z} \right) + \xi_\psi S_{\psi n} + \eta_\psi \text{sign}(S_{\psi n}) \right. \\ &\quad \left. - w_\psi \right] \frac{I_z}{I} + k_{D\psi} \frac{S_{\psi n}}{|S_{\psi n}| + \delta} \\ U_{Xe q} &= [\lambda \dot{e}_{xn} + \ddot{X}_D + \xi_x S_{xn} + \eta_x \text{sign}(S_{xn}) - w_x] \frac{m}{\mu_{1n}} + k_{Dx} \frac{S_{xn}}{|S_{xn}| + \delta} \\ U_{Ye q} &= [\lambda \dot{e}_{yn} + \ddot{Y}_D + \xi_y S_{yn} + \eta_y \text{sign}(S_{yn}) - w_y] \frac{m}{\mu_{2n}} \\ &\quad + k_{Dy} \frac{S_{yn}}{|S_{yn}| + \delta} \quad (32) \end{aligned}$$

V. SIMULATION RESULTS

In this section the simulation results and the data used for simulation are given. The simulations are carried out in MATLAB and from the output the effectiveness of controller

is verified. The system parameter values are listed in table 1 and simulation parameter values are listed on table 2.

TABLE I. VALUES OF SYSTEM PARAMETERS

System Parameters	Values
Mass of each Drone	1.8 kg
Mass of Payload	1.8 kg
Arm length of Drone(<i>l</i>)	0.42m
Inertia about X axis (<i>I_x</i>)	2.16x10 ⁻³ kg. m ²
Inertia about Y axis (<i>I_y</i>)	2.16x10 ⁻³ kg. m ²
Inertia about Z axis (<i>I_z</i>)	0.33x10 ⁻³ kg. m ²

TABLE II. VALUES OF SIMULATION PARAMETERS

Simulation Parameters	Value for x,y,z tracking	Value for Φ, θ, Ψ tracking
b	5	2
ξ	5	20
η	0.1	0.1
<i>k_D</i>	100	20
p/q	5/7	5/7
λ	2	10
$\frac{\Delta_{Mxy}}{\Delta_F}$	1/400	1/400
δ	0.8	0.8
η	0.1	0.1
<i>k_D</i>	100	20
p/q	5/7	5/7
λ	2	10
Λ_{Mxy}/Λ_F	1/400	1/400
δ	0.8	0.8

Remarks 1: Since while modelling of cooperative system, it was assumed that all the quad rotors have common role in lifting the payload thus mass of payload is divided in four and added to each drone weight while carrying out the simulation. We have carried out point to point tracking in our simulation. The desired values and the initial values are considered as follows, Initial Values

Quad rotor	x	y	z	Φ	θ	Ψ
Quad rotor 1	0.1	0.0	0.0	0.35	0.4	0.65
Quad rotor 2	2.5	0.1	0.0	0.25	0.45	0.6
Quad rotor 3	2.6	2.5	0.0	0.3	0.5	0.55
Quad rotor 4	0.0	2.6	0.0	0.37	0.55	0.5

TABLE III. DESIRED VALUES(FOR ALL DRONES)

x	y	z	Φ	θ	Ψ
2	2	2	0	0	0

Figure 3, Figure 5, Figure 7 shows x, y and z displacements respectively and Figure 4, Figure 6, Figure 8 shows the errors in x, y and z direction respectively. Figure 9, Figure 11, Figure 13 shows Φ, θ and Ψ rotations respectively and Figure 10, Figure 12, Figure 14 shows the errors in roll, pitch and yaw angles respectively. From the graphs it is clearly observable that all the drone converge faster to final value with TSMC than SMC. Thus TSMC has lower settling time compared to SMC, thus it is proved to be a fast responsive system.

VI. CONCLUSION AND FUTURE WORKS

We discussed the problem of cooperative load transportation using multiple drones in three dimensions. A decentralized control law based on TSMC has been proposed. From the simulation results it is observed that system of drones has converged to desired values in finite time and thus a fast responsive system. The responses are faster in TSMC compared to conventional SMC. Also the system successfully maintained stability in the presence of wind disturbances. So the robustness of controller is verified well. The cooperative system has successfully exhibited point to point tracking. The chattering of system has been reduced to almost zero by using proper tuning parameters and discrete switching law. In future works we will be proceeding with variable point tracking and more enhanced load sharing with heavy loads.

ACKNOWLEDGMENT

We thank our mentor Prof. Dr. Madhumita Pal (Associate Professor, Department of Electrical Engineering, Institute of Engineering & Management, Kolkata) for her continuous guidance and assistance through out our project work.

REFERENCES

- [1] Daniel Mellinger, Michael Shomin, Nathan Michael, and Vijay Kumar, "Cooperative Grasping and Transport Using Multiple Quadrotors", Springer-Verlag Berlin Heidelberg 2013.
- [2] Raziye Babaie and Amir Farhad Ehyaei, "Robust optimal motion planning approach to cooperative grasping and transporting using multiple UAVs based on SDRE", Transactions of the Institute of Measurement and Control 1–18 The Author(s) 2016 Reprints and permissions: sagepub.co.uk/journalsPermissions.nav.
- [3] Madhumita Pal, Avik Paul, Maitreyee Banerjee, Swetaleena Guha, "TERMINAL SLIDING MODE CONTROL TECHNIQUE FOR FLIGHT CONTROL DESIGN OF QUADROTOR", (Unpublished).
- [4] Dan Ye , Qiang Wang , Jie Li, "Fast Sliding Mode Tracking Control of Micro Quadrotor UAV", 978-1-4673-9714-8/16/, 2016 IEEE
- [5] Guangyao Yu, Yang Chen, Zhihuan Chen, Huaiyu Wu, and Lei Cheng, "Design of terminal sliding mode controller for a quadrotor UAV with disturbance observer", Proceedings of the 39th Chinese Control Conference July 27-29, 2020, Shenyang, China.
- [6] Yurong Li, Yi Qin, Fujie Wang, Fang Guo, John T. W. Yeow, "Global Fast Terminal Sliding Mode Control for a Quadrotor UAV", 978-1-7281-5169-4/20/, 2020 IEEE
- [7] Babaie, A. F. Ehyaei, "Cooperative Control of Multiple quadrotors for Transporting a Common Payload", R. Babaie, A. F. Ehyaei, AUT J. Model. Simul., 50(2) (2018) 147-156, DOI: 10.22060/miscj.2018.14252.5100.

- [8] Hossein Rastgoftar and Ella M. Atkins, Senior Member, IEEE, "Cooperative Aerial Payload Transport Guided by an In Situ Human Supervisor", 1063-6536 © 2018 IEEE.
- [9] Yu Heng Tan, Shupeng Lai, Kangli Wang and Ben M. Chen, "Cooperative Heavy Lifting Using Unmanned Multi-Agent Systems", 2018 IEEE 14th International Conference on Control and Automation (ICCA) June 12-15, 2018. Anchorage, Alaska, USA.
- [10] Alejandro P Gómez, William Oswaldo, "Sliding Mode Control: An Approach to Control a Quadrotor" 2015 Asia-Pacific Conference on Computer Aided System Engineering, pp. 314-319, 2015.
- [11] Hamid Hassani1, Anass Mansouri, Ali Ahaitouf, "Robust autonomous flight for quadrotor UAV based on adaptive nonsingular fast terminal sliding mode control", 2019 ISA Transactions, Published by Elsevier Ltd., pp. 290-304, 2019.
- [12] Lenaick Besnarda , Yuri B. Shtessel , Brian Landruma, "Quadrotor vehicle control via sliding mode controller driven by sliding mode disturbance observer", 0016-0032/\$32.00 & 2011 The Franklin Institute. Published by Elsevier Ltd.
- [13] S. T. Venkataraman, S. Gulati , "Control of Nonlinear Systems Using Terminal Sliding Modes", JPL/California Institute of Technology 4800 Oak Grove Drive, Pasadena, CA 91109. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.
- [14] Marco Herrera, Alejandro P. Gómez, William Chamorro, Oscar Camacho, "Sliding Mode Control: An approach to Control a Quadrotor", 2015 Asia-Pacific Conference on Computer Aided System Engineering, 978-1-4799-7588-4/15© 2015 IEEE.

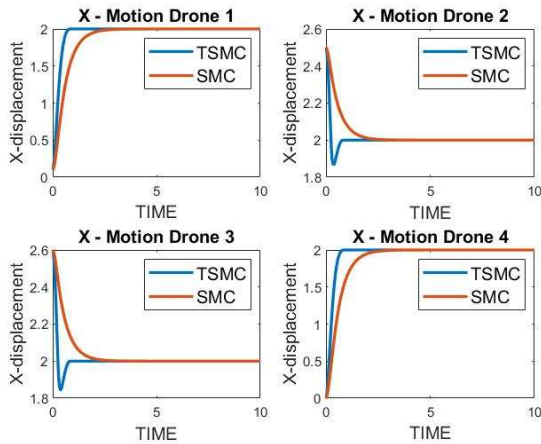


Figure 3. X motion tracking of drones

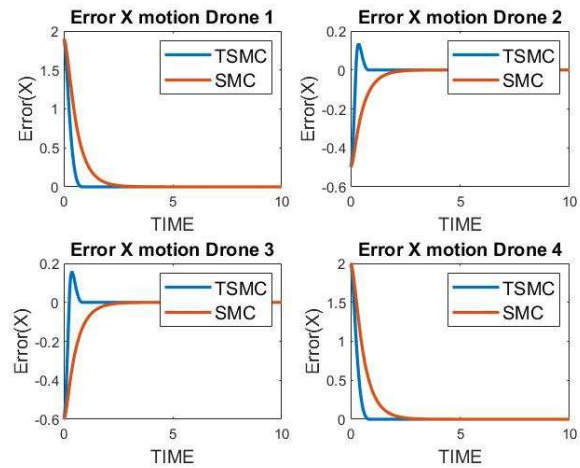


Figure 4. Error in X displacement

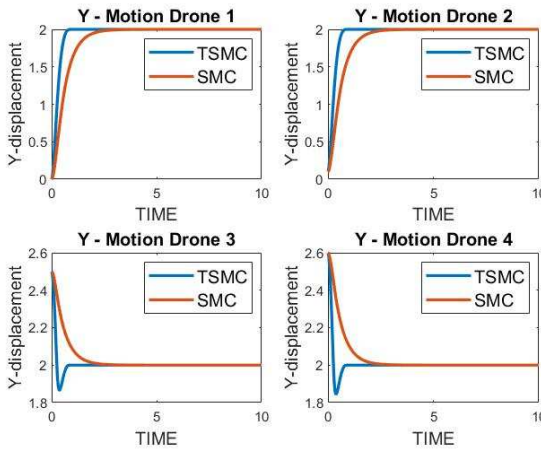


Figure 5. Y motion tracking of drones

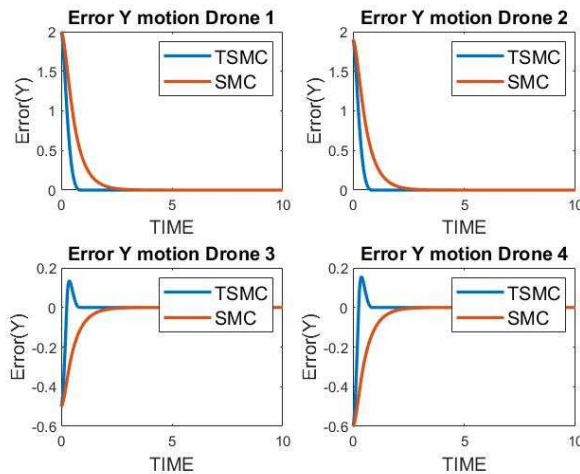


Figure 6. Error in Y displacement

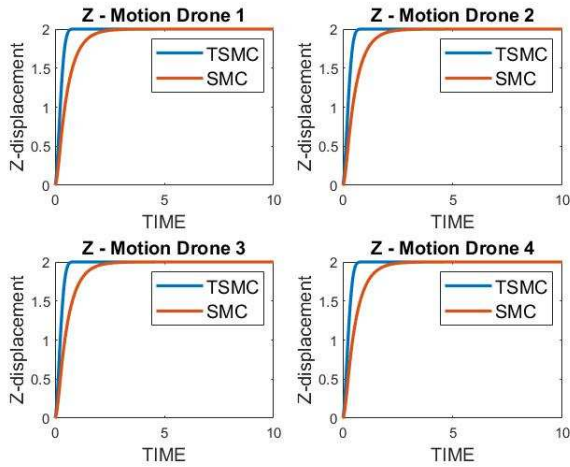


Figure 7. Z motion tracking of drones

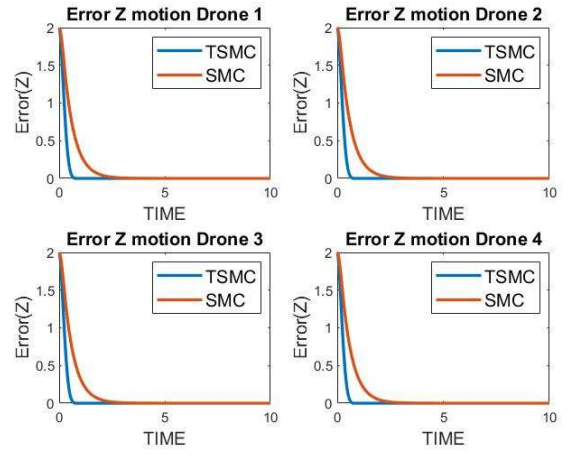


Figure 8. Error in Z displacement

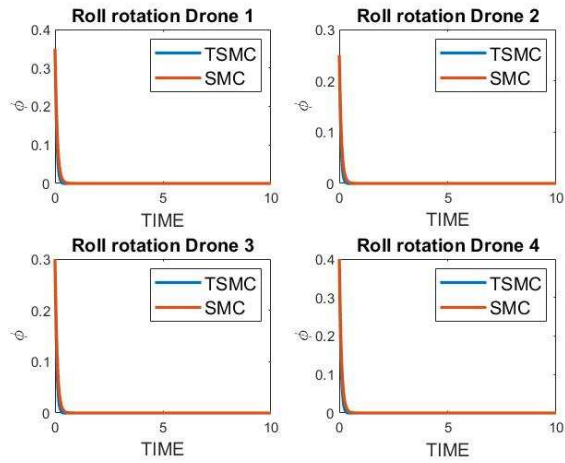


Figure 9. Roll rotation tracking of drones

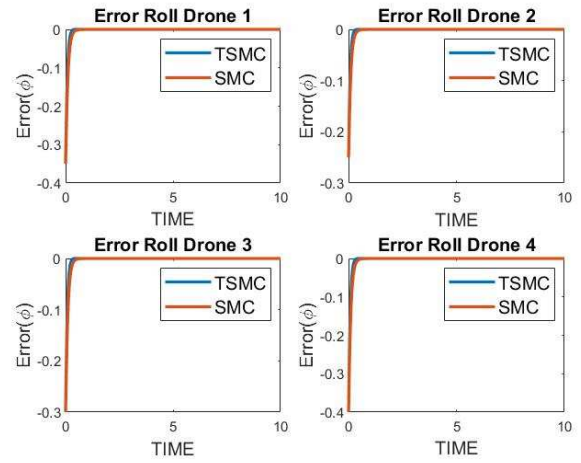


Figure 10. Error in Rolling

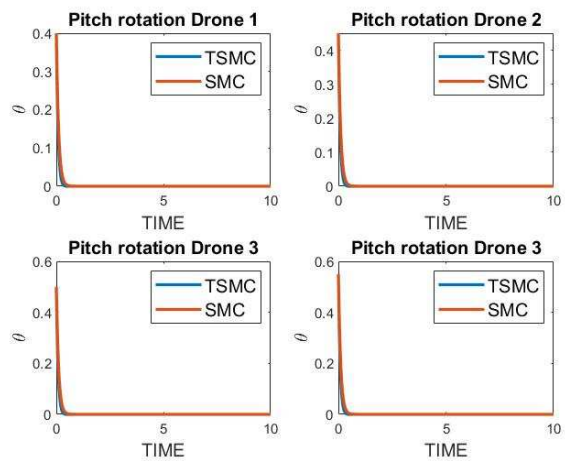


Figure 11. Pitch rotation tracking of drones

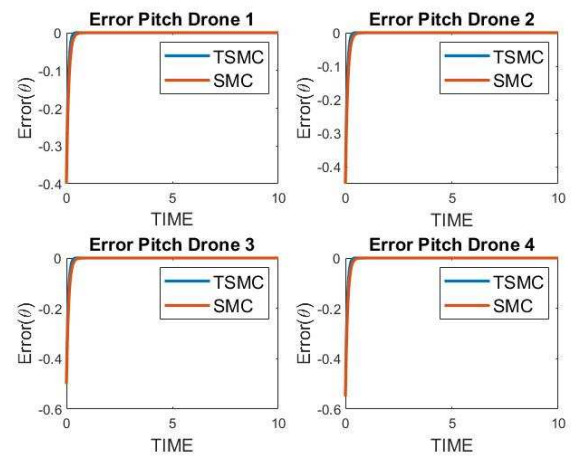


Figure 12. Error in Pitching

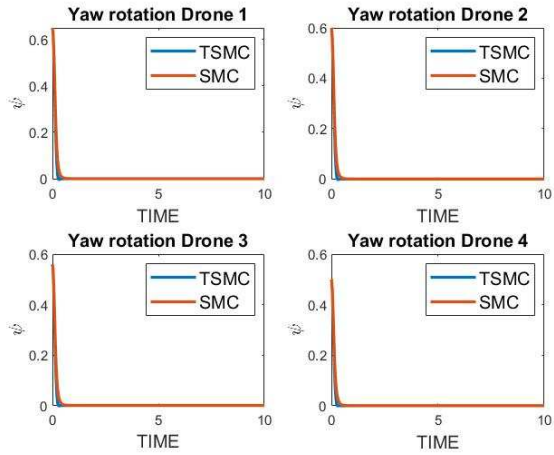


Figure 13. Yaw rotation tracking of drones

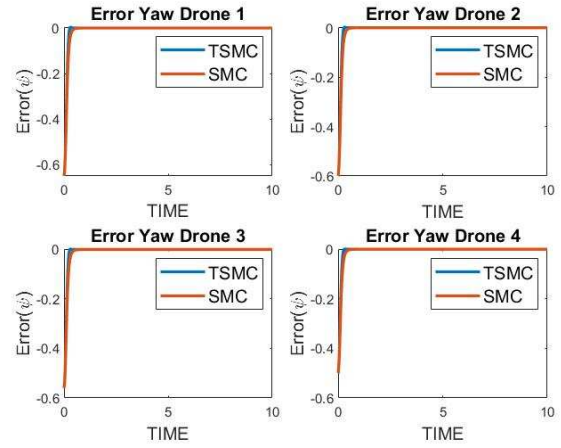


Figure 14. Error in Yawing

Cooperative transportation using SMC

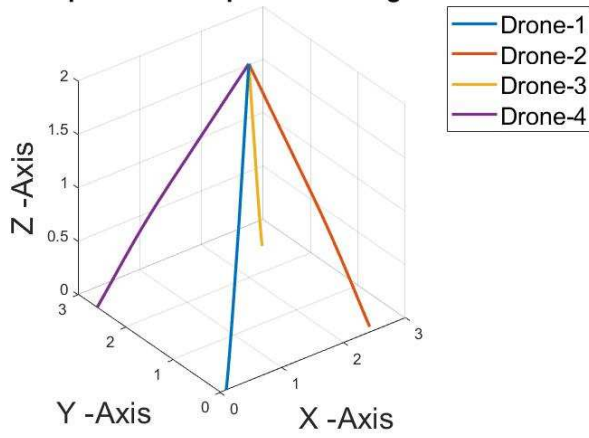


Figure 15. 3D motion tracking of drones using SMC

Cooperative transportation using TSMC

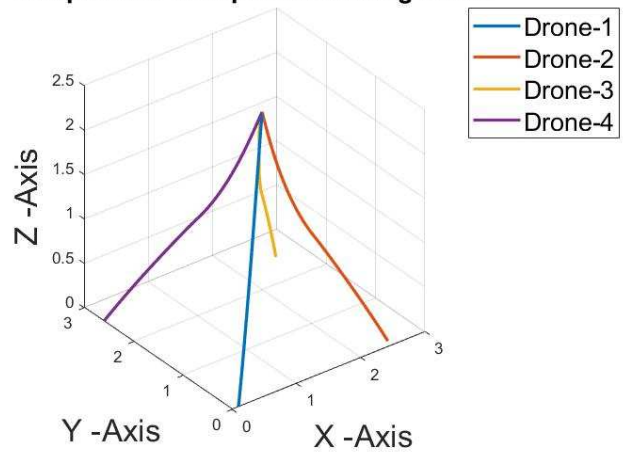


Figure 16. 3D motion tracking of drones using TSMC

Mechanical and Electronic Design of a Prototype of a Modular Exoskeleton for Lower-Limbs

Yerson Taza-Aquino
 Department of Mechatronics Engineering
 Universidad Continental
 Huancayo, Perú
 74239368@continental.edu.pe

Deyby Huamanchahua
 Department of Mechatronics Engineering
 Universidad Continental
 Huancayo, Perú
 dhuamanchahua@continental.edu.pe

Abstract—In many countries, the rehabilitation of partial disability of the lower limbs performs the process traditionally due to the high cost of implementing physiotherapy and rehabilitation centers with robotic devices. Therefore, the primary motivation of this work is to propose the first design of an exoskeleton with adjustable links that can be adapted depending on the height of the user, muscle sensors (EMG) and position are used to achieve a better response of the patient's intention of movement and thus achieve rehabilitation of the legs. The exoskeleton in question was designed using the VDI 2206 methodology, and this work presents a proposal for mechanical and electronic design with the ability to withstand the user's weight. A study of stress analysis and simulation of the electronic circuit was carried out. The electronic circuit was simulated in the Proteus software, where the correct interaction of the sensors with the motors is achieved. The results obtained show that the design of the proposed exoskeleton manages to support the weight of a person of 75 Kg with a maximum height of 170 cm. These results were obtained after being subjected to the design of the exoskeleton to the stress analysis in the SolidWorks software. Another feature of the exoskeleton design is its low weight because the material chosen is aluminum alloy 6061 T-6, which can withstand all stress tests.

Keywords— *exoskeleton, muscle sensors, rehabilitation.*

I. INTRODUCTION

In recent decades, the area of robotics has been going through a process of considerable evolution in scope and dimension [1]. Initially, the creation of robots focused on applications in industrial processes to replace labor in cyclical tasks; however, current robotic devices have significant interaction with humans [2].

Today's robots are increasingly found in the environment of the activities carried out by human beings. Robotics is in development and growth and is being applied in various branches of study such as medicine, military forces, mining, household services, etc. [3]. In medicine, mechanical structures such as exoskeletons are being created to contribute to and help human limbs in different tasks [4]. Exoskeletons are mechanical structures attached to human limbs to amplify or increase the user's strength. The main applications are in industries, military technology, and medicine [5]. The exoskeleton can be applied to perform rehabilitation therapies

on the upper and lower limbs due to a disease or accident that produces a temporary loss of muscle activity [6].

The development of exoskeletons for the lower limb began in the 60s; the reasons for their development were muscle weakness. This robotic mechanism contributed to patients with spinal cord injury recovering their gait [7].

The main reasons for disabilities are muscle diseases and diseases that affect the nervous system; in both cases, you must opt for a treatment that favors the recovery of the affected parts [8]. It is essential to perform rehabilitation therapy when suffering from temporary upper and lower limbs [9]. Research trends show that failure to perform good rehabilitation therapy results in the ultimate loss of limbs [10]. Based on the literature, exoskeletons are developed primarily for military, industrial, and medical purposes [11].

For this reason, this research aims to design and control an exoskeleton that manages to perform the hip, knee, and ankle joints, achieving 3 DoF on each side. The exoskeleton for rehabilitation of lower limbs may be used by patients whose height varies from 150 centimeters to 172 centimeters having the ability to regulate the lengths of the links.

II. METHODOLOGY

The exoskeleton design for the partial rehabilitation of lower limbs comprises the mechanical part and the electronic part. It is crucial to achieving a good selection of components and materials suitable to be subjected to simulations.

The VDI 2206 design methodology is taken as a reference, as shown in Fig. 1. Stage 1 is the technical information, which helps obtain the process of operation of the robotic system, where each stage is shown in detail in a flowchart to obtain a didactic and easy-to-understand process. Stage 2 details the design of the robotic system where the structure of the exoskeleton is shown with the respective components that make it up. Stage 3 details the electronic design showing the sensors' interaction with the motors controlled by an Arduino Uno® board, detailing all the components that make it up in Table 3.

Finally, in Stage 4, the final design is shown where the mechanical and electronic parts of the entire exoskeleton system are joined.

Wiper Actuator Model GP-WD3	
Mark	GP
Model	GP-WD63
Voltage	12V
Type of motor	DC
Power	20-200W
Engine noise	Low noise (50dB)
Relation of reduction	1:20; 1:24; 1:60; 1:68; 1:75
No-load rpm	90 rpm
Loaded rpm	60 rpm

TABLE II. TECHNICAL CHARACTERISTICS OF THE ACTUATOR

Note: the features were extracted from the actuator datasheet.

The motors that make up the exoskeleton will be controlled by the Arduino Uno® controller, which will be programmed to process the signals recorded by the sensors. In Fig. 5, the flowchart is shown with all the process blocks that will make up the operation of the exoskeleton, where the procedures that will be performed are detailed. The robotic system will be operated by an on and off button, and the power of the system will be from a 12 V source when the user performs the movements of the legs in the EMG sensor and positions the potentiometer. Subsequently, the sensors will send data to the controller to process and operate the corresponding motor; when the user does not generate any intention of moving, the device will not move; the design proposal is for patients who count partial movements of the legs.

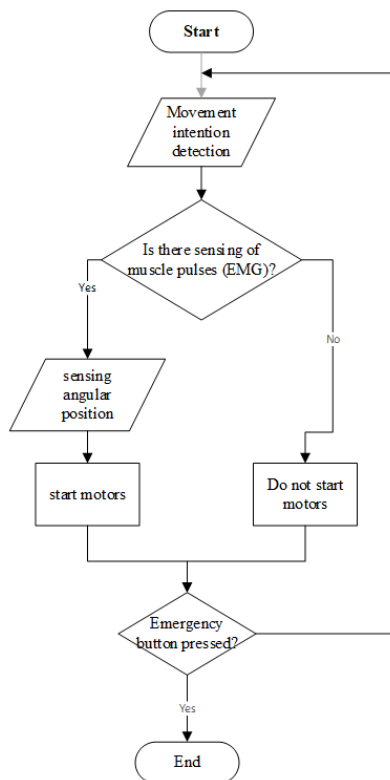


Fig. 5. Flowchart of the exoskeleton system

B. Physical design

To design the exoskeleton, the human body's standardized proportions were considered a function of the height, whereas the measurements in Fig. 6 were considered a reference.

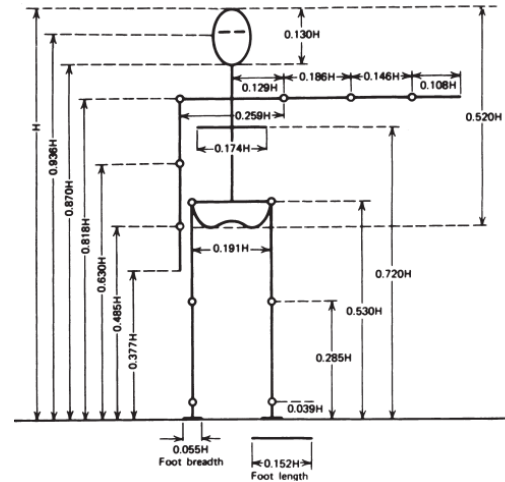


Fig.6. Standardized proportions of the female and male body as a function of height. Source: D. A. Winter, "Biomechanics and motor control of human movement", Canada: Waterloo, [2009].

Based on these standardized proportions, the prototype of the lower limb exoskeleton was designed as shown in Fig. 7, the comparison with the close links for people of a height of 150 cm and the stretched links for people with a height of 172 cm.

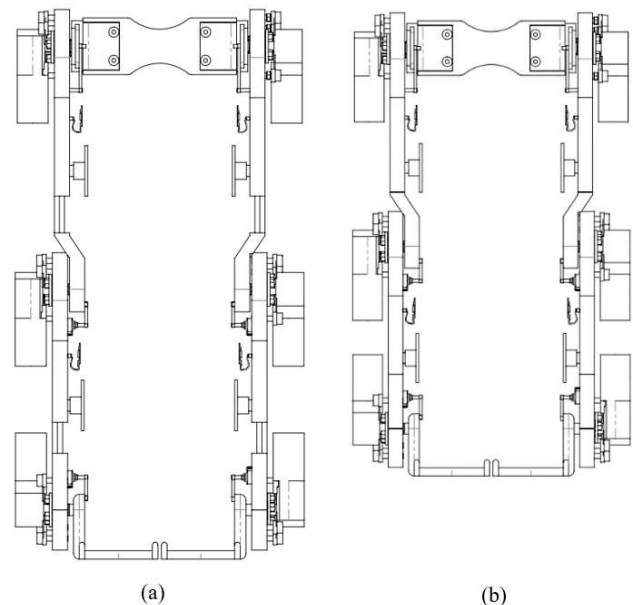


Fig. 7. 3D model of a lower limb exoskeleton (a) Exoskeleton with stretched links (for patients 172cm), (b) Exoskeleton with close links (for patients 150cm).

Fig. 8 shows the location of the motors, EMG sensors, and position potentiometers that make up the prototype of the exoskeleton structure.

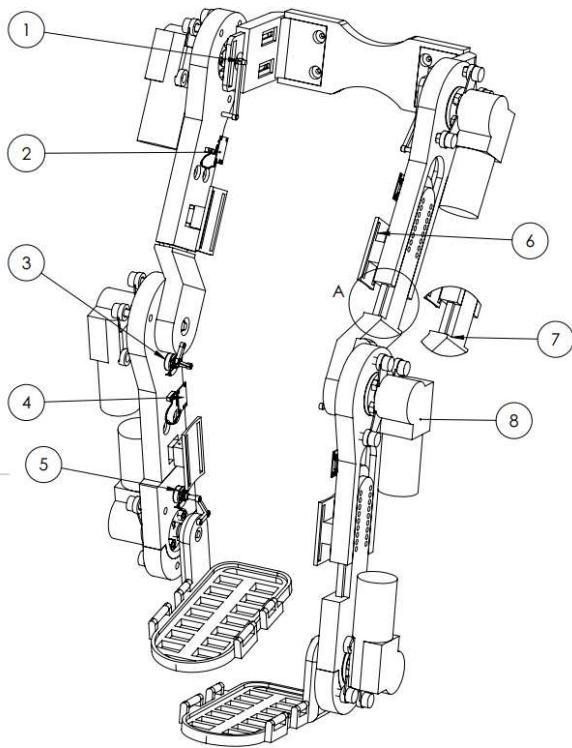


Fig. 8. Structure of the prototype design of the exoskeleton.

- (1-3-5) The position potentiometers will be in each joint (hip, knee, ankle) in total, and there are six potentiometers. These sensors will help to have better control of each motor; in this way, the independent drives are achieved depending on the force that the patient needs in each joint.
- (2-4) The EMG sensors will be located on the thigh and leg to capture the muscle pulses of the lower limbs, detecting the intention of movement; the controller will process these signals to generate the activation of the appropriate actuator.
- (6) The bases of the fasteners are variable so that they have better adaptability in the user's legs.
- (7) The links are adjustable, having the ability to cover patients between the sizes of 150 cm to 172 cm.
- (8) The exoskeleton design has six motors with their respective gearboxes, which generate a force.

C. Electronic design

The electronic design is considered a circuit connected to a 16x2 LCD screen, which allows programming a series of routines to perform the simulation in the Proteus® software Arduino libraries were linked. For the control of this screen, an Arduino Nano board and a 5V power supply were used, as can be seen in Fig. 9.

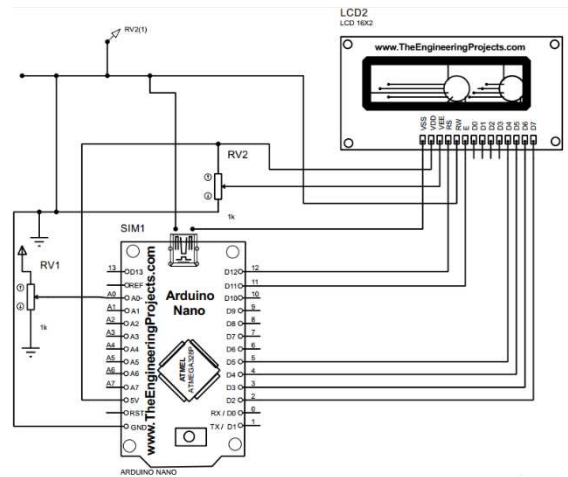


Fig. 9. LCD screen circuit.

For an electronic system of the exoskeleton prototype, it is proposed to use Arduino Uno® plates, taking into account that the movements of both legs of the design are of equal operation; that is, the circuit that is presented in Fig. 10 will be implemented in each leg of the exoskeleton, in the figure two EMG sensors are shown that will be located in the thigh and calf of each leg the signals of analog inputs of the thigh is in the A3 and calf port on the A5 port, these input signals will allow the intention of the patient's movement to be processed.

The position potentiometers enter the analog input signals through ports A0 (ankle), A1 (knee), and A2 (hip). These signals allow the exoskeleton to have better control of the motors. Each motor has the L293D driver, which inside is built the circuit of the bridge H, this allows to drive the engine in both directions anti-clockwise rotation and hourly rotation, for an exemplary operation of the motors is powered by a 12 V source. Table III shows all the electronic components that make up the system.

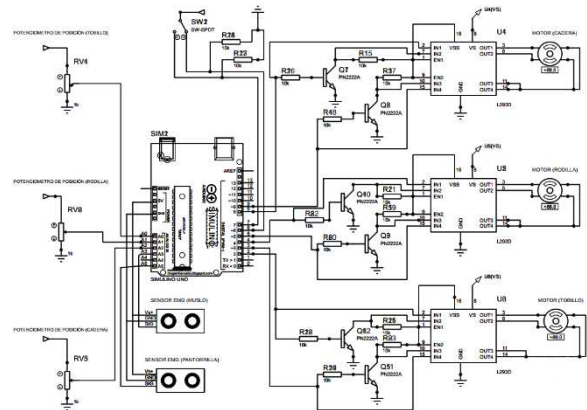


Fig. 10. Actuator control circuit.

TABLE III LIST OF ELECTRONIC COMPONENTS

Components	Description	Amount
Arduino Uno®	Motor control card	1
Arduino Nano	LCD Display Control Card	1

Sensor EMG	The capture of the muscular pulses of the leg	4
Position potentiometer	Motor motion control	6
Driver L293D	Control of the clockwise and anti-clockwise rotation of the motors	6
Screen LCD	Control and scheduling of routines	1

III. RESULTS

A. Analysis of the pieces

To verify that the design of the parts and the material are suitable, the simulation of finite element analysis is carried out with the SolidWorks program, where the essential parts of the exoskeleton structure will be subjected to critical loads. For the choice of material, it must be considered that it is lightweight, resistant, and commercially accessible; under these criteria, the aluminum alloy 6061-T6 is selected. Table IV shows all the technical specifications acquired from the SolidWorks simulator.

TABLE IV. CHARACTERISTICS OF ALUMINUM ALLOY 6061-T6

Aluminum alloy 6061-T6	
Elastic module	69 GPa
Poisson's ratio	0.33
Traction limit	310 MPa
Elastic limit	275 MPa
Bulk density	2700 kg/m ³

Note. The data acquired was from the SolidWorks simulator.

To perform the analysis of the exoskeleton system, the fastenings and connections of the hip, knee, and ankle are identified to determine the natural behavior that the structure will suffer, the maximum force and torque that will be subjected are also determined depending on the maximum weight the exoskeleton can support which is 70 kilograms. Table V shows the exoskeleton parts' weights, strength, and torque data.

TABLE V. DATA ON WEIGHT, STRENGTH, AND TORQUE

	Material Analysis			
	Hip	Thigh	Leg	Ankle
<i>P</i> _{thigh} (Kg)	7			
<i>P</i> _{leg} (Kg)	3.3	3.3		
<i>P</i> _{foot} (Kg)	1.1	1.1	1.1	1.95
<i>P</i> _{lower limb} (Kg)	11.4	4.4	1.1	1.95
<i>F</i> _{lower limb} (N)	111.83	43.16	10.79	19.13
<i>T</i> _{lower limb} (N-m)	41.5	30.5	28.5	18.5

D. System integration

The interaction of the mechanical and electronic system of the exoskeleton design would be located as shown in Fig. 11, were to observe the distribution of the electronic circuits, considering that each lower extremity of the design will have its controller that will be controlling three motors, the circuit of the LCD screen will be in the back of the hip where the specialist can program the routines according to the need of the patient. In Fig. 12, the exoskeleton can be seen in a frontal view implemented in a human body model and a side view of the exoskeleton.

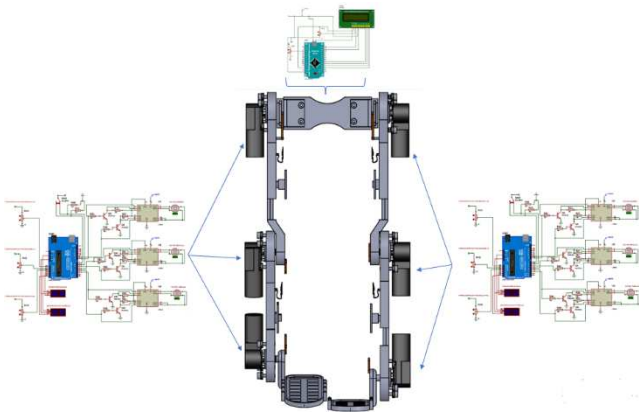


Fig. 11. Interaction of the mechanical and electronic system.

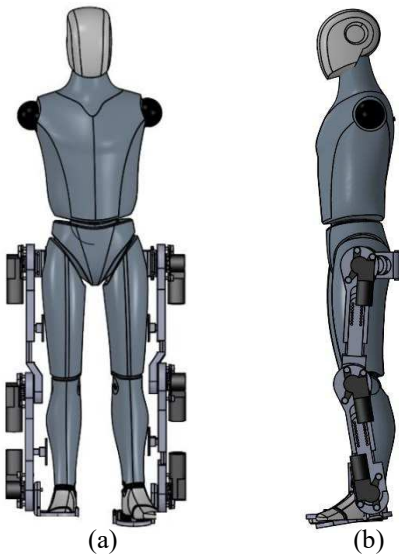


Fig. 12. View of the exoskeleton design. (a) Front view of the exoskeleton. (b) Side view of the exoskeleton.

Fig. 13 shows the finite element analysis simulation with Von Mises' theory of hip assembly.

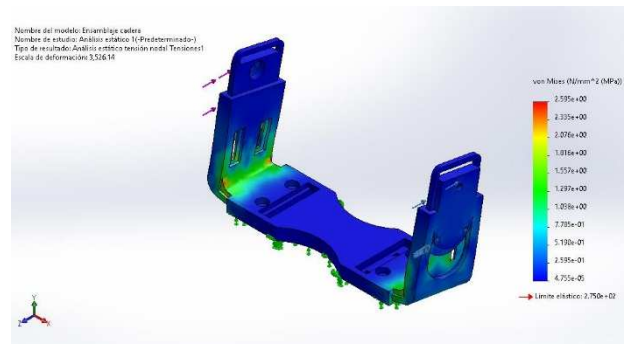


Fig. 13. Static analysis and von Mises stress of the hip assembly.

Fig. 14 shows the finite element analysis simulation with Von Mises' theory of thigh assembly.

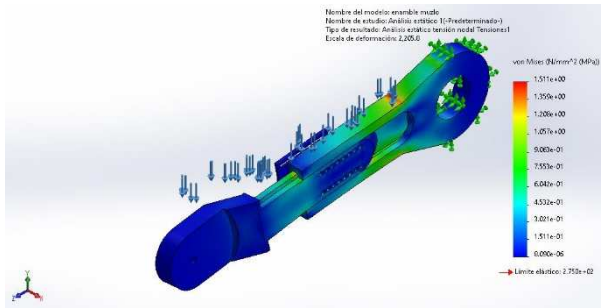


Fig.14. Static analysis and von Mises stress of thigh assembly.

Fig. 15 shows the finite element analysis simulation with Von Mises' theory of leg assembly.

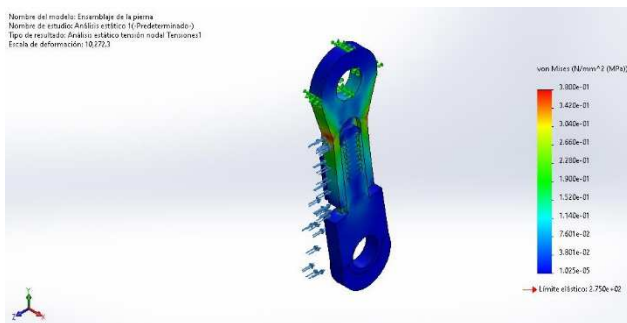


Fig. 15. Static analysis and von Mises stress of leg assembly.

Fig. 16 shows the finite element analysis simulation with Von Mises' theory of ankle assembly.

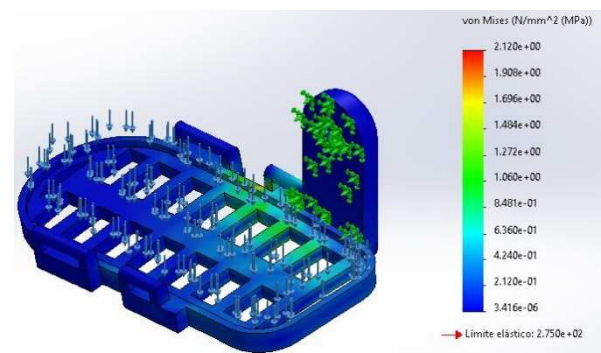


Fig.16. Static analysis and von Mises stress of the ankle assembly.

With these results of static analysis and tensions of Von Mises, the proposed prototype with the material of Aluminum Alloy 6061-T6 typically supported 70 kilograms.

IV. CONCLUSIONS

The design of the exoskeleton structure could graduate the links, being able to cover patients who have a size from 150 to 172 centimeters.

The analyses of the parts that make up the structure of the exoskeleton were subjected to an analysis of Von Mises stress, resulting in displacements and unit deformations, in the SolidWorks software, where results provided show that the alloyed material of aluminum 6061-T6 is strong enough to withstand the weight of a person and by the characteristics of the material makes the mechanical structure a lightweight design.

With the interaction of the electronic and mechanical systems, the design of the exoskeleton has different methods of control through the sensors that are part of the design, the EMG sensors control the intention of movement of the patient, and the potentiometers show the position of the motors to generate an adequate rotational rotation.

V. REFERENCES

- [1] R. López, H. Aguilar, S. Salazar, R. Lozano, and J. A. Torres, "Modelado y control de un exoesqueleto para la rehabilitación de extremidad inferior con dos grados de libertad," *Revista iberoamericana de automática e informática industrial*, vol. 11, no. 3, pp. 304–314, 2014.
- [2] Anónimo. Inter empresas. [Online]. Exoesqueletos: la edad del 'hombre de hierro'.2015 [citado el 20 de agosto del 2020]. Disponible en: <https://www.interempresas.net/Proteccion-laboral/Articulos/211884-Exoesqueletos-lacdad-del-hombre-de-hierro.html>.
- [3] Cisneros C. Maquinas especifica EXOESQUELETOS. [Online]. Sites.google.com.2016. [citado el 20 de agosto del 2020]. Disponible en: <https://sites.google.com/site/fgtce04equipo03tgigestion/funcionamiento-de-los-exoesqueletos>.
- [4] C. Pais-Vieira, M. Allahdada, J. Neves-Amado, A. Perrotta, E. Morya, R. Molioli, E. Shapkova, and M. Pais-Vieira, "Method for positioning and rehabilitation training with the exoatlet® powered exoskeleton," *MethodsX*, vol. 7, p. 100849, 2020.
- [5] Zhang, W. Chen, Y. Chai, J. Wang, and J. Zhang, "Gait graph optimization: Generate variable gaits from one base gait for lower-limb rehabilitation exoskeleton robots," *arXiv preprint arXiv:2001.00728*, 2020.
- [6] K. A. Witte and S. H. Collins, "Design of lower-limb exoskeletons and emulator systems," in *Wearable Robotics*. Elsevier, 2020, pp. 251–274.
- [7] T. Pan, C.-C. Chang, Y.-S. Yang, C.-K. Yen, Y.-H. Kao, and Y.-L. Shiu, "Development of MMG sensors using PVDF piezoelectric electro-spinning for lower limb rehabilitation exoskeleton," *Sensors and Actuators A: Physical*, vol. 301, p. 111708, 2020.
- [8] J. H. Hernandez, S. S. Cruz, R. Lopez-Gutierrez, A. Gonzalez-Mendoza, and R. Lozano, "Robust nonsingular fast terminal sliding-mode control for the sit-to-stand task using a mobile lower limb exoskeleton," *Control Engineering Practice*, vol. 101, p. 104496, 2020.
- [9] L. Zhou, W. Chen, W. Chen, S. Bai, J. Zhang, and J. Wang, "Design of a passive lower limb exoskeleton for walking assistance with gravity compensation," *Mechanism and Machine Theory*, vol. 150, p. 103840, 2020.
- [10] D. A. Tibaduiza Burgos, P.-A. Aya-Parra, and M. Anaya, "Exoesqueleto para rehabilitación de miembro inferior con dos grados de libertad orientado a pacientes con accidentes cerebrovasculares," *INGE CUC*, 2019.
- [11] V. G. P. Lugo, A. G. Betancourt, I. M. Panecat, and R. E. L. Torres, "Exoesqueleto para hipotrofia en miembro inferior con asistencia de electroestimulación," *Dra. Lucia Marquez Perez Ing. Wendolin Jacinto Diaz*, p. 193.
- [12] D. Huamanchahua, J. Ortiz-Zacarias, J. Asto-Evangelista and I. Quintanilla-Mosquera, "Types of Lower-Limb Orthoses for Rehabilitation and Assistance: A Systematic Review," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0705-0711, doi: 10.1109/UEMCON53757.2021.9666710.
- [13] D. Huamanchahua, J. C. Vásquez-Frías, N. Soto-Conde, D. Lopez-Meza and Á. E. Alvarez-Rodríguez, "Mechatronic Exoskeletons for Hip Rehabilitation: A Schematic Review," *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, 2021, pp. 381-388, doi: 10.1109/RCAE53607.2021.9638869.
- [14] D. Huamanchahua, Y. Taza-Aquino, J. Figueroa-Bados, J. Alanya-Villanueva, A. Vargas-Martinez and R. A. Ramirez-Mendoza, "Mechatronic Exoskeletons for Lower-Limb Rehabilitation: An Innovative Review," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1-8, doi: 10.1109/IEMTRONICS52119.2021.9422513.

Algorithmic Formalization of Risk Synthesis based on Functioning Table

1st Anvar Kabulov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 anvarkabulov@mail.ru

2nd Inomjon Yarashov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 timprivate345@gmail.com

3rd Alisher Tukhtakulov
*Faculty of Applied Mathematics
 and Intelligent Technologies*
National University of Uzbekistan
 Tashkent, Uzbekistan
 aaroonalduo@gmail.com

Abstract—In today’s world of information technology, information security is an important issue. In addressing information security issues, identification of attacks on the system, search for countermeasures to eliminate them, analysis of the causes of attacks, algorithmic review and algorithmization are carried out. Attacks were taken as an example in the analysis of attacks. Issues such as traffic changes and packet overload are investigated in the origin of attacks. Processes can be fully described through algorithmization in research. Identifying a specific set of information security threats by analyzing Functioning tables for algorithmic models and synthesizing and verifying incoming and outgoing data in their different versions allows any complex protection to be achieved through algorithmic descriptions. This is especially noteworthy. The diversity of system blocks means their ability to grow automatically to increase data size and provide redundant data protection. Therefore, additional safety precautions should be taken, as this article discusses the use of certain types of Functioning tables.

Index Terms—formalization, functioning table, information security, algorithmization, analyzing.

I. INTRODUCTION

The 21st century has entered human life with the rapid development of information technology and its deep application in all spheres of human activity. Thanks to the rapid development of computer technology [1]–[9], people can access a variety of information, exchange information and communicate in real time anywhere in the world. For free navigation in information flows, a modern specialist of any profile must receive, process, and use data using computers, telecommunications, and other means of communication. But to do that, you need to know the rules of navigation through the vast amount of information available and have a certain cultural background [10]–[14].

The emergence of new information technologies and their penetration into various spheres of society has led to the emergence of a new field of informatics, the science of social informatics. Social informatics studies the following issues: spacing

- Laws and problems of information society development;
- Information resources as socio-economic and cultural factors of society development;
- A person in an informed society;
- Information culture;

- Information security.

Information has always played an important role in the life of society as well as in the life of the individual. In the history of mankind, the development of the means of collecting, storing and transmitting information has not been smooth, and there have been several events of global significance in the field of information called the "Information Revolution". While information technology [15]–[20] is one of the most important pillars of society, information itself has become one of the most valuable assets of today. Man has always taken care of his valuables, not allowing them to be stolen or damaged. The above points show that information technology and information are the most important assets to protect. This is because information technology refers not only to systems for automated processing [21]–[25] of information, but also to any system (even if it is based on manual labor) that is weighed on the basis of information in the implementation [26]–[28] of an arbitrary process, related to planned activities. provided. Therefore, the security of information and information systems is a topical issue of our time.

No one in this century can question which smart robots, advances and other technical gadgets will be made that people cannot indeed envision. The advancement of information technology will lead to an increment in efficiency, quality and, over all, high efficiency in each organization. Present day information and communication technologies and the creation of conveniences moreover posture modern challenges. In later a long time, the fast improvement of information and telecommunications technologies within the world has altogether impacted the worthy place of states within the worldwide information society.

Nowadays, the risk to the security [29]–[33] of data put away in databases and circulating in telecommunications frameworks is developing quickly. As a result, the issue of information security has gotten to be a topical issue for the world. To date, one of the foremost solid devices in guaranteeing information security [34]–[36] is cryptographic protection of data. Within the world, this heading is creating quickly. New cryptographic systems, algorithms, measures are being created and connected in different areas. It is known that the data assets of any organization are one of the components

determining its financial and military potential. The successful utilize of this asset will guarantee the security of the country and the effective arrangement of a equitably educated society. In such a society, the speed of information exchange [10] will increment, the application of progressed information and communication technologies for the collection, capacity, preparing and utilize of information [18] will be widespread.

II. MATERIALS AND METHODS

A systematic study of the organizational and functional structure for algorithmic analysis of information protection [37], [38] has made it possible to highlight a number of general rules arising from the specific characteristics of the activity that must be followed at all stages of information protection development and implementation. Electronic key management system.

The main task of creating an information protection system is to choose the principles of management (objectives, criteria and restrictions on the operation of various parts of the system); distribution of functions between system levels and selection of appropriate decision schemes; establishing the right relationship between levels; coordinating the goals of subsystems at different levels and optimally stimulating their work; distribution of rights and obligations; in the distribution of functions performed between executors and technical means; selection of a set of technical means for data transmission and processing.

The most important principles to be applied in the operation of the management system are taken into account when creating an information protection system. The most important principles:

The principle of legality. Decisions made using automated information systems must be clear and unambiguous in accordance with applicable law, which is ensured by all entities (task manager, programmer, etc.) that interact in the process of creating them.

The importance of this principle can hardly be overestimated, although it is very difficult to ensure its scrupulous observance in our time due to the imperfection of legislation, which does not have time to properly regulate the mechanism of legal regulation of perestroika processes. place in the World. The principle of system security(Fig 1). A logically understandable principle that assumes the universality of protective measures against all forms of threats to objects of protection, which is a consequence of the uncertainty of the protection process. This principle is usually associated with the presence of a "human factor", since it is not clear: who, when, where and how can violate the security of the protected object.

This principle is also necessary to protect the system (Fig 2) from unauthorized access to it by unauthorized persons. An illustration of following this principle is the availability of means of protection against attempts to unauthorized changes to the software, which, in turn, implements some kind of protective function in information systems, for example, access control. Another example is an intruder alarm device that functionally supports a security line and is in turn protected by

some special means (eg masking). Closing channels for unauthorized information retrieval should begin with controlling user access to information system resources. This challenge is being tackled on the basis of a number of fundamental principles(Fig 3). Access validity principle. This principle

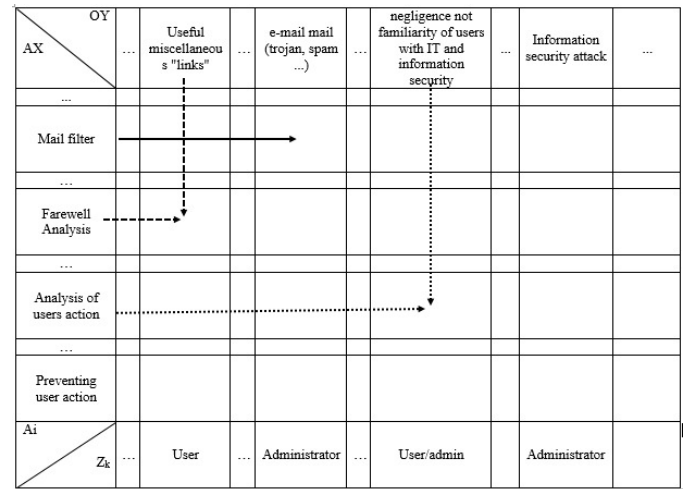


Fig. 1. Formalization scheme of system components by means of control algorithm

consists in the mandatory fulfillment of two basic conditions: the user must have a sufficient "form of clearance" to obtain information of the required level of confidentiality, and this information is necessary to perform his production functions. Note that in the field of automated information processing, active programs and processes, as well as information carriers of varying degrees of aggregation, can act as users. Then the access system involves the definition for all users of the appropriate software and hardware environment or information and software resources that will be available to them for certain operations.

III. RESULT AND DISCUSSION

The evolution of information technology (IT) is associated with intelligent systems in which there are processes of origin, adaptation and development. The systems approach defines the methodology and principles of building IT systems. The principle of the possibility of modeling helps prevent design errors in cybernetic systems. The principle of connectivity in the development of an effective system considers the object of protection comprehensively, combining the object of protection, the external environment, means of protection and threats of the attacker and taking into account the relationship: source of threat - factor (vulnerability) - threat (action) - consequences (attack).

Building a security system is a prerequisite for ensuring the security of confidential information stored and processed in the information system. Requirements for the information security system are formed based on the results of the survey of the information system and are focused on neutralizing system vulnerabilities.

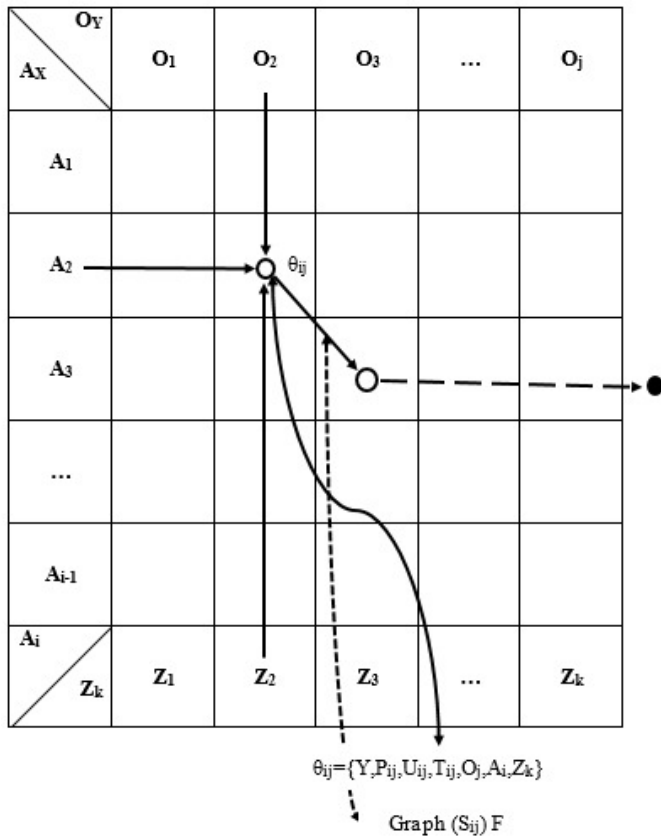


Fig. 2. A model showing the dynamics and algorithmic description of the system in terms of functionality

$AB-\theta_{ij} = \{Y, P_{ij}, U_{ij}, T_{ij}, O_j, A_i, Z_k\}$ coordinates of the initial threat to IS and steps to prevent:

- 1) To proceed to the next stage, you must always have at least 3 inputs: 1) the source of the threat; 2) counteraction; 3) user privilege level.
- 2) In each cell, the function $\theta_{ij} = \{Y, P_{ij}, U_{ij}, T_{ij}, O_j, A_i, Z_k\}$ will be considered from the beginning.

$\{X, Y, A, O, \Theta, T, U, S, F, P\}$ ACS for ensuring IS security, as well as preventing any kind of threats to IS and IR.

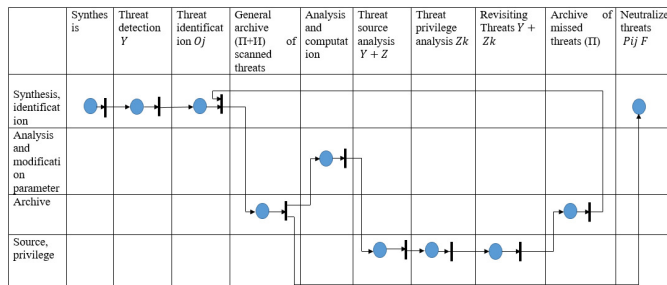


Fig. 3. Schematic of an algorithmic model of threat synthesis and dynamic modification

Y a set of possible threats $Y\{O_j\}$;

- X a set of threat prevention solutions $X\{A_i\}$;
- A Defined threat prevention solution;
- O Defined action of threats;
- Θ Coordinates between A_i and O_j ;
- T Time (to prevent the threat and successfully implement the threat);
- U External impact (on $\Theta_{ij}\{A_i, O_j\}$);
- S Transitions graph (transition from one Θ_{ij} to another $\Theta_{i+n, j+m}$);
- F Transition process;
- P Computational and logical operations of input, output and control;
- Z Privilege.

This algorithmic model includes the following processes and components: source blocking; some action of the program (set by the IS policy by the administrator) has been stopped; ineffective work of the program and inhibition of the action; ACS (computer and other equipment); revision of the program's action by the OS; no convenience and uninstallation of the program; analysis of the life of the program, adding to the archive as a threat; open access to IR; changing OS parameters; changing the parameters of installed programs; downloading other programs; modifying and deleting files; transfer information; prevention of automatic control; prevent automatic change; analysis of incoming traffic; analysis of outgoing traffic.

IV. CONCLUSION

Thus, the algorithmic models allow to perform arbitrary complex protection by analyzing the Functioning tables designed to synthesize and verify the incoming and outgoing information in their analog version, as well as to describe a certain part of the threats to information security through algorithmic formalization. It can be emphasized. The variability of elemental piece structures means their ability to grow automatically to increase data volume and introduce redundant data protection. It is therefore necessary to take additional protective measures, as this article discusses the use of specific types of Functioning Tables. In addition to the above material scope, it can be applied to other areas: the socio-economic system and various scientific fields, the possibility of their use will undoubtedly help to develop algorithmic descriptions to solve effective protection problems.

REFERENCES

- [1] Rashid, Aqsa, Masood, Asif and others, RC-AAM: blockchain-enabled decentralized role-centric authentication and access management for distributed organizations, Cluster Computing, vol. 24, pp. 3551–3571, 2021.
- [2] Fugkeaw, Somchart, Sanchol and Pattavee, A Review on Data Access Control Schemes in Mobile Cloud Computing: State-of-the-Art Solutions and Research Directions, SN Computer Science, vol. 3, pp. 1–11, 2022.
- [3] L. Vu, C. Bui, Q. Nguyen and D. Rossi, A deep learning based method for handling imbalanced problem in network traffic classification, December 2017, pp. 333339.
- [4] G. Aceto, D. Ciunzo, A. Montieri and P. A. Multi-classification approaches for classifying mobile app traffic, Journal of Network and Computer Applications, vol. 57, pp. 131145, 2018.

- [5] P. Wang, C. Xuejiao, Y. Feng and S. Zhixin, A survey of techniques for mobile service encrypted traffic classification using deep learning, *IEEE Access*, vol. 7, pp. 5402454033, 2019.
- [6] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, 2018, pp. 108116.
- [7] P. N. Matheus, F. C. Luiz, L. Jaime and L. P. Mario, Long short-term memory and fuzzy logic for anomaly detection and mitigation in software-defined network environment, 2020, pp. 8376583781.
- [8] B. Naveen and S. Manu, Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting DDoS attacks, *Romanian journal of information science and technology*, vol. 23, no. 3, p. 250–261, 2020.
- [9] S.E.Mahmoud, L.Nhien-An, D.Soumyabrata and D.J.Anca, Ddosnet: A deep-learning model for detecting network attacks, in 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). IEEE, 31 Aug.-3 Sept. 2020, pp. 18.
- [10] M. S. Yin, P. A. Pye and S. H. Aye, A slow DDoS attack detection mechanism using feature weighting and ranking, *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore*, pp. 45004509, March. 7-11, 2021.
- [11] A. H. Lashkari, D. G. Gerard, M. M. Mamun and A. A. Ghorbani, Characterization of TOR traffic using time-based features, 2017, pp. 253262.
- [12] N. Miloslavskaya, A. Tolstoy and S. Zapechnikov, Taxonomy for unsecured big data processing in security operations centers, Aug.2224 2016, pp. 154159.
- [13] N. Miloslavskaya and A. Makhmudova, Survey of big data information security, vol. 8, Aug.22-24 2016, pp. 133138.
- [14] D. Khasanov, K. Khujamatov, B. Fayzullaev and E. Reypnazarov, WSN-based Monitoring Systems for the Solar Power Stations of Telecommunication Devices, *IJUM Engineering Journal*, 2021, 22(2), p. 98118.
- [15] I. Yarashov, Algorithmic Formalization Of User Access To The Ecological Monitoring Information System, 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-3.
- [16] N. G. Zagoruiko, I. A. Borisova and O. A. Kutnenko, Constructing a concise description of data using the competitive similarity function, *Siberian Journal of Industrial Mathematics*, vol.1, no.16, pp.2941, 2013.
- [17] G. Juraev and K. Rakhimberdiyev, Modeling the decision-making process of lenders based on blockchain technology, 2021 International Conference on Information Science and Communications Technologies (ICISCT), pp. 16, 2021.
- [18] I. Kalendarov, Algorithm for the Problem of Loading Production Capacities in Production Systems, *Lecture Notes in Networks and Systems*, vol. 246, pp. 887896, 2022.
- [19] DDoS evaluation dataset (cic-ddos2019), 2018. [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>
- [20] I. Sharafaldin, A. H. Lashkari, H. Saqib and A. Ghorban, Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy, in *Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCSST)*. IEEE, Oct. 1-3, pp. 18.
- [21] A. Kabulov, I. Normatov, E. Urunbaev and F. Muhammadiev, Invariant continuation of discrete multi-valued functions and their implementation, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [22] A. Kabulov and I. Saymanov, Application of IoT technology in ecology (on the example of the Aral Sea region), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [23] A. Kabulov, I. Saymanov, I. Yarashov and F. Muxammadiev, Algorithmic method of security of the Internet of Things based on steganographic coding, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 Proceedings.
- [24] A. Kabulov and I. Yarashov, Mathematical model of Information Processing in the Ecological Monitoring Information System, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-4.
- [25] A. Kabulov and M. Berdimurodov, Optimal representation in the form of logical functions of microinstructions of cryptographic algorithms (RSA, El-Gamal), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [26] A. Kabulov, I. Saymanov and M. Berdimurodov, Minimum logical representation of microcommands of cryptographic algorithms (AES), *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [27] A. Kabulov, I. Normatov, I. Kalendarov and I. Yarashov, Development of An Algorithmic Model and Methods for Managing Production Systems Based on Algebra over Functioning Tables, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-6.
- [28] A. Kabulov, I. Kalendarov and I. Yarashov, Problems of Algorithmization of Control of Complex Systems Based on Functioning Tables in Dynamic Control Systems, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*, pp. 1-4.
- [29] A. Kabulov, E. Urunboev and I. Saymanov, Object recognition method based on logical correcting functions, *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2020*, pp. 1-4.
- [30] A. Kabulov, A. Babadzhanov and I. Saymanov, Completeness of the linear closure of the voting model, *AIP Conference Proceedings*, 2022 (accepted).
- [31] A. Kabulov, A. Babadzhanov and I. Saymanov, Correct models of families of algorithms for calculating estimates, *AIP Conference Proceedings*, 2022 (accepted).
- [32] A. Kabulov, I. Normatov, S. Boltaev and I. Saymanov, Logic method of classification of objects with non-joining classes, *Advances in Mathematics: Scientific Journal*, 2020, 9(10), p. 8635–8646.
- [33] A. Kabulov, I. Kalendarov and I. Saymanov, Models and algorithms for constructing the optimal technological route, group equipment and the cycle of operation of technological modules, *Smart transport conference 2022 Conference*, pp. 1-11.
- [34] A. Kabulov, Number of three-valued logic functions to correct sets of incorrect algorithms and the complexity of interpretation of the functions, *Cybernetics*, 1979, 15(3), p. 305–311.
- [35] A. Kabulov and G. Losef, Local algorithms simplifying the disjunctive normal forms of Boolean functions, *USSR Computational Mathematics and Mathematical Physics*, 1978, 18(3), p. 201–207.
- [36] A. Kabulov, Local algorithms on Yablonskii schemes, *USSR Computational Mathematics and Mathematical Physics*, 1977, 17(1), p. 210–220.
- [37] M. Shaw, N. Mandal and M. Gangopadhyay, A compact polarization reconfigurable stacked microstrip antenna for WiMAX application, *International Journal of Microwave and Wireless Technologies* this link is disabled, 2021, 13(9), p. 921-936.
- [38] A. Panda, P. Bhowmick, S.K. Bishnu, A. Ganguly and M. Gangopadhyaya, Derivative based Kalman filter and its implementation on tuning PI controller for the Van de Vusse reactor, 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings, 2021, 9422585

Intelligent Monitoring and Control of Wind Turbine Prototype Using Internet of Things (IoT)

Md Jishan Ali
 Dept. of Electrical Engineering
 Aliah University, Kolkata-700160
 West Bengal, India
 jishan.een.au@gmail.com

Ashim Mondal
 Dept. of Electrical Engineering
 Aliah University, Kolkata-700160
 West Bengal, India
 ashim.au.een@gmail.com

Pallav Dutta
 Dept. of Electrical Engineering
 Aliah University, Kolkata-700160
 West Bengal, India
 pallav.dutta@aol.com

Abstract— The wind power industry has made the most recent age of wind turbines (WTs) on all platforms, from small windmills to ocean wind turbines. An emerging approach to reducing overall costs is to make wind turbines smarter. At the time of diagnosing fault concern personnel have to reach the particular location. As human beings are more prone to committing errors, electronic gadgets such as microcontrollers and sensors could be entrusted. To ensure the structural integrity of the tower and its components to assure the low cost of energy and safety of the surroundings, the Control Centre is responsible for Monitoring and Controlling WTs in those locations. Various parameters like oil level, speed of the generator, level of vibration, environmental conditions like rain, humidity, and temperature are to be monitored and controlled for the proper working of WTs. This paper describes monitoring and controlling of speed and vibration of a WT with the assistance of IoT. By the use of this described system, the speed of the turbine, temperature, voltage, and vibration can be monitored and controlled from anywhere, or just through a PC or smartphone. To show the described model in the hardware prototype, simply a DC motor is used as a WT and IoT is implemented through Arduino Uno microcontroller and Bolt IoT Wi-Fi Module. The proposed framework is based on the Bolt-Arduino (Boltuino) IoT platform that provides Wireless Fidelity (Wi-Fi) and a cloud facility. One of the major objectives behind the integration of web and cloud networks is that the user controls all devices and data from far away over a web, local area network (LAN), or cloud. Also, by using the Bolt IoT module in this proposed system, code size can be decreased by around 80%.

Keywords — *Wind turbine, vibrations, temperature measurements, internet of things, remote condition monitoring & control.*

I. INTRODUCTION

Nowadays, the world is developing towards renewable energy at a great rate because of the quantitative and qualitative impact of fuels on nature. Instead of fuel, wind can be replaced with renewable energy, which is still pure. Wind energy is a source of renewable energy from blowing air on the surface of the earth. Since the wind energy system has some valuable and

useful advantages over solar energy, so the tracking of the wind and wind turbine system is very essential [1]. Currently, wind power plants are unique and the uppermost sources of power generation across the globe. The Government of India is focusing on accomplishing 227 GW of sustainable energy capacity (including 114 GW of solar capacity addition and 67 GW of wind power capacity) by 2022 [2].

The reason behind wind turbine failures due to faults in generator and gearbox in practical turbines and the downtime due to such failures is more significant. So, the appropriate operation and conservation techniques are required to provide reliable and worthwhile green energy [3]. In general, wind turbine systems are more liable to fault, and maintenance of wind turbines in remote locations is really a difficult task. Hence, it is required a proper operating approach to measure and acknowledge crucial parameters for industrial purposes. Condition monitoring and investigation strategies for wind turbines are a universally utilized tool for early detection of failure, which reduces time and increases system efficiency and pro-activeness [4]. The prime area of wind turbine downtime is due to bearing screw-ups, especially inside the generator and gearbox [5]. It is necessary to assemble an authentic condition monitoring and controlling system to enhance maintenance efficiency. These limitations can be controlled by installing vibration sensors in the gear box of the WT [6] and the speed of the WT generator can be maintained by installing an infrared sensor (IR) in front of the generator. Also, temperature and voltage sensors are used to get updated information about atmospheric temperature and generated voltage accordingly. Monitoring, controlling, and updating Arduino programs can be done from the central station. Also, with the help of simple Bolt IoT program, the activity can be monitored via any web-connected internet browser of Computer or Smartphone or Tablets. Through the ideas of IoT, the link between bridging the gap and wind turbines that are situated at a distance from the controlling center can be controlled through software or a new tool. In the following section configuration of the proposed system is discussed briefly.

II. LITERATURE SURVEY

In some recent work, several attempts have been made to improve the existing wind turbine (WT) system so that a cost-effective controlling and monitoring system can be made for all. So, executing a more reliable system can reduce the manpower

requirements, and time as well. However, many methods are still operating; some of these are discussed below:

H. F. Liew et al.[4] 2020 introduced a system where they described an IoT application and ESP 8266 to track, take data, and determine issues in wind turbines. This permits the user to manage the whole framework remotely through a protected web-connected internet. This framework assists end-users to control energy sources, physically and remotely just by simply using smartphones or PCs. The parameters that have been taken utilizing ESP 8266 are current, voltage, wind speed, and power. When the wind turbine generates electricity, it will show the results in think-speak. The major drawback of this framework is that it requires writing more code to take data in think-speak. As well as it provides service up to a limited time period.

In 2016, D. Kalyanraj et al. [6] proposed a wind turbine (WT) monitoring and control system using the Internet of Things (IoT) where they describe a minimal expanse framework having data logger facilities. Parameters of the particular wind plant like the magnitude of voltages, currents, amount of power generation, turbine speed, temperature, humidity, and level of vibration can be observed by utilizing this system. In this very paper, control of the turbine is embraced based on the turbine vibration and IoT is executed by Arduino Uno and Raspberry Pi. However, this framework suffers from overheating and many users complain that it is not able to run in the most user-friendly Windows OS. This system is costly as Raspberry Pi is used as a cloud server storage device.

In Celal Bayar University Journal of Science 2019, Ersin Akyuz et al. [7] raised that crucial parameter are measured for the performance analysis of a small wind turbine model. Estimating the performance of the framework and avoiding faults in the framework measures can be done. Basic parameters like air temperature, wind speed, battery voltage, and current were recorded and calculated through a data logger. These calculation results were sent back to the Microsoft Azure cloud computing framework and taken down in it. Simultaneously, representation with the guide of the cloud system was taken place and visualized on the web browser through the Power BI (Microsoft) platform. But according to users, the main disadvantage of this Power BI platform is that it takes a little more than normal time or even hangs while processing millions of lines and segments of information. Also, this framework does not provide information about the vibration level of WT.

Peter J. Tavner et al. [8] in the year 2010 described a paper where a wind turbine (WT) condition monitoring strategy utilizes the generator output power and speed of rotation to determine a fault identification signal. The fault calculation utilizes a continuous wavelet-change-based adaptive filter to follow the energy in the time-fluctuating fault-related frequency bands in the power signal. The frequency of the filter is constrained by the speed of the generator, and the filter bandwidth is adapted to the fluctuation of speed. Utilizing this method, fault characteristics can be removed, with the low calculation of times. But in this model, MQTT utilizes the

transmission control protocol (TCP) protocol, which requires more memory and more processing power.

In the year 2017, Fran Lizza. M et al. [9] published a paper in the International Journal of Advanced Research in Management, Architecture, Technology, and Engineering where they proposed a monitoring framework with IoT and UART (Universal Asynchronous Receiver/Transmitter) to check and determine the issue in the WT application. In this system, the major drawback can be noticed that the speed for data transfer is less compared to parallel communication.

III. CONFIGURATION OF THE PROPOSED SYSTEM

Figure 1 describes the overall block diagram of the proposed framework. To save costs, give more usability to consumers, and avoid economic problems. The microcontroller has been considered a major and significant system. This system consists of three parts i.e., sensing unit, control unit and monitoring unit. The entire process of monitoring and control will be controlled by the connected microcontroller. The sensor will detect various information and send it to the microcontroller and then the microcontroller will send the data to the Bolt IoT cloud. According to the data, the fault can be diagnosed or the program can be changed. If the vibration data of WT is more than the threshold value then the relays will trip and also give alarms to alert the operator; then WT will turn OFF and ON accordingly.

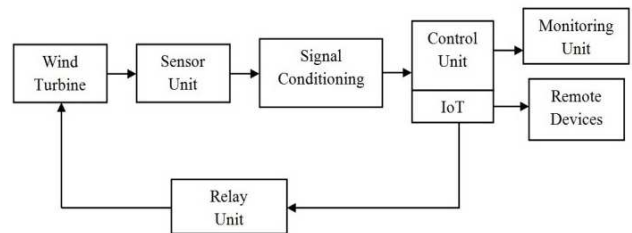


Figure 1: Overall Block Diagram of the Proposed System

Here in figure 2 shows the Schematic Diagram of the Proposed System. The infrared, temperature, voltage and vibration sensors are interfaced with the Arduino that will be responsible for taking the data from the sensors and for communicating with Wi-Fi module. Bolt IoT Wi-Fi module is mainly used to implement the IoT technology that will take data to the Bolt cloud to control the process from anywhere to reduce the man power requirements and save time. Also, a manual process will be available to display the data through LCD/LED display from a particular central station.

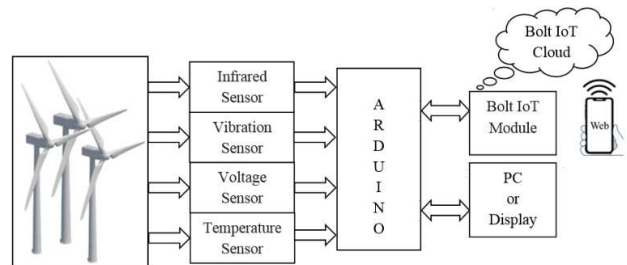


Figure 2:- Schematic Diagram of the Proposed System

The functional block diagram of the system is shown in figure 3. In the sensing unit, parameters like generator speed, vibration level of turbine, voltage, and atmospheric humidity and temperature are calculated using the respective sensors. Detected signals collaborate with the Arduino for safe activity. Bolt Wi-Fi module and Arduino are connected through USB ports accordingly. The control unit is associated with a PC or a cloud server. The values of the control unit are compared with the reference values occasionally. Whether the actual given value crosses the maximum secure limit, the control unit will give necessary signals to the corresponding relays in order to keep the ideal system performance. In the proposed framework, turbine control is acquired based on the level of vibration in the wind turbine (WT). Since turbine vibration is an extremely critical factor, therefore, the monitoring system should give essential protection to the turbine [10]. The turbine could get higher vibrations when the speed of air is beyond the cut-off values. During this circumstance, reliable and fast protection is mandatory, or else the entire system might get harmed. Under this condition protection of wind turbines is provided by relays by turning OFF, the generation of power.

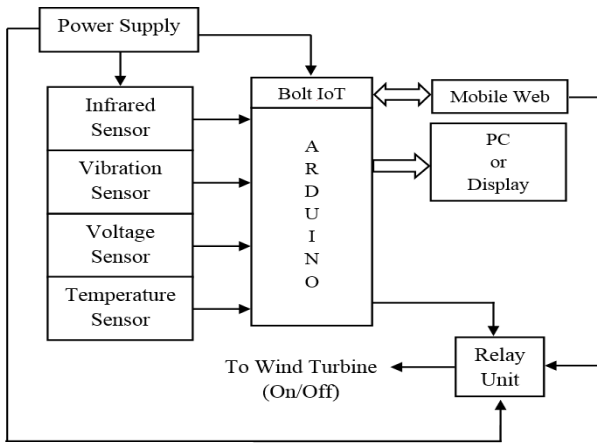


Figure 3: Functional Block Diagram

IV. SYSTEM DETAILS & HARDWARE ARCHITECTURE

The system comprises of Arduino Uno microcontroller, Bolt IoT Wi-Fi module, Bolt cloud, and few essential electronic components i.e. breadboard, relay, buzzer, vibration sensor, voltage sensor, temperature sensor infrared sensor, Connecting wires etc. A DC motor is used as a prototype for a wind turbine (WT).

A single system is capable of controlling the speed of the wind turbine and vibration which generally occurs in the gear box of a WT. At the same time, generated voltage and atmospheric temperature can be monitored from the web browser. Connection required for the proposed system that is shown in figure 4 where a prototype wind turbine (WT) is connected to the Arduino Uno microcontroller pin number 10 through a single channel relay to control the WT. The Arduino Uno microcontroller board is powered up by a battery or PC through a USB cable. The RX and TX of Bolt IoT Wi-Fi module are interfaced with Arduino Uno microcontroller digital

pin numbers 9 and 8 respectively for the purpose of serial communication. Here the four sensors, vibration sensor, IR sensor, temperature sensor, and voltage sensor are interfaced to the board. The particular VCC and ground pins of all the sensors are connected to the 5V and pins of the Arduino Uno microcontroller, respectively. A buzzer is connected for alerting purpose when there are some problems in WT, then the buzzer turns ON at the same time relay gets interrupt and wind turbine turns OFF. At the same time, information about temperature and generated voltage can also be taken from the respective sensors [11]. The purpose of connecting Bolt IoT Wi-Fi module is to take the data to the Bolt cloud, which can be accessed and controlled over laptop /mobile/tablet through HTTP & HTTPS protocols [7]. A message alerting system will be there to inform the surveyor if some crucial fault occurs in the system. Moreover, the data or the information of the system can also be monitored, as well as updating new programs to the system can be done manually from the central stations by PCs or LCDs. The entire circuit and hardware connection diagram is shown in figure 4. Flowcharts for the infrared (IR) and the vibration sensors as well as relays that are programed on the microcontroller are presented in figure 5.

A. Bolt IoT Wi-Fi Module

The entire framework depends on the Bolt IoT. The user-friendly Bolt IoT module is an IoT platform that provides Wi-Fi (Wireless Fidelity) and in-built cloud connectivity to the connected sensors. This platform is based on the ESP-8266 Wi-Fi module [12].

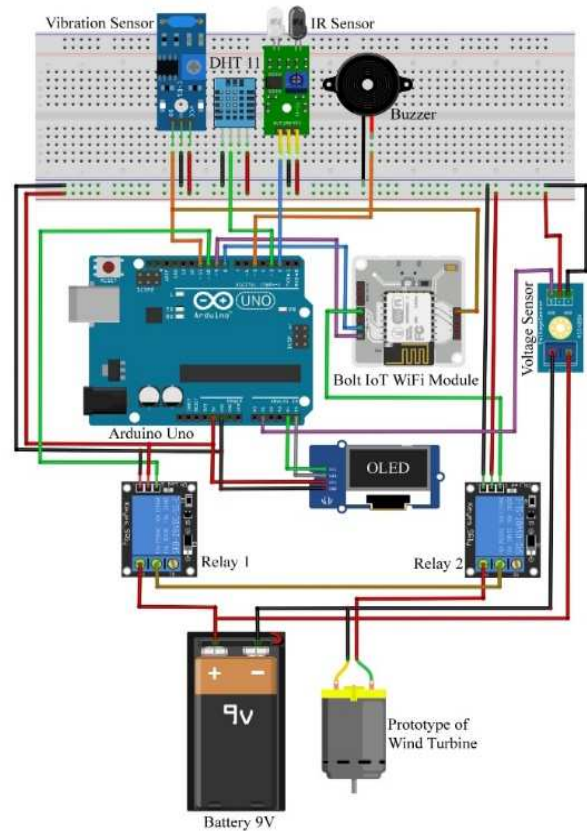


Figure 4:- Connection Diagram

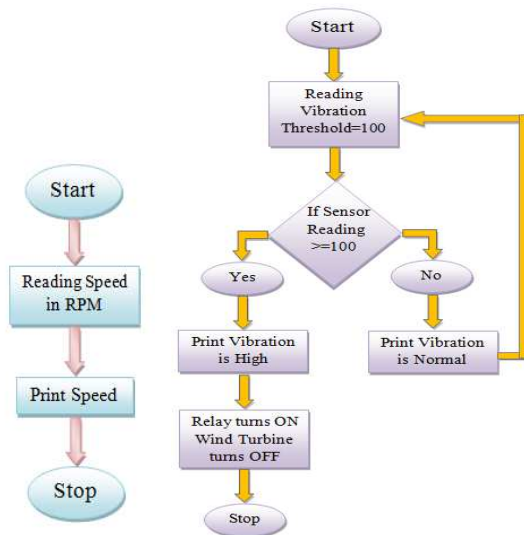


Figure 5: Flowchart for Infrared and Vibration Sensors

The entire framework depends on the Bolt IoT. The user-friendly Bolt IoT module is an IoT platform that provides Wi-Fi (Wireless Fidelity) and in-built cloud connectivity to the connected sensors. This platform is based on the ESP-8266 Wi-Fi module [12]. It has fundamental Machine Learning (ML) algorithms that can be simply incorporated with Bolt IoT projects for detecting sensor values. The reason behind choosing the Bolt IoT module is that it has the capacity to reduce the entire code size by about 80% [13]. Table I below portrays the specifications of the Bolt IoT Wi-Fi module [14] and Figure. 6 shows the pin details of the module.

Table I: Some Details of Bolt IoT Wi-Fi Module

Parameters	Features
Processing and Connectivity Module:	ESP8266 with custom firmware
MCU:	32-bit RISC : Tensilica Xtensa LX106
Power:	5V & GND pins or 5V/1A DC- Micro-USB port
Operating Voltage:	3.3V
GPIO pins:	5 Digital pins (0,1,2,3,4)
Clock Frequency (CPU):	80 MHz
PWM:	All 5 Digital pins capable of PWM
Boot Time:	Less than 1 (< 1) second

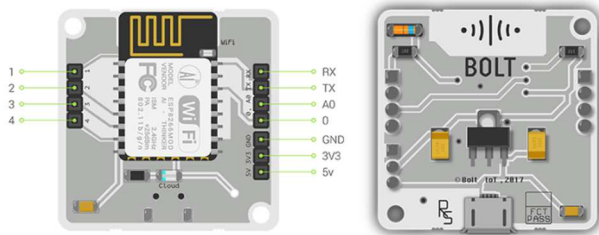


Figure 6: Pin Details of Bolt IoT Wi-Fi Module

B. Arduino Uno

Arduino Uno [15-17] is a single-board microcontroller based on the ATmega328 microcontroller. In this work, Arduino Uno plays a crucial role. It has 14 digital input and output pins, 6 analog inputs, a USB connection, a 16 MHz quartz crystal, a power jack, an ICSP header, a reset button, a USB connection, and a reset switch. It contains everything needed to help a microcontroller, simply interface it to a computer through a USB cable or battery or power it with an AC-to-DC adapter to get started.

C. Vibration Sensor

The vibration sensor module based on the vibration sensor SW-18010P is used to detect vibrations. When it is in static mode the switch is open circuit OFF-state, when external force to touch and corresponding vibration, or movement speed achieved at that instant, the circuit gets ON-state; when external force disappears, switch back to open circuit OFF-state. Here vibration sensors are used to take the vibration in the gear box of a wind turbine due to some internal faults or excessive wind [16]. This sensor helps to protect the costly wind turbine systems from unbearable internal conditions or odd atmospheric situations.

D. Infrared (IR) Sensor

An IR sensor is an electronic component that can transmit light to detect some object in its surroundings [13]. An infrared sensor (IR) can evaluate the heat of an object as well as detect motion. It includes an emitter (IR-Led) and receiver (Photo-diode). In this work, an IR sensor is used to detect the motion of the wind turbine (WT) prototype.

E. Voltage Sensor

Voltage sensors are utilized to monitor or measure and compute the amount of voltage [6]. This module is a basic and exceptionally valuable module that utilizes a potential divider to lessen the value of the input voltage by a factor of 5. For instance, with a 0V - 5V Analog info range, one can measure a voltage up to 25V.

F. Temperature Sensor

The Temperature & Humidity Sensor based on the DHT11 sensor module [12] with a digital signal output. Here this sensor gives updates on atmospheric conditions. It includes a resistive-type humidity measurement component and a Negative Temperature Coefficient (NTC) measurement component, and connects to the high-performance microcontroller, providing quick response, and cost-effectiveness.

G. Buzzer

A buzzer is an audio signaling electronic component; it might be piezoelectric or mechanical or electromechanical type. The main principle of the device is to convert the signal from audio to sound. In this very project buzzer is used as an alerting device when there is any fault in the gear box of WT it produces some vibration if that vibration is more than the threshold value then it gets interrupt and gives an alert signal.

H. Relay

A relay is an automatic switch that operates between the inputs operating voltage 0-5V [9]. In this proposed work relay plays a crucial role. When faults occur in the wind turbine relay automatically switches ON and the turbine gets disconnected from the main circuit to reduce the damages and after diagnosing the fault the WT turns ON and relay turns OFF in this very manner controlling of the system can be done.

V. RESULTS & DISCUSSIONS

In this IoT-based monitoring and controlling work, a DC motor is used as a prototype of a practical Wind Turbine. The complete hardware configuration is shown in figure 7. A 9V D.C voltage is taken from a variable voltage source. Here speeds will take from the infrared sensor, generated voltage will be taken from the respective voltage sensor, in the same way, information about the atmospheric weather and humidity will collect from temperature and humidity sensor and a vibration sensor is installed to take the vibration of gear box at every instant. The purpose of this initiative is to provide reliable condition monitoring and control of a costly wind energy generation system from a remote location.

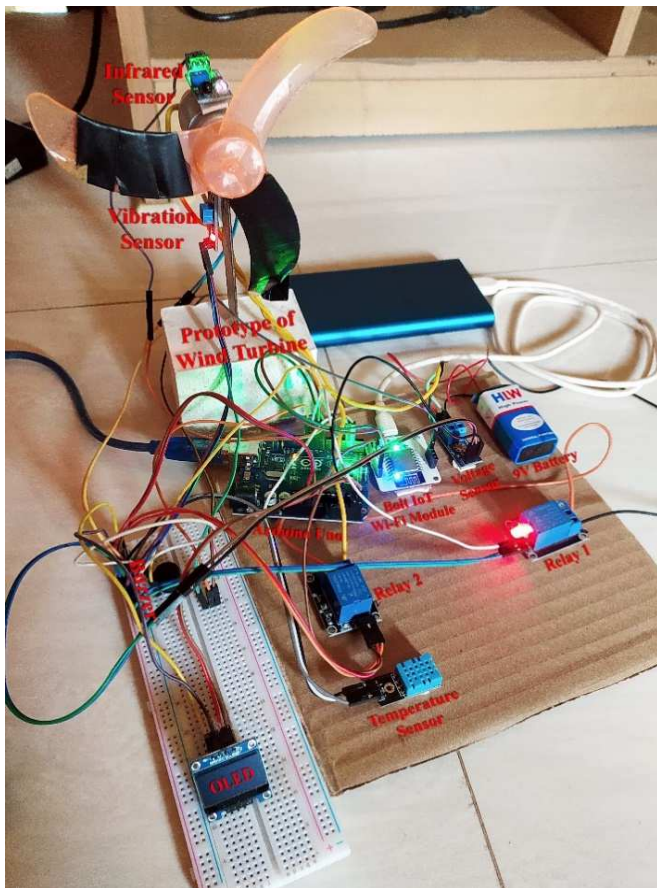


Figure 7: Prototype of the Proposed System

Wind turbine failures due to the cause of excessive vibration over the gear box and bearings, generally above 80km/hr. The wind speed causes massive vibration in the WT system. This problem can be overcome by counting vibrations. The vibration

sensor is programmed up to a threshold value of 100th count; after this very threshold value, the relay gets opened and the turbine gets stopped automatically to diagnose the fault in the turbine. At the same time, a notification is shown on the OLED as Vibration High and a preset message alert [18] will be sent to the operator's mobile/tablet/PC. After diagnosing the fault, the turbine turns ON automatically.

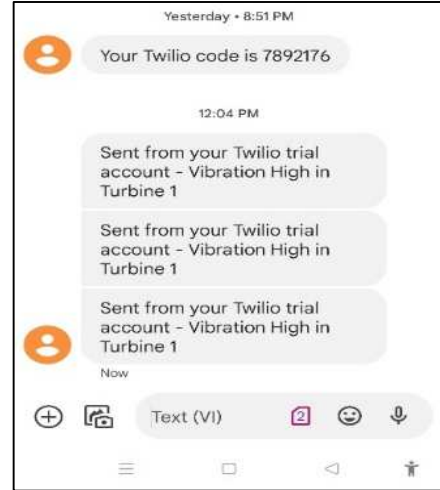


Figure 8: Fault SMS alert with Turbine Number [18]

Wind Turbine Monitoring & Control using IoT

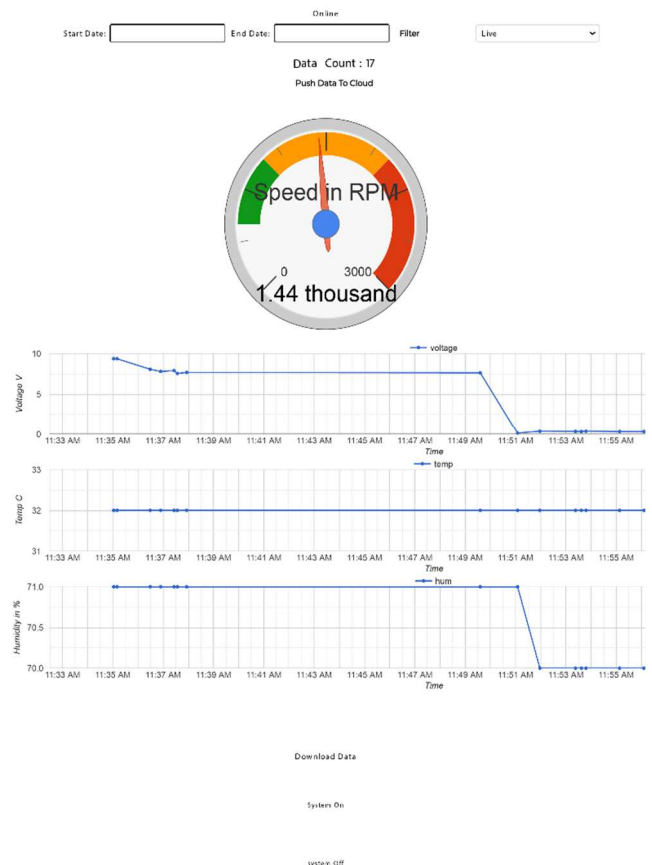


Figure 9: Data & graph obtain from Bolt IoT Cloud

To compute the turbine speed, an IR sensor is used; an IR sensor is attached to the rotor side of the WT (wind turbine) to calculate the wind turbine speed instantaneously. Vibration sensor will attach to the gearbox of the WT. At the moment, voltage and atmospheric temperature and humidity will also record in Bolt Cloud.

VI. CONCLUSION

Most of the existing systems utilize old memory cards or PCs for data logging. Only a specific PC alone can access the stored data. Today, with the assistance of IoT Technology, the restriction can be overcome. By the utilization of this proposed system, various parameters such as speed of wind turbine (WT), vibration level, power generation, voltage, currents, temperature, and humidity can be controlled, monitored effectively from anywhere, just by a web browser. The proposed system effectively reduces time, manpower requirements, and costs as well. Therefore, various industries can acquire this system, as this could be an attractive arrangement for wind turbine control & monitoring. In this system, turbine control is adopted based on the level of vibration, speed of the generator, generated voltage, atmospheric temperature and humidity. The IoT structure is implemented through Arduino uno microcontroller and Bolt IoT module. The proposed system secures system reliability and provides safe operation.

VII REFERENCES

[1] Z. Fu, M. Zhao, Y. Luo and Y. Yuan, "Self-healing strategy for wind turbine condition monitoring system based on wireless sensor networks," 2016 11th International Conference on Computer Science & Education (ICCSE), 2016, pp. 544-549, doi: 10.1109/ICCSE.2016.7581639.

[2] <https://www.ibef.org/download/Renewable-Energy-September-2020.pdf> [Accessed: 31- Mar-2022].

[3] Vishal Kumar Singh, Dr. Mallikarjun B.C, Rishav, Uttam Kumar Ray, 0, IOT based Windmill Monitoring System, International Journal Of Engineering Research & Technology (IJERT) NCESC – 2018 (Volume 6 – Issue 13),”.

[4] H F Liew et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 932 012080 “Wind Characterization By Three Blade Savonius Wind Turbine Using IoT” doi:10.1088/1757-899X/932/1/012080.

[5] S. J. Watson, B. J. Xiang, W. Yang, P. J. Tavner and C. J. Crabtree, "Condition Monitoring of the Power Output of Wind Turbine Generators Using Wavelets," in IEEE Transactions on Energy Conversion, vol. 25, no. 3, pp. 715-721, Sept. 2010, doi: 10.1109/TEC.2010.2040083.

[6] D. Kalyanraj, S. L. Prakash and S. Sabareswar, "Wind turbine monitoring and control systems using Internet of Things," 2016 21st Century Energy

Needs - Materials, Systems and Applications (ICTFCEN), 2016, pp. 1-4, doi: 10.1109/ICTFCEN.2016.8052714.

[7] B. Demircan and E. Akyüz , "IoT and Cloud Based Remote Monitoring of Wind Turbine", Celal Bayar University Journal of Science, vol. 15, no. 4, pp. 337-342, Dec. 2019, doi:10.18466/cbayarjbs.540812.

[8] W. Yang, P. J. Tavner, C. J. Crabtree and M. Wilkinson, "Cost-Effective Condition Monitoring for Wind Turbines," in IEEE Transactions on Industrial Electronics, vol. 57, no. 1, pp. 263-271, Jan. 2010, doi: 10.1109/TIE.2009.2032202.

[9] R.Raj Mohan Fran Lizza.M, Anitha.S International Journal Of Advanced Research In Management, Architecture, Technology And Engineering (IJARMATE) IJERTCONV6IS13225 ISSN (ONLINE): 2454-9762, ISSN (PRINT): 2454-9762 “Iot Based Wind Turbine Monitoring, Fault Diagnosis And Control Using Uart”.

[10] Mathew L. Wymore, Jeremy E. Van Dam, Halil Ceylan, Daji Qiao,"A survey of health monitoring systems for wind turbines, Renewable and Sustainable Energy Reviews," Volume 52, 2015, Pages 976-990, ISSN 1364-0321, doi.org/10.1016/j.rser.2015.07.110.

[11] Murtala Zungeru, Adamu; Chuma, Joseph M.; Lebekwe, Caspar K.; Phalaagae, Pendukeni; Gaboitaolelwe, Jwaone (2020). Green Internet of Things Sensor Networks (Applications, Communication Technologies, and Security Challenges) : Springer, 10.1007/978-3-030-54983-1(), -. doi:10.1007/978-3-030-54983-1.

[12] S. M. Sorif, D. Saha and P. Dutta, "Smart Street Light Management System with Automatic Brightness Adjustment Using Bolt IoT Platform," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp.1-6, doi:10.1109/IEMTRONICS52119.2021.9422668.

[13] D. Saha, S. M. Sorif and P. Dutta, "Weather Adaptive Intelligent Street Lighting System With Automatic Fault Management Using Boltuino Platform," 2021 International Conference on ICT for Smart Society (ICISS), 2021, pp. 1-6, doi: 10.1109/ICISS53185.2021.9533234.

[14] B. I. T . (I. P. Limited), “IoT Platform,” BOLT. [Online]. Available: <https://www.bolttiot.com/techspecs>. [Accessed: 31- Mar-2022].

[15] M. S. Pramod, P. N. Naveen, N. R. Chaitra, K. Ranjith and G. H. S. Vikas, "Monitoring of highway wind power parameter and controlling highway light through IOT," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017, pp. 1750-1753, doi: 10.1109/RTEICT.2017.8256899.

[16] Fausto Pedro García Márquez, Andrew Mark Tobias, Jesús María Pinar Pérez, Mayorkinos Papaelias, "Condition monitoring of wind turbines: Techniques and methods, Renewable Energy" Volume 46, 2012, Pages 169-178, ISSN 0960-1481, doi.org/10.1016/j.renene.2012.03.003.

[17] Q.F. Hassan, “Internet of Things A to Z Technologies and Applications”, ISBN:978-1-119-45674-2: Wiley.

[18] “Communication APIs for SMS, Voice, Video and Authentication,”Twilio[Online].Available: <https://www.twilio.com/>. [Accessed: 31- Mar-2022].

Sentiment Analysis and NLP models for Identifying Biases of Online News Stations

Anuska Acharya
aachary4@gmu.edu

Grace Cox
gcox4@gmu.edu

Abstract—This work attempts to identify potential reporting bias surrounding recent controversial decisions for articles published from August 2019 to October 2021 within four major news organizations: FOX, CNN, NBC, and NPR. This potential bias is determined by conducting a Sentiment Analysis using NLTK’s (Natural Language Tool Kit) VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer. Copious amounts of literature have been published regarding sentiment analysis and bias identification of news articles, though none employ VADER. The team determines the overall sentiment for an article using the Polarity Compound calculated by the Sentiment Intensity Analyzer, which then corresponds to a political tone indicated within the verbiage and context of the article. Upon completion of the analysis, it was found that CNN, NBC, and NPR tend to have the most negative sentiment surrounding this topic, while FOX tends to be more neutral though still on the positive side. This translates to the surprising identification of a slightly democrat tone for articles published by FOX, and a more republican tone for those articles published by NPR, CNN, and NBC.

Index Terms—Machine learning, Sentiment analysis, NLP, News

I. PROBLEM DESCRIPTION

The U.S. War in Afghanistan informally began in 1999 when the United Nations declared al-Qaeda and the Taliban to be terrorist entities and, “impose[d] sanctions on their funding, travel, and arms shipments” [1]. With that being said, the United States became formally involved in the War in Afghanistan following the Terrorist Attacks on the World Trade Center and Pentagon on September 11, 2001, that were sparked by the assassination of Ahmad Shah Massoud, the commander of an anti-Taliban coalition called the Northern Alliance. The September 11 terrorist attacks on the United States resulted in a rapid increase in the number

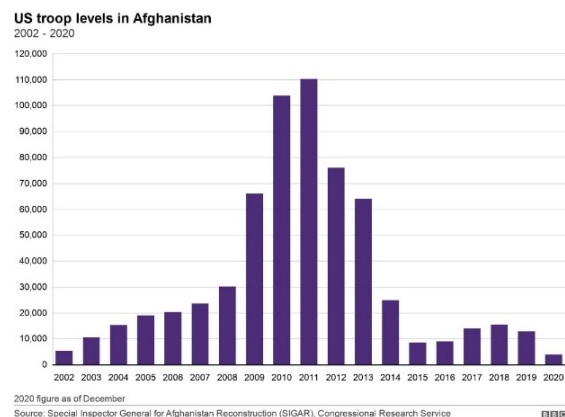


Fig. 1: Levels of US Troops in Afghanistan from 2002-2020

of troops on the ground in Afghanistan, up until 2012 when U.S. involvement in the middle east began to see a rapid decline (Figure 1). Although the U.S. has been decreasing the troop levels in Afghanistan since 2012, the complete removal of all American troops from Afghanistan on August 30, 2021, sparked major political upset, both in the Middle East and here in the United States.

The following diagram (Figure 2) displays the timeline of events surrounding the United States’ complete withdrawal from Afghanistan, beginning with the signing of the Doha Agreement by President Donald Trump and the Taliban in February 2020, and ending with Joe Biden’s decision to pull all remaining troops from Afghanistan on August 30, 2021 [2].

Being that the United States’ complete withdrawal from Afghanistan is major international news, there is value in understanding the potential biases that lie within the many news organizations reporting on this topic throughout the United States.

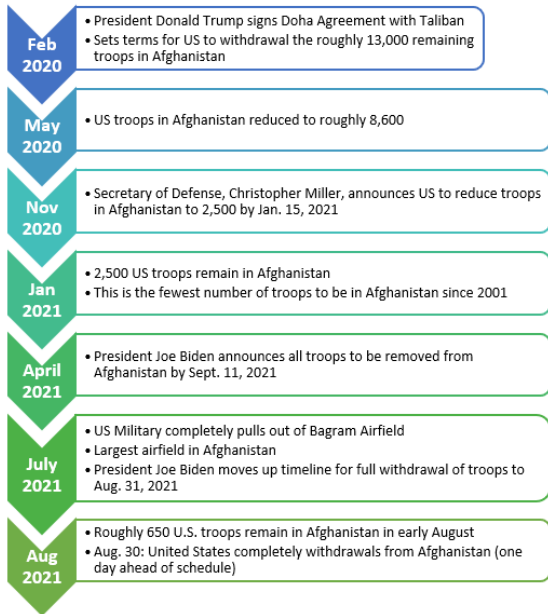


Fig. 2: Timeline of the United States' Withdrawal from Afghanistan

This project will focus conducting a sentiment analysis of 30 news articles written and released by each of four major news stations, to include: FOX, NBC, CNN, and ABC, regarding the United States' withdrawal from Afghanistan earlier this year. This sentiment analysis will aid in identifying potential biases that lie within these different organizations when reporting on a major political event.

II. IMPORTANCE OF PROBLEM

The decision by the U.S. to withdrawal from Afghanistan is a history-making decision that has ended the United States' nearly 20-year involvement in the war against terrorism, al-Qaeda, and the Taliban in Afghanistan and was met with mixed reactions of support and opposition from the general American public (Figure 3) [3]. We note that roughly 70% of individuals identifying as Democrat and Independent support this decision, mirroring the 70% overall support this decision received. This is contrasted by the near split in support/opposition responses seen within individuals who identify as Republican.

The Republican and Democrat parties appear to be on opposing sides of this issue, with the majority

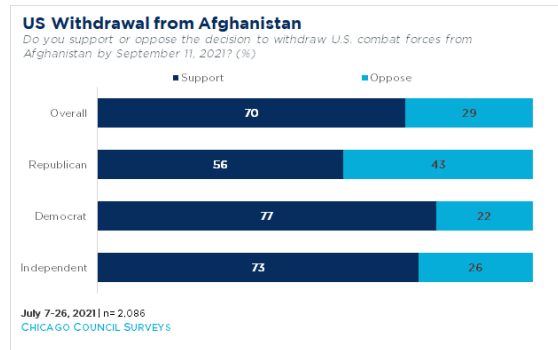


Fig. 3: Political Support regarding the US Withdrawal from Afghanistan

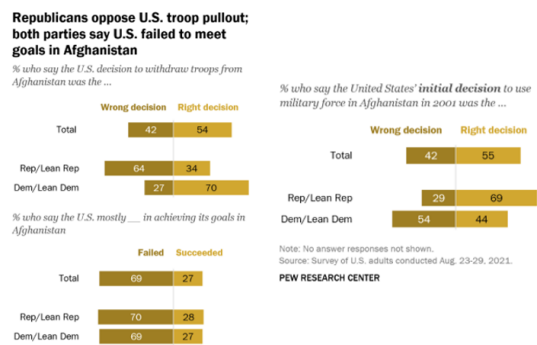


Fig. 4: PEW Research Center Poll regarding US Troop Pullout

of individuals with Republican leanings believing that the United States made the wrong decision in completely pulling out of Afghanistan and the majority of Democrats believing that the U.S. decision to withdraw troops from Afghanistan was the right decision (Figure 4) [4]. Despite the public's varying levels of support surrounding the United States' decision to withdrawal from Afghanistan, there has been an abundance of news coverage regarding this hot topic. Machine learning models for fake news detection and text analysis could help to better understand the online posts [?], [?], [?], [5]–[17].

III. PRELIMINARY LITERATURE REVIEW

Many researchers have worked in sentiment analysis on different news stations. Research conducted by Reis, Benevenuto, Melp, Prates, Kwak collected a total of 69,907 headlines from four different news sources: BBC News Online, Daily Mail Online, The

New York Times, and Reuters Online [18]. The primary purpose of the analysis was to identify the sentiment of the headlines of various news articles produced by these big and popular news sources. The research focused on identifying if such positive or negative sentiment of the headlines was able to generate more clicks. This research found that the polarity of the headlines of the article impacted the popularity of the news article leading to more clicks. The study also revealed the headline with a positive or negative tone attracted more readers compared to the headlines with a neutral sentiment.

Similarly, in another research article by Islam, Ashraf, Abir and Mottailb [19] conducted sentiment analysis to detect the polarity based on sentence structure and dynamic dictionary. The research used detection of sentences and the use of a library for the classification of the news article. The library was defined using a list of reserved words. The proposed algorithm selected online news articles and extracted paragraphs, sentences, phrases, and words. The end of the sentence was detected if the sentence included the “.” full stop sign. Once the end of the sentence was detected, the algorithm then searched for positive or negative words, sentences and phrases and determined the polarity of the news articles. The researchers utilized java programming languages and NetBeans IDE to develop the interface for the algorithm. The experiment included the extraction of 56 random news articles from The Independent, The Telegraph, and The Daily Star which resulted in only an 8.93% margin of error which is only 5 out of 56 articles. Further analysis of the error showed these articles had a smaller number of sentences which resulted in difficulty in determining the polarity.

A research article by Jagdale, Deshmukh, and Shirsath [20] explained the three levels of sentiment analysis and advanced online text analysis [21]–[33]. These levels are document level, sentence level, and entity and aspect level. In document-level sentiment analysis, the whole document is analyzed, and the polarity of the entire document is identified. Similarly, in sentence-level sentiment analysis, each and all sentences are analyzed to determine the polarity; positive, negative, or neutral. Entity and aspect level sentiment analysis is based on opinion. The research involved 2225 documents

TABLE I

News Organizations Covered in Analysis and Number of Articles Used	
News Organization	Number of Articles Scraped
FOX	20
NBC	20
CNN	20
NPR	20

TABLE II

Sentiment Analyses to be Conducted on News Articles Covering U.S. Withdrawal from Afghanistan	
1	2
Positive	Democrat
Negative	Republican
Neutral	Independent

from the BBC. The methodology used in this research was tokenization, stop words, stemming, and then assigning scores based on sentiments.

IV. PROPOSED APPROACH

To complete our Sentiment Analysis successfully and adequately, the team will focus on four major news organizations, previously listed, and will scrape a sufficient number of articles from the websites of each organization. The following table (Table I) provides an overview of the four major news organizations analyzed throughout this project as well as the number of articles regarding the United States’ withdrawal from Afghanistan that were scraped for use throughout our Sentiment Analysis.

Two types of Sentiment Analysis (Table 2) will be conducted throughout this project. The first is a general Sentiment Analysis that describes the overall tone of the article as either Positive, Negative, or Neutral, which will aid the team in determining the overall attitude within each news organization regarding the United States’ complete withdrawal from Afghanistan. The second Sentiment Analysis will describe the political tone of the article as being Democrat, Republican, or Independent, which will aid in highlighting the overall political biases present within each news organization surrounding the United States’ withdrawal from Afghanistan.

V. PROPOSED METHOD FOR EVALUATION

The team intends to take advantage of various Natural Language Processing (NLP) techniques using programs such as Python and R to conduct a

sentiment analysis on each of the 120 articles covering the United States' withdrawal from Afghanistan published by one of four news stations (FOX, CNN, NBC, NPR). The team will then use the results of this sentiment analysis to identify any potential biases surrounding major political events that are apparent within each of these major news organizations. Several text preprocessing steps must be completed on all text to successfully complete a sentiment analysis on these articles published by each of the four major news organizations. These text preprocessing steps can include tokenization, stop word and punctuation removal, lemmatization, stemming and more, and are described in detail below.

A. Text Preprocessing Definition

Preprocessing text data refers to the process of transforming the text input into a, "predictable and analyzable" [34] form for the task at hand and the ultimate goal of cleaning and preprocessing text data is to, "... reduce the text to only the words that you need for your NLP goals" [35]. It should be noted that different tasks require an emphasis on different steps within the preprocessing procedure. Some common types of text preprocessing techniques include tokenization, punctuation/noise removal, and text normalization (lowercase tokens, stop word removal, and lemmatization/stemming).

B. Tokenization

Tokenization refers to the preprocessing task of, "breaking up text into smaller components of text (known as tokens)" [36]. A token may be an entire word, a part of a word, or characters like punctuation. Tokenization is one of the most important parts of NLP preprocessing, as it defined what our models can express. In this project, tokenization was done using SpaCy [37], a Python package often used in NLP settings. SpaCy not only provides generic tokenization functions, but also allows the user to, "customize the tokenization process to detect tokens on custom characters" [38]. This custom tokenization in SpaCy could be used for words including hyphens or apostrophes that should be processed as a single token.

TABLE III

Raw Text Data	Lowercase Text Data
CaMeL	camel
UPPER	upper
lower	lower

C. Punctuation & Noise Removal

Punctuation in text data does not add much, if any, value to the data and the meaning behind it and thus is typically removed from the raw textual data during the preprocessing steps. Punctuation includes characters including, but not limited to commas, periods, exclamation/question marks, hyphens, and apostrophes.

D. Lowercase Tokens

Text data usually contains characters that have different cases, some of which may not be conducive to the Natural Language Processing procedure. In order to further normalize textual data for ease of analysis, all tokens are typically converted into a lowercase capitalization scheme. The following table (Table III) displays a few examples for creating lowercase tokens from raw textual data.

E. Stop Word Removal

Stop words are commonly used words in a language that provide little to no information to the text. Some examples of stop words in the English language can include, but are not limited to: "the", "is", "a", "are" [34]. There are many packages available within Python that are capable of detecting and removing stop words from the provided text, with the most popular being the NLTK package. Stop word removal is one of few text preprocessing tasks that are used for text normalization.

F. Lemmatization & Stemming

Stemming refers to the NLP preprocessing task that is concerned with, "bluntly removing word affixes (prefixes and suffixes)" [2]. There are two major errors that can arise from Stemming Algorithms: Over Stemming and Under Stemming. Over stemming occurs when two words that have different stems are stemmed to the same root word; for example, the words "universal", "university", and "universe" are stemmed to "univers" which

```
[am, is, are] → be
[walk, walked, walking] → walk
[watches, watching, watched] → watch
```

Fig. 5: Lemmatization Examples

would not be correct since their modern meanings are in different domains and are generally not synonymous [39]. Under stemming occurs when, “two words that should be stemmed to the same root are not” [39]. For example, the words “alumnus”, “alumni”, and “alumnae” should all be stemmed to the same word, but typically are not. Textual data can contain tokens that are different forms of a certain word (i.e. walk, walked, walking) and condensing all of the tokens in a text to their root word increases the ease of analysis while at the same time reducing inflectional forms. Lemmatization refers to this process of taking each token and bringing it down into its root form. The goal of lemmatization is to “... remove inflectional endings only and to return the base or dictionary form of a word” [40]. The dictionary form of a particular word is referred to as the lemma for that word. The example below (Figure 5) displays lists of possible words contained in text data as well as the respective lemma for each list of words.

VI. RESULTS

Prior to conducting this sentiment analysis, date metadata was scraped from each of the 80 articles to determine the month and year that each article was published in (Figure 6). This date metadata was provided in a schema.org annotation format within @type: NewsArticle and will provide the team with further context for the sentiment analysis to be conducted. We note all articles, except for one, were published in 2021, with the majority of articles being published in August and September 2021. This coincides with the previously mentioned timeline regarding the United States’ complete withdrawal from Afghanistan.

This sentiment analysis of four major news organizations is based on 20 articles published by each of FOX, NBC, CNN, and NPR regarding the United States’ withdrawal from Afghanistan; the links for each of the articles used within this analysis can be found in Table 1 within Appendix A. It should

be noted that though the team initially explored this text data upon completion of textual pre-processing, due to time constraints, the team decided to approach the problem from a different angle. Each webpage is scraped using the BeautifulSoup and urlopen libraries within Python, then using NLTK’s VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer, a Polarity Compound score indicating the articles overall sentiment is calculated. This Polarity Compound score is a normalized sum of the Positive, Negative, and Neutral scores that ranges in value from -1 to 1, with positive values indicating a positive sentiment, negative values indicating a negative sentiment, and values close to 0 indicating a more neutral sentiment. It should be noted that VADER’s Sentiment Intensity Analyzer is pretrained and thus does not require any training effort, making it ideal considering the time constraints placed on this project.

The following Figure 7 displays the Polarity Compound scores for each of the 20 articles regarding the United States’ withdrawal from Afghanistan that were published online by FOX News. This figure displays that FOX is relatively split between articles classified as having a positive sentiment and those having a negative sentiment. With that being said, the articles with positive sentiment tend to have a stronger positive sentiment than those articles with negative sentiments, in which case we see weaker negative sentiments (indicated by polarity compound scores that are closer to 0 e.g., -0.158, -0.271).

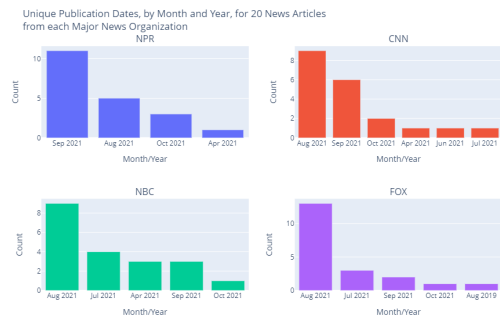


Fig. 6: Publication Dates for Scraped Articles, by News Organization

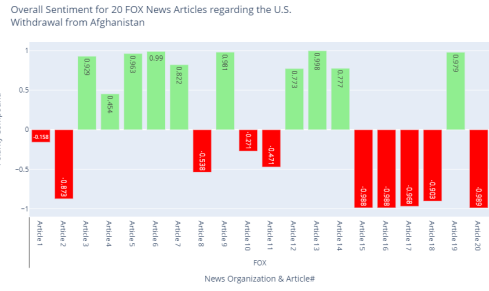


Fig. 7: Sentiment Analysis for FOX News Articles

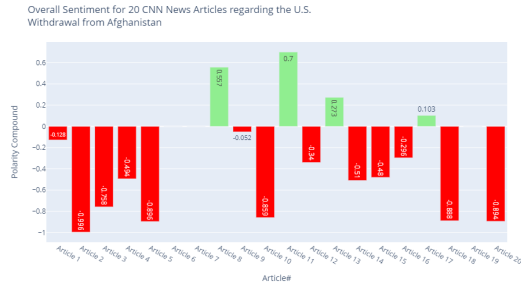


Fig. 9: Sentiment Analysis for CNN News Articles



Fig. 8: Sentiment Analysis for NBC News Articles

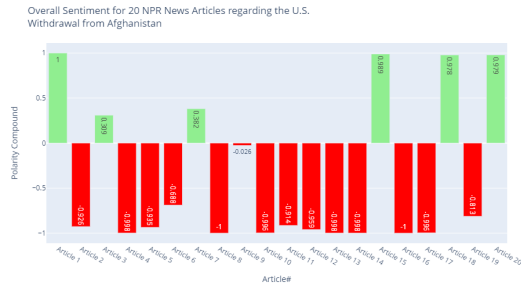


Fig. 10: Sentiment Analysis for NPR News Articles

The following Figure 8 displays the Polarity Compound scores for each of the 20 articles regarding the United States’ withdrawal from Afghanistan that were published online by NBC News. This figure displays that NBC tends to have a largely negative sentiment within articles regarding this controversial political decision. There were only two articles out of the 20 that were used within this analysis that were classified as having a positive sentiment. We note that of the selected articles published, the majority have a very strong negative sentiment, indicated by Polarity Compound Scores that are either -1 or very close to -1.

The following Figure 9 displays the Polarity Compound scores for each of the 20 articles regarding the United States’ withdrawal from Afghanistan that were published online by CNN News. We note that, like NBC, CNN tends to have a largely negative sentiment within the published articles regarding this monumental decision. Only four of the 20 articles were classified as having a positive sentiment. Also, we noted that the articles with a negative sentiment are typically stronger than those

with a positive sentiment for the articles published by CNN.

The following Figure 10 displays the Polarity Compound scores for each of the 20 articles regarding the United States’ withdrawal from Afghanistan that were published online by NPR. We note that, like NBC and CNN, NPR tends to have a largely negative sentiment within the published articles regarding this monumental decision. Only six of the 20 articles were classified as having a positive sentiment. Also, we noted that the articles with a negative sentiment are typically stronger than those with a positive sentiment, and most of the Polarity Compound scores for these articles are both very strong (i.e. close to 1) and negative.

The following Figure 11 displays the Average Sentiment for the four major news organizations FOX, NBC, CNN, and NPR. The Average Sentiment of all 20 articles is determined by calculating the average polarity compound score across all articles published by each of the major news organizations previously mentioned. This figure aids in highlighting the overall bias that is present within

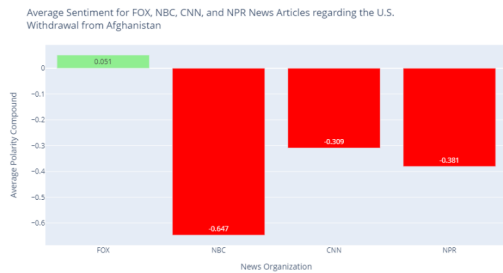


Fig. 11: Average Sentiment Analysis by News Organization

the articles published by FOX, NBC, CNN, and NPR regarding the United States’ withdrawal from Afghanistan. We see that while FOX tends to be relatively neutral when reporting on this topic, there is still a slightly positive overall bias present. In contrast, NBC, CNN, and NPR all have a negative bias, with NBC having a strong negative bias when reporting on the United States’ decision to withdrawal completely from Afghanistan, indicating their strong disapproval of this decision, and CNN & NPR having a relatively moderate negative bias’.

In addition to determining the overall general sentiment, we want to determine whether the articles published by each of these four major news organizations tend to have a Democrat, Republican, or Neutral leaning. Figure 12 depicts the Polarity Compound scores for each of the 20 articles for ABC, CNN, NPR, and FOX, with blue shades indicating a more Democrat leaning, red shades indicating a more Republican leaning, and white shades indicating a Neutral leaning. It should be noted that the political leaning of an article is determined by the value of its polarity compound calculated by VADER. Polarity Compound values closer to 1 indicate a democrat tone and those values closer to -1 indicate a republican tone. CNN appears to have the most neutral articles (though still Republican in leaning), with only two articles having strong Democrat leanings, while NBC and NPR are clearly the most Republican leaning in nature as there are more deep red shades present in Figure 10. One surprising result from both Figure 11 and Figure 12 is that FOX typically tends to be more Republican in nature, while NBC and CNN tend to lean more Democrat. This result could be

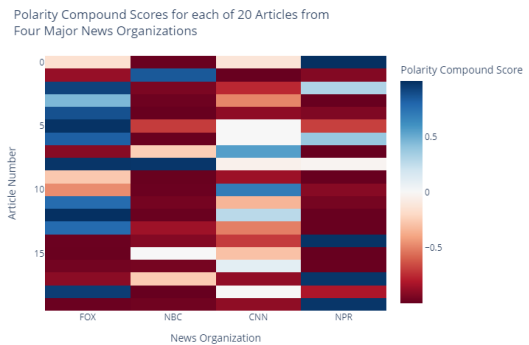


Fig. 12: Political Tone Analysis for Individual Articles by News Organization

explained by the fact that there was a change of presidency during the process of the United States’ withdrawal from Afghanistan. Another explanation for this surprising result for FOX News stems from the information displayed above in Figure 3, which shows that Republicans are fairly split regarding the United States’ decision to completed withdrawal from Afghanistan, with only 56% of surveyed republicans supporting this decision.

VII. FUTURE WORK

One potential area of future work regarding discovering biases within articles published by major news organization regarding the United States’ complete withdrawal from Afghanistan includes creating a model that is unique to the verbiage and context of these articles. Due to the time constraints of the course, the team was unable to both train and employ a Natural Language Processing model with adequate accuracy. Since the team decided to use the Polarity Compound scores for each article to represent the articles political tone, another area for future work would entail creating dictionaries and lists of keywords that are representative of each political leaning (Democrat, Republican, or Independent).

REFERENCES

[1] C. on foreign relations. “the u.s. war in afghanistan”. [Accessed December 9, 2021]. [Online]. Available: <https://www.cfr.org/timeline/us-war-afghanistan>

- [2] E. Kiely and R. Farley. "timeline of u.s. withdrawal from afghanistan". [Online]. Available: <https://www.factcheck.org/2021/08/timeline-of-u-s-withdrawal-from-afghanistan/>
- [3] D. Smeltz and E. Sullivan. (August 9, 2021) "us public supports withdrawal from afghanistan". [Online]. Available: <https://www.thechicagocouncil.org/commentary-and-analysis/blogs/us-public-supports-withdrawal-afghanistan>
- [4] T. V. Green and C. Doherty. "majority of u.s. public favors afghanistan troop withdrawal; biden criticized for his handling of situation". [Online]. Available: <https://www.pewresearch.org/fact-tank/2021/08/31/majority-of-u-s-public-favors-afghanistan-troop-withdrawal-biden-criticized-for-his-handling-of-situation/>
- [5] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, "Addressing health-related misinformation on social media," vol. 320, no. 23, p. 2417, Dec. 2018. [Online]. Available: <https://doi.org/10.1001/jama.2018.16865>
- [6] L. Cui and D. Lee, "Coaid: COVID-19 healthcare misinformation dataset." *CoRR*, vol. abs/2006.00885, 2020. [Online]. Available: <https://arxiv.org/abs/2006.00885>
- [7] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, 2017, pp. 963–972. [Online]. Available: <https://doi.org/10.1145/3041021.3055135>
- [8] J. H. Fetzer, "Disinformation: The use of false information," vol. 14, no. 2, pp. 231–240, May 2004. [Online]. Available: <https://doi.org/10.1023/b:mind.0000021683.28604.5b>
- [9] E. Dolgin, "COVID vaccine immunity is waning — how much does that matter?" *Nature*, vol. 597, no. 7878, pp. 606–607, Sep. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02532-4>
- [10] U.S. Food and Drug Administration. "covid-19 vaccines". [Accessed November 6, 2021]. [Online]. Available: <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines>
- [11] N. Sallahi, H. Park, F. E. Mellouhi, M. Rachdi, I. Ouassou, S. Belhaouari, A. Arredouani, and H. Bensmail, "Using unstated cases to correct for COVID-19 pandemic outbreak and its impact on easing the intervention for qatar," *Biology*, vol. 10, no. 6, p. 463, May 2021. [Online]. Available: <https://doi.org/10.3390/biology10060463>
- [12] M. El-Harbawi, B. B. Samir, M.-R. Babaa, and M. I. A. Mutalib, "A new QSPR model for predicting the densities of ionic liquids," *Arabian Journal for Science and Engineering*, vol. 39, no. 9, pp. 6767–6775, Jun. 2014. [Online]. Available: <https://doi.org/10.1007/s13369-014-1223-3>
- [13] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," vol. 359, no. 6380, pp. 1094–1096, Mar. 2018. [Online]. Available: <https://doi.org/10.1126/science.aao2998>
- [14] Q. Su, M. Wan, X. Liu, and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective," vol. 1, no. 1-2, p. 1, 2020. [Online]. Available: <https://doi.org/10.2991/nlpr.d.200522.001>
- [15] S. Akon and A. Bhuiyan, "Covid-19: Rumors and youth vulnerabilities in bangladesh," 07 2020.
- [16] M. Fernandez and H. Alani, "Online misinformation." ACM Press, 2018. [Online]. Available: <https://doi.org/10.1145/3184558.3188730>
- [17] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai, "Fight under uncertainty: Restraining misinformation and pushing out the truth." *IEEE*, Aug. 2018. [Online]. Available: <https://doi.org/10.1109/asonam.2018.8508402>
- [18] J. dos Reis, F. Benevenuto, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," *CoRR*, vol. abs/1503.07921, 2015. [Online]. Available: <http://arxiv.org/abs/1503.07921>
- [19] M. U. Islam, F. B. Ashraf, A. I. Abir, and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," in *2017 20th International Conference of Computer and Information Technology (ICCIIT)*. IEEE, Dec. 2017. [Online]. Available: <https://doi.org/10.1109/iccitechn.2017.8281777>
- [20] V. S. Shirsat, R. S. Jagdale, and S. N. Deshmukh, "Document level sentiment analysis from news articles," in *2017 International Conference on Computing, Communication, Control and Automation (ICCCUBEA)*. IEEE, Aug. 2017. [Online]. Available: <https://doi.org/10.1109/iccubea.2017.8463638>
- [21] M. Heidari, S. Zad, M. Malekzadeh, P. Hajibabae, S. HekmatiAthar, O. Uzuner, and J. H. J. Jones, "Bert model for fake news detection based on social bot activities in the covid-19 pandemic," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.
- [22] A. Ain, "The WHO is right to call a temporary halt to COVID vaccine boosters," *Nature*, vol. 596, no. 7872, pp. 317–317, Aug. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02219-w>
- [23] E. Callaway, "COVID vaccine boosters: the most important questions," *Nature*, vol. 596, no. 7871, pp. 178–180, Aug. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02158-6>
- [24] A. Weatheron. "health expert says booster shot could be needed after getting covid-19 vaccine". [Accessed June 8, 2021]. [Online]. Available: <https://www.13newsnow.com/article/life/booster-shot-may-be-needed-after-covid-19-vaccine/291-49a8966c-3d91-48ad-99a0-02905c5593cc>
- [25] M. Malekzadeh, P. Hajibabae, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "Review of graph neural network in text classification," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.
- [26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational

- Linguistics, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
- [27] P. Hajibabae, M. Malekzadeh, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, “An empirical study of the graphsage and word2vec algorithms for graph multiclass classification,” in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.
- [28] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, “Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection,” *Nature Medicine*, vol. 27, no. 7, pp. 1205–1211, May 2021. [Online]. Available: <https://doi.org/10.1038/s41591-021-01377-8>
- [29] J. Havey. “pharma research progress hope;”. [Online]. Available: https://catalyst.phrma.org/a-year-and-a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm_campaign=2021-q3-cov-innutm_medium=pai_srh_cpc-ggl-adtutm_source=gglutm_content=clk-pol-tpv_scl-geo_std-usa-dca-pai_srh_cpc-ggl-
- [30] P. R. Krause, T. R. Fleming, R. Peto, I. M. Longini, J. P. Figueroa, J. A. C. Sterne, A. Cravioto, H. Rees, J. P. T. Higgins, I. Boutron, H. Pan, M. F. Gruber, N. Arora, F. Kazi, R. Gaspar, S. Swaminathan, M. J. Ryan, and A.-M. Henao-Restrepo, “Considerations in boosting COVID-19 vaccine immune responses,” *The Lancet*, vol. 398, no. 10308, pp. 1377–1380, Oct. 2021. [Online]. Available: [https://doi.org/10.1016/s0140-6736\(21\)02046-8](https://doi.org/10.1016/s0140-6736(21)02046-8)
- [31] J. H. Kim, F. Marks, and J. D. Clemens, “Looking beyond COVID-19 vaccine phase 3 trials,” *Nature Medicine*, vol. 27, no. 2, pp. 205–211, Jan. 2021. [Online]. Available: <https://doi.org/10.1038/s41591-021-01230-y>
- [32] E. C. Fernández and L. Y. Zhu, “Racing to immunity: Journey to a COVID-19 vaccine and lessons for the future,” *British Journal of Clinical Pharmacology*, vol. 87, no. 9, pp. 3408–3424, Jan. 2021. [Online]. Available: <https://doi.org/10.1111/bcp.14686>
- [33] S. Zad, M. Heidari, P. Hajibabae, and M. Malekzadeh, “A survey of deep learning methods on semantic similarity and sentence modeling,” in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.
- [34] K. Ganesan. (April 2019) “all you need to know about text preprocessing for nlp and machine learning.”. [Online]. Available: <https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html>
- [35] Code Academy. “text preprocessing”. [Online]. Available: <https://www.codecademy.com/courses/text-preprocessing/lessons/text-preprocessing/exercises/introduction>
- [36] ——. “natural language processing/text preprocessing”. [Online]. Available: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-text-preprocessing/cheatsheet>
- [37] spaCy. “spacy 101: Everything you need to know”. [Online]. Available: <https://spacy.io/usage/spacy-101>
- [38] Real Python. “tokenization in spacy”. [Online]. Available: <https://realpython.com/natural-language-processing-spacy-python/#tokenization-in-spacy>
- [39] T. Srivastava. (August 6, 2019) “nlp: A quick guide to stemming.”. [Online]. Available: <https://medium.com/@tusharsri/nlp-a-quick-guide-to-stemming-60f1ca5db49e>
- [40] stanford. (2009) “stemming and lemmatization”. [Online]. Available: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

Machine Learning models for Customer Relationship Analysis to Improve Satisfaction Rate in Banking

Patricia Fabijanczyk

Lahari Bagam

Nan Jia

Ebrima Ceesay

Abstract—The purpose of this research project was to analyze customer complaint data from financial institutions and identify areas of opportunity for these institutions to improve their customer satisfaction rate. In addition to pointing out areas for improvement, this paper also looks into similar research and tries to understand if themes found in this analysis are consistent with those done by other researchers. Banking is an essential piece to everyday life for all people across the world. Banks need to ensure that their products and processes are simple and accessible to all. Although banks have a monopoly on our financial needs their desire to retain existing customers and gain new ones drives the necessity of providing excellent and timely customer service.

The study was conducted using a dataset of over two million customer complaint records and examining what were the top three financial institutions receiving complaints and which products received them. In addition, other aspects of complaints such as state of origination was also looked at. Analysis was done using machine learning, python, tableau and other tools to show the data points and their correlation. Understanding the top financial institutions methods of handling customer complaints, we are able to make recommendations for further product improvements to increase customer satisfaction. Concluding the research project is a list of challenges and opportunities for further research projects. In addition, there are recommendations for the financial institutions investigated in this project on how to move forward from analyzing customer complaint data.

Index Terms—Machine learning, Finance, NLP, online information

I. INTRODUCTION

With the development and increased use of technology providing feedback has become instant and easy. Feedback can be negative or positive but to companies it should be what they constantly ask

for and want to receive. Customer feedback can also make or break a company [1]. With either positive or negative customer feedback a company can capitalize on what customers offer to them via multiple channels and incorporate it back into their innovation chain. Time and time again we have seen companies that do not listen to their customers fail – whether it be they continue to exist but struggle to gain momentum or they fail completely and are forced to shut their doors. One prime example of this is the sad fate of Blockbuster; although they had plenty of opportunities to reimagine the way they brought movies to families across the United States they took the stance that they knew their customers and they would stick to what they knew – brick and mortar stores with VHS tapes [1]. Netflix entered the scene and completely changed the landscape of at-home entertainment and eliminated their first competition, Blockbuster [1]. Although it is difficult to compare services between Netflix and Blockbuster to financial institutions, fundamentally they serve to provide a service to us the customers, whether it be entertainment or keeping our money secure in a checking account or helping us finance a home [1]. The Blockbuster example is a prime example of why companies should ask for and utilize customer feedback. Financial institutions are a staple to our lives but with banks competing to differentiate themselves there are institutions that can fold to others by not evolving to meet customer needs.

A. Research Paper's Structure

This research paper is structured to explain the importance of customer satisfaction to a financial

institution. Section two focuses on explaining the problem in detail and its significance to the market. Section three highlights the dataset used to conduct the analysis and how it was prepared for processing. Section four discusses the approach and methods used to analyze the large dataset. Section five, and six dive deeper into published research found over the course of the project that guided and informed this research. Section seven investigates the data analysis and tools used to conduct the research, such as using machine learning to develop a series of graphs from the data set. Section ten rounds off the research project by highlighting some challenges in the project and opportunities to expand the research while provide recommendations to the top three banks on where focus should be for improvements within their product lines to ensure customer satisfaction remains high and drives higher customer retention in return.

Banking is a necessity everyone must interact with on a regular basis. The days of cash only transactions are over and financial institutions Capital One, Bank of America, Wells Fargo, BB&T just to name a few, now take on the role of intermediary when it comes to us and our assets. Each of the financial institutions given as an example and many more provide us with a number of services when it comes to finances – checking and savings accounts, loans, mortgages, credit cards. If we do not like something in the services, we receive we naturally turn to the company we are experiencing problems with and voice our dislikes in hopes that it will alter the way we do business with them moving forward and avoid similar issues in the future. Complaining to a financial institution can be done through multiple channels – via the web, over the phone, even sometimes over their app or on social media.

The market today offers numerous companies to pick from when it comes to banking – there are small town banks, credit unions, and large financial institutions that offer a variety of products to customers. When customers have a number of options companies strive to differentiate themselves in the market and win customers – either by providing competitive rates for loans, checking rates with a higher than average interest rate, rates on mortgages, incentives for opening up credit cards, and

quality customer service [2]. Customer satisfaction counts for a lot today than all of the products a company can provide [2]. “If your customer is not satisfied, he or she will stop doing business with you [2].” Customer satisfaction is defined as the customer’s “perception that his or her expectations have been met or surpassed [2].” For banking, customers expect their funds to be available to them when needed, that includes making sure the technology that supports customers is working properly. When customers have a good experience with a company with a product it is more likely that they will return the next time they have a need for that product or a new one [2]. Financial institutions want customers to build portfolios with them, it results in greater profits for them, so customer satisfaction is what every bank strives for.

Since interacting with financial institutions is such a fundamental part of everyone’s lives it is important to do further research and analysis into what customers struggle with to identify areas of opportunity for banks – not only for their product development but also for overall customer service and interaction. There is plentiful literature available on the importance of customer satisfaction and the impact on business if it is low. Each financial institution might define success has something different, but it is at the heart of a bank to serve a customer for all their financial needs so in our analysis and research we will look to provide insight into areas that they can focus on improving to hopefully reach new customers or encourage existing customers with other products to expand their services.

II. DATASET

The first dataset was sourced from DataWorld [3]. The dataset consists of fifteen thousand records. Information that is provided includes: date complaint was received, product and sub-product, issue with product and complaint narrative, company response, company name, state and zip code of complaint origination, complaint submission method, time response, and whether the consumer disputed or not. The data fields that will be used for analysis will be product and sub-product, issue with the product, company response, company name, state of complaint origination, complaint submis-

sion method, and consumer disputed field. All records that are not complete with these fields will be excluded from evaluation. Complaint narratives will be used for a more detailed description of the issue submitted. The dataset spans the years 2011 to 2019 allowing us to do a year-over-year comparison of complaint volumes in certain products and states.

The second dataset was sourced from the Consumer Financial Protection Bureau and has over two million records to analyze for this project [4]. Fields from the first dataset can be mapped to the fields in the second dataset thus creating a larger sample to run analysis on. Things to note about the CFPB dataset are that complaints in this set are only published after the company responds and the relationship with the consumer is confirmed [4]. Data points collected and that appear in both data sets are: company name, company’s response to the consumer, whether the response was timely (a yes or no response), date and state received, the product the complaint is against, and the issue, some complaints have additional details provided in the form of sub-issue [4]. A full list of data provided can be seen in Table 1.

Not all datapoints in the dataset were used during this research project. The most important datapoints and those that were used in the analysis were: company, product, sub-product, issue, sub-issue, complaint ID, and state. Timely response was considered as an additional aspect for correlation but responses were only limited to ‘yes’ and ‘no’ which do not provide enough to draw out any conclusions. Additional details for timely response would have required criteria that would define what timely is for each bank.

A. Data Cleaning, Processing and Input Tools

We have processed our data in order to find if there are any duplicate or null values present in our dataset. While processing the data, we found that there were no duplicate values but there were two null values that were removed in order to visualize the data in Tableau.

For initial data analysis we used Tableau (Table 2 for initial code) to create some visualizations for a subset of the data set. Moving forward we plan to make use of Python Jupyter Notebook and some of their libraries like Seaborn and Pandas for achiev-

TABLE I: Complete List of Complaint Data Consumer Financial Protection Bureau publishes. [1: direct pull of descriptions from CFPB]

Field Name	Description
Date received	The date the CFPB received the complaint. For example, “05/25/2013.”
Product	The type of product the consumer identified in the complaint. For example, “Checking or savings account” or “Student loan.”
Sub-product	The type of sub-product the consumer identified in the complaint. For example, “Checking account” or “Private student loan.”
Issue	The issue the consumer identified in the complaint. For example, “Managing an account” or “Struggling to repay your loan.”
Sub-issue	The sub-issue the consumer identified in the complaint. For example, “Deposits and withdrawals” or “Problem lowering your monthly payments.”
Consumer complaint narrative	Consumer complaint narrative is the consumer-submitted description of “what happened” from the complaint. Consumers must opt-in to share their narratives. We will not publish the narrative unless the consumer consents, and consumers can opt out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer.
Company public response	The company’s optional, public-facing response to a consumer’s complaint. Companies can choose to select a response from a pre-set list of options that will be posted on the public database. For example, “Company believes complaint is the result of an isolated error.”
Company	The complaint is about this company. For example, “ABC Bank.”
State	The state of the mailing address provided by the consumer.
ZIP code	The mailing ZIP code provided by the consumer. This field may: i) include the first five digits of a ZIP code; ii) include the first three digits of a ZIP code (if the consumer consented to publication of their complaint narrative); or iii) be blank (if ZIP codes have been submitted with non-numeric values, if there are less than 20,000 people in a given ZIP code, or if the complaint has an address outside of the United States).
Tags	Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers. For example, complaints where the submitter reports the age of the consumer as 62 years or older are tagged “Older American.” Complaints submitted by or on behalf of a service member or the spouse or dependent of a servicemember is tagged “Servicemember.” Servicemember includes anyone who is active duty, National Guard, or Reservist, as well as anyone who previously served and is a veteran or retiree.
Consumer consent provided?	Identifies whether the consumer opted in to publish their complaint narrative. We do not publish the narrative unless the consumer consents, and consumers can opt-out at any time.
Submitted via	How the complaint was submitted to the CFPB. For example, “Web” or “Phone.”
Date sent to company	The date the CFPB sent the complaint to the company.
Company response to consumer	This is how the company responded. For example, “Closed with explanation.”
Timely response?	Whether the company gave a timely response. For example, “Yes” or “No.”
Consumer disputed?	Whether the consumer disputed the company’s response.
Complaint ID	The unique identification number for a complaint.

identifying the areas of issues and will help with predicting product and issues using the historical complaint data. From this, we will be able to provide plotting graphs, scatter plots, histograms, and correlation matrix.

IV. REASON FOR BIG DATA SOLUTION

A big data solution will allow for greater detection about customer sentiment regarding a company's products or services. Using Machine Learning will allow for a more robust way to process information [27]. The machine power can process the dataset in order to calculate various types of variables from the population [27] at a faster and more efficient way. In addition to overall sentiment, we will be able to look at multiple facets of the complaints received, such as, channel complaint was received through, the company that received the complaint, the state from which the complaint originated from, and whether or not it was resolved in a timely manner. Similarly, the Consumer Financial Protection Bureau uses the large dataset of complaints to analyze it and help govern and guide companies with more informed financial laws, rules and regulations [4].

V. PUBLISHED RESEARCH

To understand how a company can be successful one needs to begin by understanding what complaints most are received. Merriam-Webster defines complaint as “[an] expression of grief, pain, or dissatisfaction”. “Customer experience is the emotion felt by customers¹ when they come into any contact with a company – no matter how or by what means. It is what customers remember from their interaction with a business. [28]” There are a number of studies that have been done that show that customers that have a positive experience with a company are 86% more likely to return and do business with that company again and goes up to 92% when that customer is already an existing customer [28]. In addition to those higher percentages, losing customers can be very costly for a business. In another study done, 50% of customers said they left because they did not feel valued and had a poor experience with customer service [28]. Another staggering number is that 80% of customers would willingly pay more for a product or service if the

customer experience is better [28]. If those numbers are not enough to convince a company the value that customer satisfaction provides then it has been analyzed that acquire new customers can cost five to twenty-five times more than retaining an existing customer [28].

A. Customer Satisfaction & Measurement

To effectively measure and predict customer satisfaction one must define how to measure it. An important piece to measuring satisfaction is also understanding quality as it plays a key role in how the customer determines their expectations of the product they are interacting with [2]. Quality is difficult to define in the realm of customer satisfaction for the simple fact that it can vary person-to-person – thus making it more difficult to quantify [2]. One way to gather measurement for customer satisfaction is through the use of surveys [29]. Surveys give companies the insights directly from their customers. “Strong relationships lead to higher levels of loyalty with customers [thus] resulting in profitability [29]”. Service quality, when looking at the “Relationship Survey Framework” (Figure ??), depends on the following customers experiences: installation, complaint, billing, purchase, pre-sales, and any other ones they interacted with [29]. When all those experiences are positive then it builds or improves upon the relationship resulting in an overall better product experience, price, and increase in corporate reputation [29].

B. Customer Complaint Handling on Customer Satisfaction

Another study was done on banks in Asia on the “effect of bank commitment, bank communication, and handling customer complaint on customer loyalty through customer satisfaction” was completed by obtaining a sample from customers [30]. “The commitment of a service provider greatly impacts on customer loyalty. In this case, when the service provider is highly committed to providing services to its customers, it will make customers satisfied and then will not switch to other competitors and continue to use the service or product” [30]. Similarly, as seen by the top three banks identified in this research project, the banks are heavily focused on providing exceptional service to their customers.

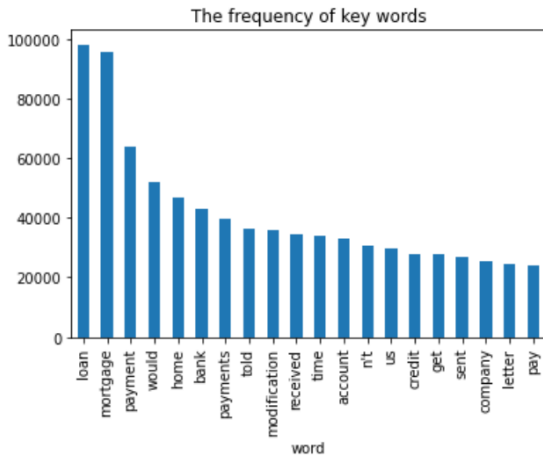


Fig. 3: Frequency of Key Words

The study found that commitment, communication and the way incidents are handled all have a great impact on one another [30].

A third aspect of the study done focused on the bank’s communication and its impact on customer satisfaction. The research data showed that bank communication also had a positive and significant effect on customer satisfaction [30]. “Communication is very important in a business.

From the two million data records natural language processions identified the top complaints that companies received: top three being around loans, mortgages and payments(Figure 3). To identify areas of opportunities analysis of the top three financial institutions will be compared to the three financial institutions that have the least number of complaints against them. In addition to the bar chart in Figure 3 that shows the frequency of key words, the word cloud in Figure ?? clearly illustrates which words are most likely used in the complaints data.

The three companies that received the most complaints in our dataset against them were Equifax, Experian, and TransUnion regarding credit reporting. Since these three companies are not financial institutions and only serve as credit reporting bureaus they will be excluded as part of this analysis.

C. Bank of America

Bank of America first came into the market in 1923 but was established as the leading bank with the merger between BankAmerica and NationsBank

in 1998 [31]. Bank of America offers a large variety of products to their consumers, to include personal banking such as checking and savings accounts, home and personal loans, credit cards, small business accounts, wealth management activities, and products and support to larger businesses and institutions [32]. In 2019 Bank of America received the J.D. Power U.S. Retail Banking Advice Study highest ranking for customer satisfaction [33]. “The study measures satisfaction across twenty-three of the largest banks across the U.S.” [33].

D. Wells Fargo

Wells Fargo was first founded in 1852 but similar to Bank of America current day Wells Fargo was established in 1998 with the merger of Wells Fargo & Company and Norwest Corporation [34]. Wells Fargo customers have many products to choose from including banking and credit cards, loans, investing and retirement support, wealth management, services to small businesses, and commercial products as well [35]. Wells Fargo is committed to transforming their business and practices, with a focus on customer experience and customer-focused innovation [36]. One of Wells Fargo’s goals is customer service and advice and like Bank of America, Wells Fargo ranked third in customer satisfaction in the J.D. Power 2019 U.S. Retail Banking Advice Study [36].

E. Comparison of Top Three Companies with Highest Complaints

Initial word processing on the dataset, we were able to compare Bank of America, Wells Fargo, and JPMorgan Chase & Co. on the volume of complaints they received. As seen in Figure 4 the most seen words in customer’s complaints for the three companies include loan, mortgage, and payments. Figure 5 further breaks the words most used in the complaints across the companies and looks to show how frequently it was used.

VI. HOW ANALYSIS WILL HELP CUSTOMER SATISFACTION

Understanding where customers complain the most will helps guide financial institutions towards improvements that will hopefully alleviate the pain points. A Bain & Company brief discusses the

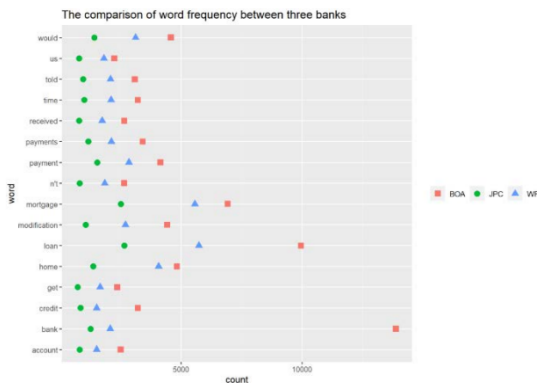


Fig. 4: Comparison of Word Frequency between Bank of America, Wells Fargo, and JPMorgan Chase & Co.

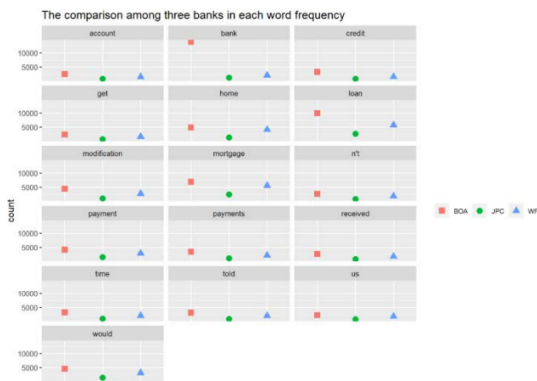


Fig. 5: Comparison of Word Frequency between Bank of America, Wells Fargo, and JPMorgan Chase & Co.

importance of analytics in deepening customer relationships in banking [37]. They outline that analytics “enable banks to activate, not just retain, their high-value customers” and harnessing the full power of analytics works best with five principles: “segmenting customers by value, automating forecasts, predicting loyalty, understanding what causes churn, and taking a test-and-learn approach” [37]. Our focus on complaints will help to inform why a customer might leave thus understanding customers’ struggles. Our analysis of complaint data will lead to predicting behavior of customers on certain products and issues. Bain & Company discusses how a predictive model using a data set with known drivers of Net Promoter Score (NPS) [37]. “A predictive model starts with basic features

present in ordinary segmentation, such as channel usage, the frequency and nature of sales and service interactions, product usage, and revenue. It then taps more advanced sources of data, such as natural language processing of contact center conversations, including volume and tone [37].”

As seen in Figure ??, Bain & Company did a study on the impact of predictive NPS model and “using a predictive model achieves 70% predictive accuracy and a 30% success rate on commercial campaigns, compared with a roughly 5% success rate for the average campaign” [37]. Using the support of these successful number we will look to predict possible outcomes for Bank of America, Wells Fargo, and JPMorgan Chase & Co.

A. Complaints Handling on Bank Brand

In a study on “the impact of the magnitude of service failure and complaint handling on satisfaction and brand credibility in the banking industry” there is a correlation between customer satisfaction and how complaints are handled [38]. The study also points out the variability in what is considered “effective” handling; the fact that humans are involved introduces the variable that cannot be controlled [38]. An individual, despite what the bank may consider timely, can have differing opinions on how a complaint should be handled and what timeframe it should be handled in [38]. “Davidow (2003) concluded that there are six aspects of responsiveness: timeliness, redress, apology, credibility, attentiveness, and facilitation. Among these factors, timeliness is a controllable element that customers consider and judge firms regarding it since failure occurs” [38].

VII. THE RESEARCH

The research was conducted using machine learning and there were several graphs and matrixes produced. In the following sub-sections graphs generated will be discussed and a code snippet used to generate the graphs is shown as well.

The histogram (code in Figure ??) is an acquainted graphical presentation for addressing the frequency of a batch of data. The scope of the information is divided into intervals and the number of qualities falling into each interval is counted. The histogram then, at that point, comprises of

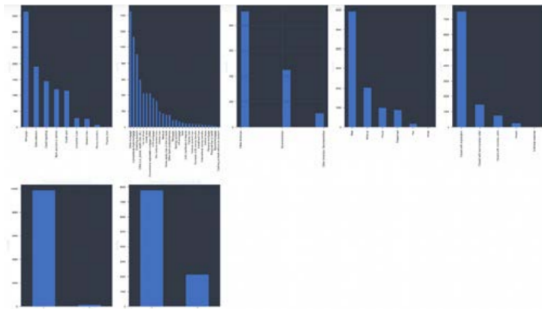


Fig. 6: Distribution graphs

a progression of square shapes whose widths are characterized by as far as possible inferred by the bin widths, and whose statures rely upon the number of qualities in each bin.

In the above graphs (Figure 6), we are identifying the distribution graph for columns from the dataset which has no NAN values. The histograms are displayed with the frequency of counts in y-axis while x-axis contains the values of the column picked.

A. Correlation Graph & Matrix

The correlation graph (code in Figure ??) is a $(K \times K)$ square and even framework whose ij passage is the connection between the sections I and j of X . Huge data in this graph show genuine collinearity between the factors involved (Figure 7). In any case, the nonexistence of outrageous relationships doesn't suggest the absence of collinearity. The regressor factors for numerous relapses can be profoundly multicollinear even though no pairwise connections are huge. In the above dataset (Figure 7), we only have complaint-id as numerical with int64 data type. All other columns are objects. So, the graph is linear showing for only complaint-id.

B. Scatter Plot/Density Plot

A scatterplot (code in Figure 8) is quite possibly the most impressive yet basic visual plot accessible (Figure 9). In a scatterplot, the data points focused are set apart in Cartesian space with qualities of the dataset lined up with the directions. The credits are as a rule of nonstop information type. One of the critical perceptions that can be closed from a scatterplot is a connection between two ascribes under



Fig. 7: Correlation Matrix on Consumer Complaints dataset on Complaint ID

```
# Scatter and density plots
def plotScatterMatrix(df, plotSize, textSize):
    df = df.select_dtypes(include=[np.number]) # keep only numerical columns
    # Remove rows and columns that would lead to df being singular
    df = df.dropna('columns')
    df = df[[col for col in df if df[col].nunique() > 1]] # keep columns where there are more than 1 unique values
    columnNames = list(df)
    if len(columnNames) > 10: # reduce the number of columns for matrix inversion of kernel density plots
        columnNames = columnNames[:10]
    df = df[columnNames]
    ax = pdy.plotting.scatter_matrix(df, alpha=0.75, figsize=(plotSize, plotSize), diagonal='kde')
    corrs = df.corr().values
    for i, j in zip(*plt.subplots_indices_from(ax, k = 1)):
        ax[i, j].annotate('Corr. coef = %.3f' % corrs[i, j], (0.8, 0.2), xycoords='axes fraction', ha='center', va='center', size=textSize)
    plt.tight_layout('Scatter and Density Plot')
    plt.show()
```

Fig. 8: Snippet of Code for creating Scatter and Density Plot

request. Assuming the qualities are directly connected, then, at that point, the information focuses adjust more like a fanciful straight line; in case they are not associated, the information focuses are dispersed. Aside from essential relationships, scatterplots can likewise demonstrate examples or gatherings of groups in the information and distinguish exceptions in the data.

VIII. CHALLENGES & OPPORTUNITIES

Getting access to data was straightforward and it was plentiful in quantity once downloaded. Access to customer complaints is made easy through the collection process of the Consumer Financial Protection Bureau. To really calculate customer satisfaction a deeper understanding of the resolution would have been needed, such as the definition of what timely response means and whether the customer actually received a resolution to their

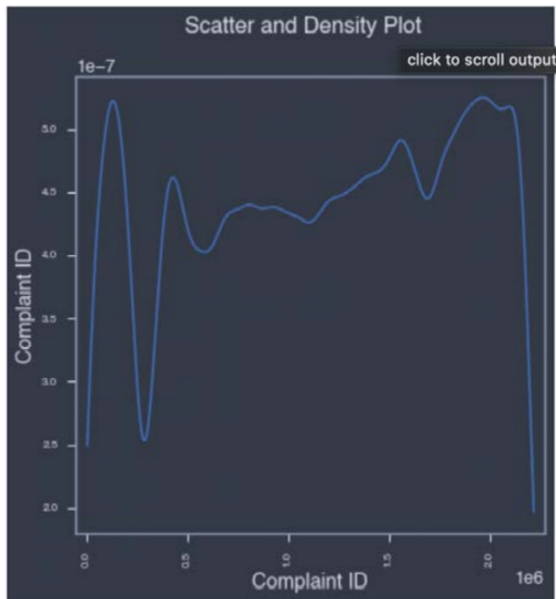


Fig. 9: Scatter and Density Plot for Complaint ID

complaint that did not require any further follow-up. This data would be gathered easily via a survey to customers. Companies use a CSAT, a customer satisfaction score, as a performance indicator for how their service is doing for certain products [39]. CSAT is measured through customer feedback and is done so by allowing customers a scale of one to five, ranging from Very Unsatisfied to Very Satisfied [39]. Once responses are collected, they are averaged out to provide a composite customer satisfaction score [39].

Another challenge to the research was having no demographic attributes to run analysis one in correlation with complaints. As seen in previous studies done on Indonesian banks, they were able to further extract insights based on age, gender, and education level. Such information could prove to be valuable to companies in identifying the root cause of the issue and adjusting their marketing towards those demographics that need it based on the data showed.

To further develop this research topic, it would recommend collecting more data on the customers having the problems. In addition to more data on the customers themselves, expanding the banks in this research project would be helpful in under-

standing if these issues are only limited to one company or whether there is an overarching problem with the industry.

REFERENCES

- [1] K. States. (September 2, 2011) “three businesses that failed for lack of customer service”. Accessed December 9, 2021. [Online]. Available: https://www.insidetucsonbusiness.com/news/on_guard/three-businesses-that-failed-for-lack-of-customer-service/article_42816a5e-d4df-11e0-9b64-001cc4c002e0.html
- [2] R. F. Gerson, *Measuring customer satisfaction*. Menlo Park, Calif: Crisp Publications, 1993.
- [3] Data World. “there are 802 open data datasets available on data.world.”. Accessed December 9, 2021. [Online]. Available: <https://data.world/datasets/open-data>
- [4] Consumer Financial Protection Bureau. “consumer complaint database.”. Accessed December 9, 2021. [Online]. Available: <https://www.consumerfinance.gov/data-research/consumer-complaints/>
- [5] L. Cui and D. Lee, “Coaid: COVID-19 healthcare misinformation dataset,” *CoRR*, vol. abs/2006.00885, 2020. [Online]. Available: <https://arxiv.org/abs/2006.00885>
- [6] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, “The science of fake news,” vol. 359, no. 6380, pp. 1094–1096, Mar. 2018. [Online]. Available: <https://doi.org/10.1126/science.aao2998>
- [7] Q. Su, M. Wan, X. Liu, and C.-R. Huang, “Motivations, methods and metrics of misinformation detection: An NLP perspective,” vol. 1, no. 1-2, p. 1, 2020. [Online]. Available: <https://doi.org/10.2991/nlpr.d.200522.001>
- [8] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, 2017*, pp. 963–972. [Online]. Available: <https://doi.org/10.1145/3041021.3055135>
- [9] C. Yang, R. C. Harkreader, and G. Gu, “Empirical evaluation and new design for fighting evolving twitter spammers,” *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013. [Online]. Available: <https://doi.org/10.1109/TIFS.2013.2267732>
- [10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
- [11] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, “Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection,” *Nature Medicine*, vol. 27, no. 7, pp. 1205–1211, May 2021. [Online]. Available: <https://doi.org/10.1038/s41591-021-01377-8>

- [12] J. Havey. "pharma research progress hope,". [Online]. Available: https://catalyst.phrma.org/a-year-and-a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm_campaign=2021-q3-cov-innutm_medium=pai_srh_cpc-ggl-adjutm_source=gglutm_content=clk-pol-tpv_scl-geo_std-usa-dca-pai_srh_cpc-ggl- [26]
- [13] N. Sallahi, H. Park, F. E. Mellouhi, M. Rachdi, I. Ouassou, S. Belhaouari, A. Arredouani, and H. Bensmail, "Using unstated cases to correct for COVID-19 pandemic outbreak and its impact on easing the intervention for qatar," *Biology*, vol. 10, no. 6, p. 463, May 2021. [Online]. Available: <https://doi.org/10.3390/biology10060463> [28]
- [14] M. El-Harbawi, B. B. Samir, M.-R. Babaa, and M. I. A. Mutalib, "A new QSPR model for predicting the densities of ionic liquids," *Arabian Journal for Science and Engineering*, vol. 39, no. 9, pp. 6767–6775, Jun. 2014. [Online]. Available: <https://doi.org/10.1007/s13369-014-1223-3> [29]
- [15] P. R. Krause, T. R. Fleming, R. Peto, I. M. Longini, J. P. Figueroa, J. A. C. Sterne, A. Cravioto, H. Rees, J. P. T. Higgins, I. Boutron, H. Pan, M. F. Gruber, N. Arora, F. Kazi, R. Gaspar, S. Swaminathan, M. J. Ryan, and A.-M. Henao-Restrepo, "Considerations in boosting COVID-19 vaccine immune responses," *The Lancet*, vol. 398, no. 10308, pp. 1377–1380, Oct. 2021. [Online]. Available: [https://doi.org/10.1016/s0140-6736\(21\)02046-8](https://doi.org/10.1016/s0140-6736(21)02046-8) [30]
- [16] J. H. Kim, F. Marks, and J. D. Clemens, "Looking beyond COVID-19 vaccine phase 3 trials," *Nature Medicine*, vol. 27, no. 2, pp. 205–211, Jan. 2021. [Online]. Available: <https://doi.org/10.1038/s41591-021-01230-y> [31]
- [17] E. C. Fernández and L. Y. Zhu, "Racing to immunity: Journey to a COVID-19 vaccine and lessons for the future," *British Journal of Clinical Pharmacology*, vol. 87, no. 9, pp. 3408–3424, Jan. 2021. [Online]. Available: <https://doi.org/10.1111/bcp.14686> [32]
- [18] Code Academy. "natural language processing/text preprocessing". [Online]. Available: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-text-preprocessing/cheatsheet> [33]
- [19] S. Akon and A. Bhuiyan, "Covid-19: Rumors and youth vulnerabilities in bangladesh," 07 2020. [34]
- [20] "suicide statistics". [Accessed December 9, 2021]. [Online]. Available: <https://www.befrienders.org/suicide-statistics> [35]
- [21] T. V. Green and C. Doherty. "majority of u.s. public favors afghanistan troop withdrawal; biden criticized for his handling of situation". [Online]. Available: <https://www.pewresearch.org/fact-tank/2021/08/31/majority-of-u-s-public-favors-afghanistan-troop-withdrawal-biden-criticized-for-his-handling-of-situation/> [36]
- [22] M. U. Islam, F. B. Ashraf, A. I. Abir, and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," in *2017 20th International Conference of Computer and Information Technology (ICCIIT)*. IEEE, Dec. 2017. [Online]. Available: <https://doi.org/10.1109/iccitechn.2017.8281777> [37]
- [23] E. Kiely and R. Farley. "timeline of u.s. withdrawal from afghanistan". [Online]. Available: <https://www.factcheck.org/2021/08/timeline-of-u-s-withdrawal-from-afghanistan/> [38]
- [24] Real Python. "tokenization in spacy". [Online]. Available: <https://realpython.com/natural-language-processing-spacy-python/#tokenization-in-spacy> [39]
- [25] D. Smeltz and E. Sullivan. (August 9, 2021) "us public supports withdrawal from afghanistan". [Online]. Available: <https://www.thechicagocouncil.org/commentary-and-analysis/blogs/us-public-supports-withdrawal-afghanistan> [40]
- spaCy. "spacy 101: Everything you need to know". [Online]. Available: <https://spacy.io/usage/spacy-101> [41]
- D. Deshpande and M. Kumar, *Artificial Intelligence for Big Data - Complete guide to automating Big Data solutions using Artificial Intelligence techniques*. Packt Publishing, Limited, 2018.
- D. Lafrenière, *Delivering fantastic customer experience: how to turn customer satisfaction into customer relationships. 1st edition*. New York: Productivity Press., 2019.
- A. Rao and S. Chandra, *The Little Book of Big Customer Satisfaction Measurement*. SAGE Publications India, 2013.
- R. S. R. Widiyanto and B. Rachmat, "Effect of bank commitment, bank communication and handling customer complaint on customer loyalty through customer satisfaction at PT bank central asia tbk of mojobahit mojoberto sub-branch office," *International Journal of Multicultural and Multireligious Understanding*, vol. 6, no. 3, p. 49, Jun. 2019. [Online]. Available: <https://doi.org/10.18415/ijmmu.v6i2.756> [42]
- Wikipedia. "bank of america". Accessed December 9, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Bank_of_America [43]
- "about bank of america - our people, our passion, our purpose.". Accessed December 9, 2021. [Online]. Available: <https://about.bankofamerica.com/en/newsroom>. (February 4, 2019) "bank of america tops j.d. power ranking for retail banking advice.". Accessed December 9, 2021. [Online]. Available: <https://newsroom.bankofamerica.com/press-releases/awards-and-recognition/bank-america-tops-jd-power-ranking-retail-banking-advice> [44]
- Wikipedia. "wells fargo.". Accessed December 9, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Wells_Fargo [45]
- "wells fargo personal."wells fargo – banking, credit cards, loans, mortgages & more. Accessed December 9, 2021. [Online]. Available: <https://www.wellsfargo.com/> [46]
- (2016) "wells fargo's transformation". Accessed December 9, 2021. [Online]. Available: <https://www08.wellsfargomedia.com/assets/pdf/commitment/progress-report.pdf> [47]
- P. Baecker, M. Conde, D. Darnell, S. Narayanan, and M. Bergmann. (July 1, 2021) "how analytics can deepen banks' customer relationships". Accessed December 9, 2021. [Online]. Available: <https://www.bain.com/insights/how-analytics-can-deepen-banks-customer-relationships/> [48]
- G. Shams, M. A. Rehman, S. Samad, and R. A. Rather, "The impact of the magnitude of service failure and complaint handling on satisfaction and brand credibility in the banking industry," *Journal of Financial Services Marketing*, vol. 25, no. 1-2, pp. 25–34, Mar. 2020. [Online]. Available: <https://doi.org/10.1057/s41264-020-00070-0> [49]
- "what is csat and how do you measure it?". Accessed November 22, 2021. [Online]. Available: <https://www.qualtrics.com/experience-management/customer/what-is-csat/> [50]

Machine learning Models for Mental Health Analysis based on Religious Impact

Waseem Ashraf
washraf2@gmu.edu

Krishna Sri Dontha
kdontha@gmu.edu

Tarun Kumar Kancheti
tkanchet@gmu.edu

Abstract—Once a taboo topic, recently we have seen a greater awareness of suicide and the factors that influence the suicide rate. There are many factors that influence the suicide rate, among these factors are religious affiliation and religious diversity. Most of the research on religion's influence on the suicide rate has relied on the study of published articles retrieved from a plurality of databases and surveys performed on the selected population. In addition, hardly any research addresses the impact of religious diversity on the suicide rate. The present paper proposes studying the impact of religion on suicide using a quantitative approach. A data set containing the suicide rate, and religious affiliation rate of over 150 countries is constructed from 1990 to 2010. The countries that constituted a population with a single religion over a threshold percentage are identified as countries lacking religious diversity. The analysis indicates that different religions impact suicide differently. A baseline of suicide rate was generated using countries that are mostly affiliated with no religions. The preliminary research was limited to the top four religions of the world. Our research revealed that countries that are mostly affiliated with Christianity, Hinduism, and Buddhism had lower suicide rate compared to countries with no religious affiliation at all. Even the countries that are religiously diverse shield against suicide compared to countries that are not affiliated with any religion.

Index Terms—Machine learning, Mental Health, NLP, Online Information

I. INTRODUCTION

Every death caused by any fact or is distress for the family. The experience is even more painful when the cause of death is self-inflicted. The World Health Organization (WHO) estimates that each year approximately one million people die from suicide, which represents a global mortality rate of 16 people per 100,000 or one death every 40 seconds. It is predicted that by 2020 the rate of death will increase to one every 20 seconds [?]. Based

on different studies and statistics the suicide has been a major issue in the past, and the situation is only getting worse over the years. A lot of research has been performed on different factors that impact the suicide rate such as age, socio economics, sex, culture, region, race, mental health etc... The study off actors that impacts the suicide are complex as suicide rate could be impacted by factors ranging from parenthood to the environment; however, understanding each of the factors that could influence suicide is essential for suicide preventions. One of the factors that could potentially impact the suicide rate is religion. A handful of studies over the past ten years have examined the relationship between religion and suicide risk. Most of these have not found an association, while two found more suicide ideation among persons who gave low importance to religion [2]. Majority of these studies are based on literature reviews, and surveys performed with selected questionnaires. There is hardly any research available that takes a quantitative approach to study the impact of religious affiliation on the suicide rate. In addition, world is becoming religiously diverse, and there have not been any study performed that determines the influence of religious diversity on the suicide rate. Thus, there is a need to investigate impact of religious affiliation, and religious diversity on the suicide rate. The analysis from these study aids us to make data-driven decisions to reduce suicide rates. These decisions include (i) restrictive access to common means of suicide such as (pesticides, firearms) (ii) public awareness of common suicide signs and available help (iii) social support within communities (iv) improving living conditions (v) providing essentials like food (vi) public transportation, education, and jobs (vii) regularly implementing Mental Health

Screenings by health organizations.

II. LITERATURE

Ying and Tyler studied first on religion, suicide ideation, and suicide attempts and second investigated on religion and completed suicide. From their first analysis, they found that spirituality may reduce the risk of suicide ideation through some evidence. They also found that communal forms of religious participation such as service attendance may reduce the risk of suicidal behaviors. In their second part of the analysis, they found religious affiliations may be associated with a lower risk of suicide, but the associations vary across religious denominations and social contexts. They found stronger evidence that proves that frequent religious service attendance may reduce the risk of death by suicide. They suggest there is a need to understand the association between religion and suicide in the broader context of secularization and shift in the age of distribution of suicide from older to younger groups in many modern societies [3].

Danah and Andrew found predominant religion is significantly associated with the sex suicide ratio. They performed analysis on cultural factors such as geography, religion and life expectancy, PPP, education total fertility rate, and gender development index on male and female ratio. They also found Christianity is one of the major religions for suicide rates due to colonial history in countries that may lead to a variety of societal influences in addition to religion. Another religion was Hinduism which ranked third highest suicide mortality ratio. Additionally, they found that as societies become richer and more educated, males have a higher risk of dying as a consequence of suicide relative to females [4] and it shows in the analysis of online comments. [?], [1]–[4]

Philippe's analysis revealed religion was important for many of the patient suffering from psychotic or other psychiatric disorders [7]. emotion and sentiment analysis [5]–[10] of the online comments could help us to detect mental health patients. For this work, we used machine learning and NLP models for analytic of online texts based on the NLP models developed by our research mentors at George Mason University and other

researchers to detect mental health based on text analysis on online information. [11]–[23]

Agnus reveals income level of the region showed the strongest negative association with the suicide rate, followed by heavy drinking, the population aged 65 and had positive associations with the suicide rate [8].

Lawrence et al. examined the relationship between suicide attempts and ideation. Three hundred and twenty-one adult inpatients and outpatients with current diagnoses of Major Depressive Disorder or Bipolar Disorder were recruited for the study. Based on the diagnostic interview results, the study found past suicide attempts were more common among depressed patients with a religious affiliation. Additionally, suicide ideation was more severe among depressed patients who said religion is more important, and among those who Brito et al. examined the association between religious beliefs and observance and the prevalence of psychiatric disorders, psychotic symptoms and history of suicide attempts in the French general population. The study involved interviewing 38,694 subjects at 47 study sites. A positive association was found with psychotic disorders. Conversely, a negative association was found between religiosity and history of suicide attempts [9].

Religiosity has been shown to be associated with lower levels of aggression and hostility, drug use, and risky sexual activity, which are related to suicidal behavior. In a prospective study with depressed youth, for a more complex understanding of the relationship between depression and religiosity, finding not all religious beliefs and experiences corresponding with better mental health. Research suggests that only certain aspects of religiosity/spirituality (e.g., importance of religion, sense of connectedness) might be associated with suicidal behavior. Research regarding religion/spirituality in youth, requires more exploratory studies to further this understanding [6].

Gearing et al. discusses the impact of religion on the suicide rate [5]. The study states “an accurate understanding of a client's religious faith and participation may indicate potential suicide risk. In addition, assessing religiosity may also identify potential areas that treatment could target and enhance life affirming beliefs and expectations.” The

methodology used in this study involved searching the databases for peer reviewed articles on religion and suicide between 2008 and 2017. The study concluded that the influence of religiosity on suicide appears to vary by gender. Similarly, the research also identified that individuals with lower religious orientation are at increased risk of suicide compared to individuals with higher religious orientation. This study was restricted to males. In addition, the study identifies that research is needed to better understand the unique protective role that religions exert against suicide, specifically as it may be moderated by demographic characteristics (e.g., gender, age, and belonging to specific subgroups) and degree of religiosity (e.g., participation in religious activities).

The above papers researched on suicide rate influenced by spiritual beliefs, religious rituals, psychotic disorders, heavy drinking, tobacco, alcohol, geographical area, life expectancy, PPP, education, fertility rate, and gender development. However we use applications of machine learning in natural language processing [17], [24]–[40] in other domains such as health, security and business, and apply transfer learning models to analyse the online comments [41]–[55]. There are only a few papers that have concentrated on religion and suicide rate relation. Additionally, they conducted research based on other paper analyses. This paper aims to analyze the suicide rate based on sex, religion and few other attributes using publicly available data. There are six major religions in the world. These religions are Buddhists, Muslims, Christians, Hinduism, Jews, and folk religions. We would be identifying which religions contributes lowest and highest suicide rate among them. Additionally, we will explore which sex has contributed high suicide rate in top-ranked religion. The above analysis can help to visualize a broader context in society.

A. Problem Description

Between 2008 and 2017, publications on religion and suicide were published in the PsycINFO, MEDLINE, SocINDEX, and CINAHL databases. The influence of religious variety, as well other factors, on the suicide rate was not understood in the studies undertaken during these years. To prevent suicide and reduce the ever-increasing suicide rate,

we must first determine the relationship between the elements that influence suicide. The literature evaluation suggests that religion may have a significant role in suicide prevention; however, this is not the focus of their study. Determining and forecasting the cause of suicide will help us make data-driven decisions in the health curriculum and take proactive preventative efforts through suicide prevention organizations like NSPL, NOFA, AFSP, NIMH, SPTS, and CDC...etc Furthermore, the study will educate the public about the impact of religion on the suicide rate.

B. Problem Investigation

Investigation into the problem set included data collection for the following primaries:

- The factors caused to suicide by the religious believes and its impact when there are diversified religions.
- Is there a link between religion and suicide, the results of this study will be determined by comparing the suicide rates of various faiths to the suicide rates of non-religious persons. The greatest influences of religions on suicide.

III. METHODOLOGY

The methodology was divided into two steps:

- Dataset Preparation
- Analytical Approach

A. Dataset Preparation

We searched trusted sources for data related to suicide rate of different countries over the years, and religious affiliation of different countries over the years. We retrieved the suicide rate data from world health organization, and religious affiliation data from data world. Once datasets were received and their validity verified, the next step was to combine the datasets to build a dataset that could correlate religious affiliation with suicide rate. The dataset was merged using the countries and years as common attribute.

B. Analytical Approach

In order to analyze the impact of religions on the suicide rate, first we needed to determine the countries that are not affiliated with any country as baseline. The non-religious countries were

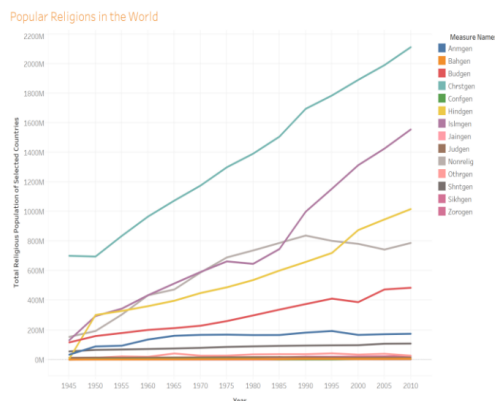


Fig. 1: Most Popular Religions in the World -[Tableau]

determined, and their suicide rate was analyzed. Next, we determined the most popular religions in the world. Once the most popular religions were determined, analysis were performed to determine the countries that belong to each of the popular religions. In order to ensure that influence of single religion is determined, we selected the countries where the predominant population belongs to a single religion, thus lacking religious diversity. For majority of popular religions, threshold of 95% and above was used to ensure that these areas are only influenced by a single religion. In addition, countries with religious diversity were also determined to analyze the influence of religious diversity on suicide. The results of individual religions were compared with each other, and the baseline of non-religious affiliation to determine the religious influence on the suicide rate.

1) *Top Five Religions in the World:* For this study we focused on top five religions in the world. Figure 1 shows the religious population over the years in the world. As can be seen from the above figure, Christianity, Islam, non-religion, Hinduism, and Buddhism makeup top five religions in the world. In addition, top five religion's population have kept increasing over the time. Christianity is still the leading religion, whereas Islam is growing on faster rate than any other religions.

2) *Countries that Lack Religious Diversity:* Next, for each of the top five popular religions in the world, we identified the countries that belong to a single religion (thus lacking religious diversity).

Countries with Dominant Christian Affiliation

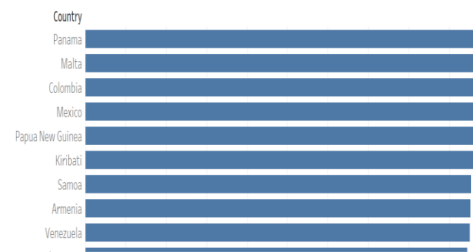


Fig. 2: : Predominant Christian Countries -[Tableau]

The countries that have a single dominant religion were selected based on religious affiliation threshold. For majority of popular religions, we were able to select countries that have at least 90% religious' affiliation.

3) *Predominant Christian Countries:* Figure 2 bar graph shows ranked snippet of the Christian affiliation for each country in the dataset for year 2010. The analysis was performed from 1995-2010 by each five years. The following table shows countries with at least 95% of the population that belongs to Christianity in 2010.

REFERENCES

- [1] J. H. Fetzer, "Disinformation: The use of false information," vol. 14, no. 2, pp. 231–240, May 2004. [Online]. Available: <https://doi.org/10.1023/b:mind.0000021683.28604.5b>
- [2] M. Fernandez and H. Alani, "Online misinformation." ACM Press, 2018. [Online]. Available: <https://doi.org/10.1145/3184558.3188730>
- [3] H. Zhang, A. Kuhnle, J. D. Smith, and M. T. Thai, "Fight under uncertainty: Restraining misinformation and pushing out the truth." IEEE, Aug. 2018. [Online]. Available: <https://doi.org/10.1109/asonam.2018.8508402>
- [4] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, "Addressing health-related misinformation on social media," vol. 320, no. 23, p. 2417, Dec. 2018. [Online]. Available: <https://doi.org/10.1001/jama.2018.16865>
- [5] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIoT)*. IEEE, 2021, pp. 0255–0261.
- [6] L. Cui and D. Lee, "Coaid: COVID-19 healthcare misinformation dataset," *CoRR*, vol. abs/2006.00885, 2020. [Online]. Available: <https://arxiv.org/abs/2006.00885>
- [7] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," vol. 359,

- no. 6380, pp. 1094–1096, Mar. 2018. [Online]. Available: [18] <https://doi.org/10.1126/science.aao2998>
- [8] Q. Su, M. Wan, X. Liu, and C.-R. Huang, “Motivations, methods and metrics of misinformation detection: An NLP perspective,” vol. 1, no. 1-2, p. 1, 2020. [Online]. Available: <https://doi.org/10.2991/nlpr.d.200522.001> [19]
- [9] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April* [20] 3-7, 2017, 2017, pp. 963–972. [Online]. Available: <https://doi.org/10.1145/3041021.3055135>
- [10] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, “A survey on concept-level sentiment analysis techniques of textual data,” in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0285–0291.
- [11] M. Heidari, H. James Jr, and O. Uzuner, “An empirical study of machine learning algorithms for social media bot [21] detection,” in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.
- [12] C. Yang, R. C. Harkreader, and G. Gu, “Empirical [22] evaluation and new design for fighting evolving twitter spammers,” *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013. [Online]. Available: <https://doi.org/10.1109/TIFS.2013.2267732> [23]
- [13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: [24] Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: [25] <https://doi.org/10.18653/v1/n19-1423>
- [14] D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport, “Neutralizing antibody [26] levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection,” *Nature Medicine*, vol. 27, no. 7, pp. 1205–1211, May 2021. [Online]. [27] Available: <https://doi.org/10.1038/s41591-021-01377-8>
- [15] J. Havey. “pharma research progress hope.”. [Online]. Available: [https://catalyst.phrma.org/a-year-and- \[28\] a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm_campaign=2021-q3-cov-innvtm_medium=paishrcpc-ggl-advutm_source=gglutm_content=clk-pol-tpvsl-geostd-usa-dca-paishrcpc-ggl- \[29\]](https://catalyst.phrma.org/a-year-and- [28] a-half-later-the-biopharmaceutical-industry-remains-committed-to-beating-covid-19?utm_campaign=2021-q3-cov-innvtm_medium=paishrcpc-ggl-advutm_source=gglutm_content=clk-pol-tpvsl-geostd-usa-dca-paishrcpc-ggl- [29])
- [16] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, “Ontology creation model based on attention mechanism for a specific business domain,” in *2021 IEEE International IOT, Electronics [30] and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.
- [17] N. Sallahi, H. Park, F. E. Mellouhi, M. Rachdi, I. Ouassou, S. Belhaouari, A. Arredouani, and H. Bensmail, “Using unstated cases to correct for COVID-19 pandemic outbreak [31] and its impact on easing the intervention for qatar,” *Biology*, vol. 10, no. 6, p. 463, May 2021. [Online]. Available: [32] <https://doi.org/10.3390/biology10060463>
- M. El-Harbawi, B. B. Samir, M.-R. Babaa, and M. I. A. Mutalib, “A new QSPR model for predicting the densities of ionic liquids,” *Arabian Journal for Science and Engineering*, vol. 39, no. 9, pp. 6767–6775, Jun. 2014. [Online]. Available: <https://doi.org/10.1007/s13369-014-1223-3>
- M. Heidari and S. Rafatirad, “Semantic convolutional neural network model for safe business investment by using bert,” in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2020, pp. 1–6.
- P. R. Krause, T. R. Fleming, R. Peto, I. M. Longini, J. P. Figueroa, J. A. C. Sterne, A. Cravioto, H. Rees, J. P. T. Higgins, I. Boutron, H. Pan, M. F. Gruber, N. Arora, F. Kazi, R. Gaspar, S. Swaminathan, M. J. Ryan, and A.-M. Henao-Restrepo, “Considerations in boosting COVID-19 vaccine immune responses,” *The Lancet*, vol. 398, no. 10308, pp. 1377–1380, Oct. 2021. [Online]. Available: [https://doi.org/10.1016/s0140-6736\(21\)02046-8](https://doi.org/10.1016/s0140-6736(21)02046-8)
- J. H. Kim, F. Marks, and J. D. Clemens, “Looking beyond COVID-19 vaccine phase 3 trials,” *Nature Medicine*, vol. 27, no. 2, pp. 205–211, Jan. 2021. [Online]. Available: <https://doi.org/10.1038/s41591-021-01230-y>
- E. C. Fernández and L. Y. Zhu, “Racing to immunity: Journey to a COVID-19 vaccine and lessons for the future,” *British Journal of Clinical Pharmacology*, vol. 87, no. 9, pp. 3408–3424, Jan. 2021. [Online]. Available: <https://doi.org/10.1111/bcp.14686>
- M. Heidari, S. Zad, and S. Rafatirad, “Ensemble of supervised and unsupervised learning models to predict a profitable business decision,” in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–6.
- M. Heidari and J. H. Jones, “Using bert to extract topic-independent sentiment features for social media bot detection,” in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2020, pp. 0542–0547.
- Code Academy. “natural language processing/text preprocessing”. [Online]. Available: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-text-preprocessing/cheatsheet>
- DeepAI. “named-entity recognition”. [Online]. Available: <https://deepai.org/machine-learning-glossary-and-terms/named-entity-recognition>
- C. on foreign relations. “the u.s. war in afghanistan”. [Accessed December 9, 2021]. [Online]. Available: <https://www.cfr.org/timeline/us-war-afghanistan>
- P. Hajibabae, M. Malekzadeh, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, “An empirical study of the graphsage and word2vec algorithms for graph multiclass classification,” in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021. “covid-19 third doses — health”. [Accessed December 5, 2021]. [Online]. Available: <https://www.fairfaxcounty.gov/health/novel-coronavirus/vaccine/third-doses>
- S. Zad, M. Heidari, P. Hajibabae, and M. Malekzadeh, “A survey of deep learning methods on semantic similarity and sentence modeling,” in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.
- S. Akon and A. Bhuiyan, “Covid-19: Rumors and youth vulnerabilities in bangladesh,” 07 2020.
- M. Heidari, S. Zad, M. Malekzadeh, P. Hajibabae, S. HekmatiAthar, O. Uzuner, and J. H. J. Jones, “Bert model for fake

- news detection based on social bot activities in the covid-19 pandemic,” in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, [47] IEEE, 2021.
- [33] “suicide statistics”. [Accessed December 9, 2021]. [Online]. Available: <https://www.befrienders.org/suicide-statistics>
- [34] T. V. Green and C. Doherty. “majority of u.s. public favors afghanistan troop withdrawal; biden criticized for his handling of situation”. [Online]. Available: <https://www.pewresearch.org/fact-tank/2021/08/31/majority-of-u-s-public-favors-afghanistan-troop-withdrawal-biden-criticized-for-his-handling-of-situation/>
- [35] M. U. Islam, F. B. Ashraf, A. I. Abir, and M. A. Mottalib, “Polarity detection of online news articles based on sentence structure and dynamic dictionary,” in *2017 20th International Conference of Computer and Information Technology (ICCIIT)*. IEEE, Dec. 2017. [Online]. Available: <https://doi.org/10.1109/iccitech.2017.8281777> [49]
- [36] M. Malekzadeh, P. Hajibabae, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, “Review of graph neural network in text classification,” in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, [50] IEEE, 2021.
- [37] E. Kiely and R. Farley. “timeline of u.s. withdrawal from afghanistan”. [Online]. [51] Available: <https://www.factcheck.org/2021/08/timeline-of-u-s-withdrawal-from-afghanistan/>
- [38] Real Python. “tokenization in spacy”. [Online]. Available: <https://realpython.com/natural-language-processing-spacy-python/#tokenization-in-spacy> [52]
- [39] D. Smeltz and E. Sullivan. (August 9, 2021) “us public supports withdrawal from afghanistan”. [Online]. Available: <https://www.thechicagocouncil.org/commentary-and-analysis/blogs/us-public-supports-withdrawal-afghanistan> [53]
- [40] spaCy. “spacy 101: Everything you need to know”. [Online]. Available: <https://spacy.io/usage/spacy-101>
- [41] Code Academy. “text preprocessing”. [Online]. Available: <https://www.codecademy.com/courses/text-preprocessing/lessons/text-preprocessing/exercises/introduction> [54]
- [42] J. Howard. “could covid-19 vaccine boosters be necessary? here’s what experts are saying”. [Accessed October 2021]. [Online]. Available: <https://www.wesh.com/article/could-covid-19-vaccine-boosters-be-necessary-heres-what-experts-are-saying/36519793> [55]
- [43] A. Ain, “The WHO is right to call a temporary halt to COVID vaccine boosters,” *Nature*, vol. 596, no. 7872, pp. 317–317, Aug. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02219-w>
- [44] E. Callaway, “COVID vaccine boosters: the most important questions,” *Nature*, vol. 596, no. 7871, pp. 178–180, Aug. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02158-6>
- [45] A. Weatheron. “health expert says booster shot could be needed after getting covid-19 vaccine”. [Accessed June 8, 2021]. [Online]. Available: <https://www.13newsnow.com/article/life/booster-shot-may-be-needed-after-covid-19-vaccine/291-49a8966c-3d91-48ad-99a0-02905c5593cc>
- [46] P. Naaber, L. Tserel, K. Kangro, E. Sepp, V. Jürjenson, A. Adamson, L. Haljasmägi, A. P. Rumm, R. Maruste, J. Kärner, J. M. Gerhold, A. Planken, M. Ustav, K. Kisand, and P. Peterson, “Dynamics of antibody response to BNT162b2 vaccine after six months: a longitudinal prospective study,” *The Lancet Regional Health - Europe*, vol. 10, p. 100208, Nov. 2021. [Online]. Available: <https://doi.org/10.1016/j.lanepe.2021.100208>
- M. Heidari, J. H. Jones, and O. Uzuner, “Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter,” in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020, pp. 480–487.
- S. J. Thomas, E. D. Moreira, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. P. Marc, F. P. Polack, C. Zerbini, R. Bailey, K. A. Swanson, X. Xu, S. Roychoudhury, K. Koury, S. Bouguermouh, W. V. Kalina, D. Cooper, R. W. Frenck, L. L. Hammitt, Özlem Türeci, H. Nell, A. Schaefer, S. Ünal, Q. Yang, P. Liberator, D. B. Tresnan, S. Mather, P. R. Dormitzer, U. Şahin, W. C. Gruber, and K. U. Jansen, “Safety and efficacy of the BNT162b2 mRNA covid-19 vaccine through 6 months,” *New England Journal of Medicine*, vol. 385, no. 19, pp. 1761–1773, Nov. 2021. [Online]. Available: <https://doi.org/10.1056/nejmoa2110345>
- E. Dolgin, “COVID vaccine immunity is waning — how much does that matter?” *Nature*, vol. 597, no. 7878, pp. 606–607, Sep. 2021. [Online]. Available: <https://doi.org/10.1038/d41586-021-02532-4>
- “virginia open data portal”. [Accessed November 6, 2021]. [Online]. Available: <https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-Vaccines-DosesAdmini/28k2-x2rj>
- U.S. Food and Drug Administration. “covid-19 vaccines”. [Accessed November 6, 2021]. [Online]. Available: <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines>
- M. Heidari and S. Rafatirad, “Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment,” in *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020, pp. 322–329.
- populationU. “populationU”. [Accessed October 10, 2021]. [Online]. Available: <https://www.populationu.com/us/virginia-population>
- V. D. of Health. “covid-19 vaccine summary – coronavirus”. [Accessed October 10, 2021]. [Online]. Available: <https://www.vdh.virginia.gov/coronavirus/covid-19-in-virginia/covid-19-vaccine-summary/>
- M. S. Berg. “what doctors wish patients knew about covid-19 herd immunity”. [Accessed October 10, 2021]. [Online]. Available: <https://www.ama-assn.org/delivering-care/public-health/what-doctors-wish-patients-knew-about-covid-19-herd-immunity>

A Survey on the Jamming and Spoofing UAV Network Attacks and How Machine learning is an Effective Against them

Faisal Alrefaei

*Electrical Engineering and Computer Science
Embry Riddle Aeronautical University
Daytona Beach, USA
Alrefaei@my.erau.edu*

Abstract—Recently, the unmanned aerial vehicle (UAV) has attracted companies, governments, and organizations to execute valuable missions. The security of UAV networks has become an urgent need to avoid disaster consequences. Jamming and spoofing attacks are the most dangerous attacks used against UAV networks. The jamming attack is a stealthy attack that is hard to detect and leads to unavailability of the service over the wireless network. In some cases, unintentional interference caused by the traffic network is similar to a jamming attack. Therefore, the efficient anomaly detection technique must be capable to distinguish them either as normal or abnormal behavior. The Global Position System (GPS) is an important critical system on which the United States relies in some missions. However, hackers target the Global Position System (GPS) signal to manipulate it to deviate the UAV to their extreme zones. Some detection methods used for GPS spoofing attack that are classified into three categories, such as digital signatures, encryption, and some characteristics of Automatic Dependable Surveillance Broadcast (ADS-B). This article reviews the impact of the jamming and spoofing attack and previous traditional security detection and defense techniques. In addition, it addresses the benefit of deep learning technology to show that deep learning technology is an effective technique when it comes to protecting UAV communication networks.

Index Terms—UAV, cyber attack, jamming attack, deep learning, orthogonal frequency

I. INTRODUCTION

In the past decades, revolution technology in computing and sensor capabilities has emerged as promising solutions to play a key role in the modern era. It is used as small devices that rely on an open network to execute a specific mission in a limited time and at low cost, such as unmanned aerial vehicles (UAVs) or drones. UAVs have been used broadly in rescue tasks, packet delivery, remote sensing, rely communication, etc. The components of the UAV network include the UAV, GPS, the ADS-B receiver, and the Ground Control Station (GCS) as shown in Figure 1 [1]. The UAV is considered an autopilot based on self-flying and control without human intervention or the presence of a pilot on the board. All of these components are completely dependent on ADS-B and GPS for navigation. Therefore, the UAV application has been increasingly used due to its unique characteristics. UAVs are present in some critical

fields to perform special tasks in extreme environments. In addition, it is easy to learn how to fly UAVs and they are easily programmed for specific missions in a short time. Therefore, the UAV has been attracting the military and civilian sectors to perform various missions.

Although the growth use of UAVs and the facilities that can be made, the network UAV system is prone and subject to malicious threats to degrade the performance of the UAV system for example, Jamming and Spoofing attacks. A UAV uses an air-ground (line of sight) signal to transmit the data. This link is exposed to be blocked by (jamming attack) or being compromised by malicious hackers who deliberately violate the integrity of the data transmission (spoofing attack). Therefore, the security of the UAV network is an urgent issue.

Jamming attack is one of the dangerous threat attack techniques used against the UAV's node network to disrupt the communication between UAVs and other legitimate entities. It is an action that violates the policies of the media access control protocol or PHY in wireless communication to disrupt transmissions [2]. It is a form of denial-of-service (DOS) attack that leads to interrupting the ongoing communication and making it unavailable. Hackers execute the jamming attack technique by sending radio signals to flood the channel node. Jamming attacks can be classified into four categories: constant, reactive, random, and deceptive [3]. Each of these attacks is executed on the basis of their knowledge and technique to execute their malicious action correctly.

GPS and ADS-B play a critical role in the UAV's navigation and position. The spoof attack is an attack that aims to deviate the UAV from its planned path to the unsafe zone. Hackers deliberately alter the GPS signal to manipulate the navigation system [4]. In ADS-B the spoofing attack subjects ADS-B of the ground or aircraft-based attacks by manipulating the international Civil Aviation Organization (ICAO) addresses that are positioned in each aircraft as a unique identifier [5]. Hackers spoof the ICAO to show a fake non-existent aircraft to mislead the Air traffic Control ATC.

Scientists and researchers have innovated and published many suggestions and strategies to detect and defend against

jamming and spoofing attacks. These traditional techniques are impractical and ineffective when applied in UAV networks. UAVs are classified as small devices that have limited resources, such as size, power consumption, and high mobility. Therefore, the researchers try to define a suitable technique that is effective taking into account the limited resources.

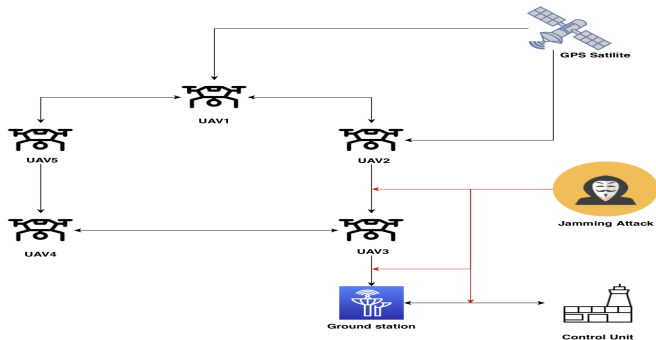


Fig. 1. Jamming Attack on the UAVs

II. JAMMING ATTACK

Jamming attack is a version of the Internet Denial of Services (DOS) attack. Data communication in the UAV network is classified into four types [6]. The first is the data link between the UAV and the GCS. The second is a UAV and another nearby UAV. The third is a UAV and the GPS. The last is the UAV and ADS-B at the ground station. Automatic Dependent Signal Broadcast (ADS-B) is a system that is being deployed on aircrafts to control air traffic. The UAV depends significantly on the ADS-B to easily drive it to its safe home land. The compromising or blocking these system’s signals Attackers use this jamming attack to disrupt UAV communication by sending unwanted radio signals. Hackers aim to distribute ongoing communication to degrade UAV communication performance by broadcasting noise radio signals to flood the channel nodes.

1) *Constant Jamming Attack*: A constant jamming attack is a malicious action that continuously sends a high power noise signal without following any protocol to prevent communication and consume the power of the receiver as shown in Figure 2. This malicious action aims directly at arbitrary achieve tow goals, which create congestion over communication links or lead the receiver to consume its resources by receiving a large number of data packets [8]. An extra number of data packets is a reason to consume power or memory size. This action prevents the legitimate entity from sending signals over the network and causes delays in the time required for the new signal to reach the receiver side channel. Therefore, hackers intentionally continue to target quality signals by sending arbitrary signals [8], disrupting quality signals to make the signal disappear and making it impossible to decode it.

2) *Deceptive Jamming Attack*: A deceptive jamming attack is the most stealthy attack among the four types of jamming attack. its intention constantly sends authentic packets to

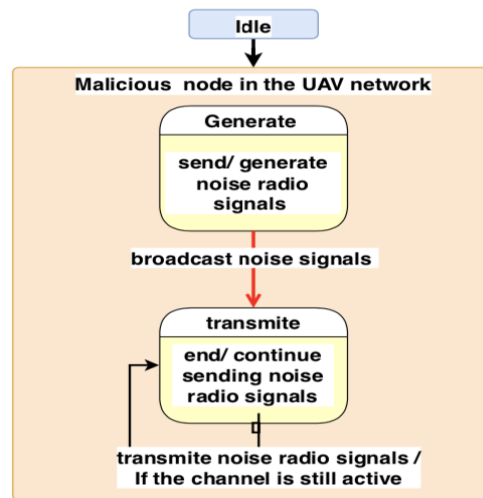


Fig. 2. Constant Jamming attack on the UAVs

mislead the other aircraft [7]. The deceptive jamming attack occurs in the ADS-B as an injection of a message attack, as shown in Figure 3. The adversaries inject empty data packets into the transmission link during the data transmission being active. They exploit the transmission gap that occurs when the sender sends the data packet [8]. The deceptive jamming attack, intending to inject its injection packet as empty useless payload in the receiver channel.

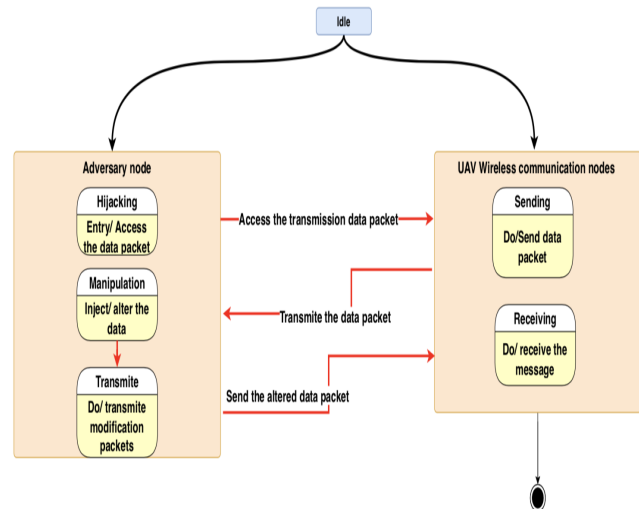


Fig. 3. Deceptive Jamming Attack on UAVs

3) *Random Jamming Attack*: The preservation of resources is the main goal of performing efficient malicious action against targets. The intermittent jamming attack preserves its resources by changing between idle and active modes [9]. It sends noise signals intermittently. This kind of jamming attack leads to not consuming its power resources by continuing to send signals as constant jamming signals, as shown in Figure 4. This attack is called random jammer, where it

is a combination of message injection attack and constant jamming attack [9]. Its technique aims to preserve its power resources.

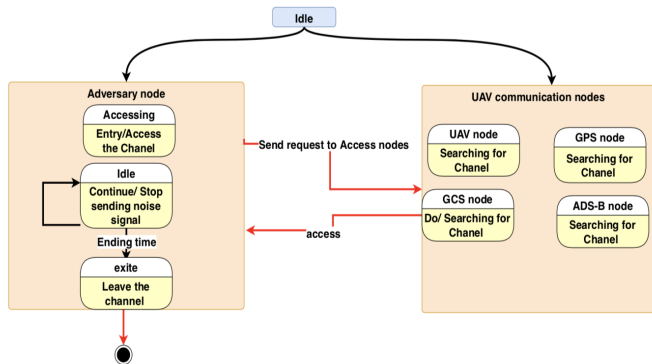


Fig. 4. Random Jamming Attack on the UAVs

4) *Reactive Jamming Attack*: The reactive jamming attack aims to sense the status of the reception channel. This technique needs to sense the activation status of the channel in the nodes of the UAV network to start its action, as shown in Figure 5. Once hackers discover that the channel is activated, they start sending the noise signal to the reception nodes [9]. Another technique used by hackers in this type of attack is that hackers can distinguish sender quality signals. once the hackers distinguish that weakness in the sender signals is achieved, they stop sending the noise signals; otherwise, it continues sending the signal unless those two conditions are not achieved [9].

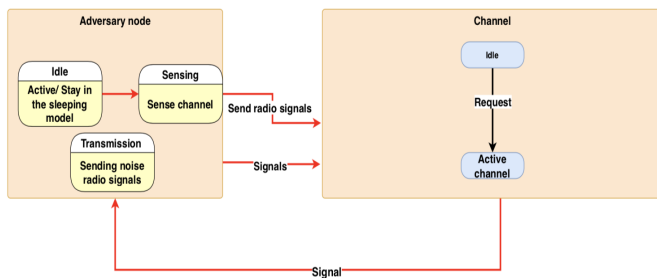


Fig. 5. Reactive Jamming Attack on the UAVs

III. SPOOFING ATTACK

Hacker lanch spoof attack on the two data links and GSC based on the GPS and ADS-B data packet, as shown in Figure 6. The components of the UAV network include the UAV, the navigation satellite system, the ADS-B receiver, and the ground station. All of these components are completely dependent on ADS-B and GPS for navigation. However, natural communication is vulnerable to being compromised by

malicious actions to violate ongoing data packet transmission integrity. The malicious action then is to collect the ADS-B and GPS data to execute their malicious action.

ADS-B is one of the main components of next-generation air transportation systems. This system plays a crucial role in the management of spaces to avoid congestion and collisions. The aircraft is equipped with ADS-B out and ADS-B in [13]. The two channels transmit and receive control to keep the airspace safer than ever before. The content of the information in the transmission packets includes the position, speed, altitude, and other information for another flight in the sky.

GPS plays a key role in the self-flying of the UAV in space. The GPS widely used by UAV recovers from the navigation and positioning system. The United States developed the GPS system to be used for navigation and flying [12]. The GPS signal is divided between two fields of military or civilian missions. Most of the signal is sent as plain text, so anyone can see it or alter it to achieve specific objectives.

GPS spoofing attack is a common technique used by hackers to inject false GPS signals into the UAV. The adversary uses the GPS signal to send it as a fake signal with higher signal power [14]. Hackers use this attack to drive the UAV to an unsafe zone or deviate the UAV from its planned flying path to cause a collision in the space. They can be gradually simulated at different locations by presenting fake positions.

Relying on non-encrypted packets of ADS-B broadcasting and GPS signals attracts attention to the vulnerability that can be exploited for malicious actions. ADS-B spoofing is divided into two categories, aircraft and ground-based [15]. The hacker found that targeting the ADS-B signal leads to misleading the GSC and other nearby UAVs. They always initiate their malicious plan by modifying or altering the data to drive the UAV to dangerous zones.

IV. SPOOFING ATTACK ON THE GPS AND ADS-B

UAV system was fully dependent on and guided by GPS and ADS-B to control its direction or perform its tasks. The UAV uses GPS to perform specific tasks such as landing in the planned zone. Additionally, the aircraft uses the ADS-B to exchange data with the nearby ground station [16]. The execution of malicious actions by the adversary can lead to catastrophic consequences. Hackers can inject malicious packets to drive the UAV to unsafe zones or a fake aircraft presence to mislead the air traffic control or the nearby UAV to cause a collision or disrupt airspace traffic [8].

A. Spoofing Attack on GPS

GPS is the main component in driving the UAV in the planned direction. UAVs rely on the data sent by GPS to fly, but the link between GPS and UAVs is prone to adversaries. The spoofing of this link is the main threat to civilians or military. GPS spoofing can be a form of eavesdropper that listens to the transmission of data between UAVs and GPS signals in space [17]. It can cause the UAV to have the wrong position or cause manipulation of the predefined position to

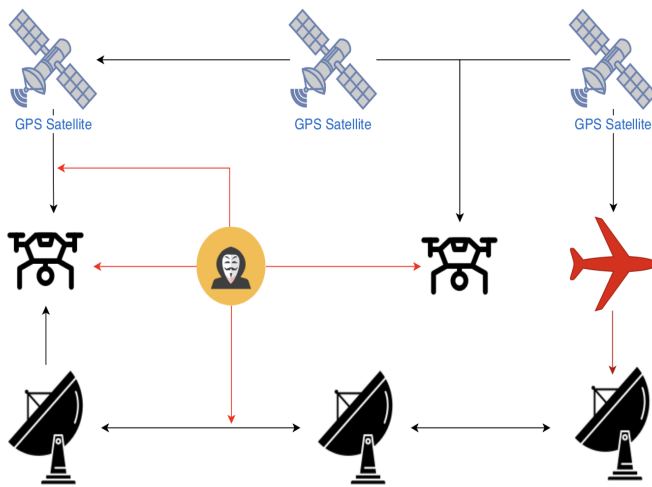


Fig. 6. Spoofing and Jamming Attack Framework on the UAVs

mislead UAV tasks. Therefore, the GPS spoofing signal can be effective in disrupting a military or civilian mission.

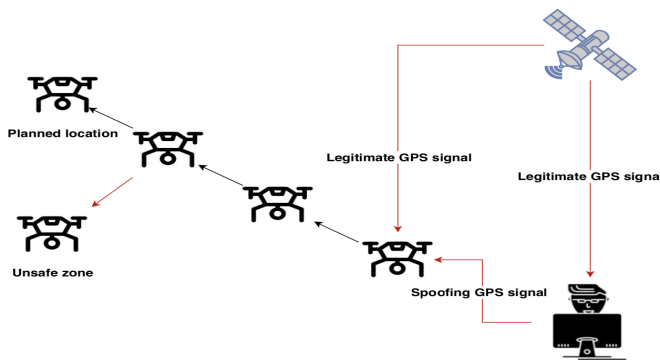


Fig. 7. Spoofing attack framework for UAVs

B. Spoofing attack on the ADS-B

ADS-B spoofing attack is executed in two forms: aircraft-based attackers and ground-based attackers [5]. Each of these attackers has a specific technique. In aircraft-based attackers, the adversary modifies the ICAO addresses to mislead the nearby aircraft or the aircraft control to present itself as a legitimate aircraft, as shown in Figure 7. In ground-based attackers, the adversary does not face any challenges, such as using high power to launch their attacks [18]. Malicious action is initiated using the low-cost SDR device.

C. Aircraft Spoofing Attack

aircraft spoofing attack hides its entity and is executed by spoofing ICAO addresses [5]. They spoof the ICAO addressee to perceive the aircraft as a legitimate entity to mislead either the air traffic control or nearby aircraft. Using this technique is an effective way where secondary radar surveillance can not detect the attack because the aircraft is physically present.

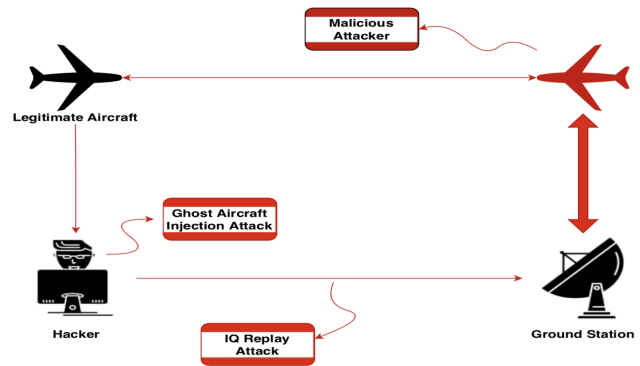


Fig. 8. ADS-B Spoofing Attack on UAVs

D. Message or IQ data replay attack

Message or IQ data replication attack is similar to the replay attack, but the way IQ data replication works is clever and stealthy. The attacker in this attack preserves the message that is re-generated and re-transmits it by SDR devices without any change or modification previously recorded in the messages [5].

E. Ghost aircraft injection attack

in the ghost aircraft injection attack, hackers send a fake ADS-B signal [5]. The adversary in this attack uses the SDR to execute this malicious action.

V. ADS-B SYSTEM BACKGROUND

Automatic dependent surveillance broadcast (ADS-B) is a technology system that was used to adopt new technology and ride of the traditional radar system [?]. It is a satellite-based system that is used to be deployed and equipped in aircraft to broadcast data. The two data links used by the ADS-B are 1090 and 987 MHz [?]. In addition, this system consists of two main components: ADS-B out and ADS-B in. The ADS-B out subsystem periodically broadcasts to the nearby entities (aircraft and ground station) the data related to the airplane identification such as ICAO addresses and the airplane statuses data such as velocity, altitude, and speed. The ADS-B on the device is used to receive a broadcast generated by a near-aircraft or ground station. The data received are spanning such as weather, ATC guidelines, and other data statuses. All these parameters lead to improve the performance of the aircraft in the sky, aircraft control, and avoid collisions in the space.

VI. PREVIEW WORKS AND DRAWBACKS

Recently, some techniques have been developed to resist jamming attacks. These jamming attack techniques vary depending on the type of jamming attack. Each of the four jamming attacks has characteristics that distinguish them from each other. For example, reactive jamming sensing the channel statuses if this channel started receiving the packets, reactive attack jamming attack emits a noise signal otherwise, still ideal [19]. Therefore, the reactive jamming attack is headed to be detected [20].

The traditional approach to jamming attacks is to redirect network traffic, change channel, frequency, or move to a safe zone to stop receiving the jamming signal again [21]. The famous technique to prevent jamming is frequency hopping and channel hopping [?]. In frequency hopping, once the disruption signal is presented over the network, the new frequency is chosen to be used over the network, so all receivers are asked to move on these new transmission frequency bands. Channel hopping is used frequently to countermeasure the integrity or blocking effect of the network [6].

Game theory has been widely used to resist against the jamming attack. Game theory models the jamming attack and the scenario to understand it and then builds resister techniques against the jamming attack [7]. In this theory, the jamming and the nodes are formulated as a follower and leader. The follower monitors the behavior of the leader and some signal features, such as signal power. Therefore, the follower launches its jamming signal based on the leader feature signals.

Innovation and the development of new techniques to protect UAV communication from being compromised are a continuous motivation of researchers to suggest methods to secure UAV communication. They innovated authentication, cryptography, and verification techniques that use the non-cryptography schema for the UAV security field [11]. The non-cryptography technique aims to verify the location of the UAV using the Kalman filter, distance bounding, etc.

Multichannel receiver techniques proposed against the jamming attack against the ADS-B jamming attack. Singular value decomposition used to use algebraic manipulation. These techniques provide different signals from different resources [9].

The increasing number of spoofing attacks and their catastrophic results encouraged researchers to continue to develop new techniques to withstand this attack. There are three methods that have been created for these goals, which are cryptography, authentication-based signals and external UAV communications [8]. Cryptography techniques are used to encrypt transmission packets and decrypt at the receiver through a shared key. In the authentication-based signal, the author [12] used two algorithms: ECC and RSA. In the third detection technique, anomaly detection uses characteristics to recognize attacks such as UAV speed and acceleration.

GPS spoofing attack to mislead the UAV or the automatic control to hijack the UAV to an unsafe zone. Adversity targets the GPS signal and counters the GPS signal to achieve its objectives. Researchers emphasize this vulnerability and innovate defense techniques to resist the GPS spoofing attack again. The GPS spoofing attack can be divided into three categories: signal processing techniques, hardware techniques, and combination of hardware and signal processing techniques [5]. In signal processing techniques, the adversary uses it to collect data from the transmitted and processed signals to use it to launch the attack. In the sequence, an attack is required to the presnet sensor and control system to lanch the attack. In the third spoofing techniques

VII. MACHINE LEARNING

Machine learning technology has been increasingly used in different aspects due to its characteristics in dealing with different missions. Machine learning is known as deep learning, which has several hidden layers that use different abstractions to learn new features [23].

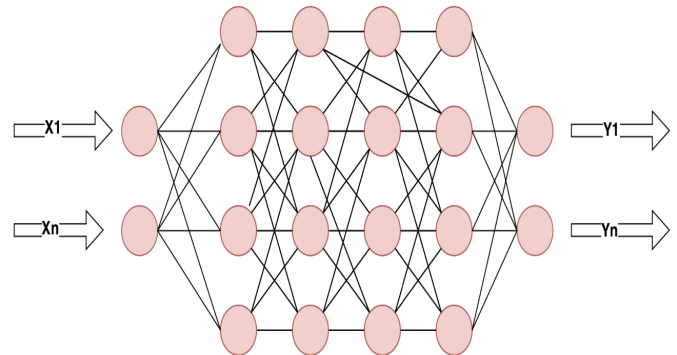


Fig. 9. Deep Learning Layer Structure

Deep learning aims to identify the attacker from unknown distributed inputs in layers [4]. This technique in deep learning is used to store new data and process them to learn new features as output. Therefore, the prevalence of deep learning is determined by the number of training examples. Deep learning performs better when it has enough hidden layers to be high. In the form section, show the classification of the DL into three sections supervised, unsupervised, and reinforcement learning.

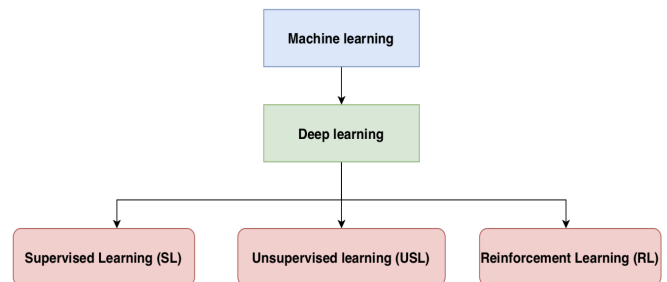


Fig. 10. Machine learning Algorithms

VIII. REINFORCEMENT LEARNING

Reinforcement learning algorithm is considered a reward algorithm. It is used to achieve optimal results. RL first uses the regimented learning process with a specific, specific parameter to achieve optimal results [?]. Security engineers use Q-learning, which is an extension of RL. They use it widely in defenses against jamming attacks.

IX. UNSUPERVISED LEARNING USL

Unsupervised learning algorithms have been widely used to identify patterns and define correlation without training in the predefined data set. It includes clustering in the USL to identify similarity between data groups [23].

X. SUPERVISED LEARNING

the supervised learning algorithm is trained on the pre-defined data set to reach specific output and precision. The predefined data set consists of two tubules label and attribute [23]. The label represents the output, while the attribute represents the input.

In classification tasks, algorithms are used to classify the input data into different classes. The security engineer uses these algorithms to classify different classes on the network under spoofing and other attackers. Also, it is useful to categorize other destructive executions such as malware and spyware.

In regression, this algorithm is used to predict specific output. In security fields, the regression algorithm is an efficient approach to predict the parameters of the data packet transferred over the network. Predicting these parameters helps security engineers distinguish the original and manipulated packets. It is useful to deny anomaly detection to recognize abnormal behavior such as a malicious log in or location. In addition, it can be used to build anomaly detection to recognize HTTP requests [23].

XI. FEATURE SELECTION

feature selection technique is an effective way to help security engineers build a robust ML model [1]. It leads to extracting details and shows how the correlation between features in the dataset. During the construction and training of the ML model, some ineffective features are presented. These features have to be removed [1]. However, these features are the reason for increasing the tiem'building model, complexity, and being the main reason for reducing accuracy. Therefore, the security engineer needs to use efficient feature selection techniques to extract usable features to overcome the ML model. Here are some feature selection:

A. Bit Error Rate (BER)

BER is used effectively to select features during non-coherent receivers. The presentation of the jamming attack implies an increase in the rate of the erroneous. for example, it is used in the ADS-B receiver and was calculated as [24]

$$BER = \frac{1}{2} \exp\left(-\frac{E_b}{2N_0}\right) \quad (1)$$

B. Spearman Correlation

The Spearman correlation is used to identify the feature after that ride of the correlated features [1]. The security engineer likes to use this technique to measure the monotonic relation that is presented between two variables. However, this technique is used to show all the details better than other feature selections. By using this technique, the security engineer has

the ability to demonstrate the strongest relationship between variables through:

$$r = \frac{Cov(X_r, Y_r)}{Std(X_r)Std(Y_r)} \quad (2)$$

C. Bad packet ratio (BPR)

The cyclic redundancy check (CRC) is the standard to recognize whether jamming attacks occur on continuous transmission or not. Given the ADS-B protocol, the CRC is used to check the received packets at the receiver if they fail, which is a sign of a jamming attack [1]. Therefore, increasing the error on the receiver side leads to increased drooping of the packet.

D. Extra Tree Classifier

Extra tree classifier is used to reduce complexity in computation and processing time. This technique selects the feature efficiently. In this technique, the pre-description is calculated by averaging the results using the decision tree. Therefore, this technique leads to the delivery of an accurate prediction, reduces variance, and overfits the control.

XII. DEEP LEARNING AND INTRUSION DETECTION

The traditional attack detection and defense methods are not suitable when it comes to resisting attacks or spoofing UAV network links. Limited resources and high mobility are the main reason. In addition, some detection techniques addressed in this article lead to updating the architecture of the ADS-B protocol. Therefore, these techniques are not effective in applying them to the communication of the UAV network.

Deep learning has been widely used in different fields with different tasks. Building intrusion detection systems using deep learning techniques has attracted researchers' attention. The deep learning algorithms are used to implement IDS accuracy to recognize the unknown attacker. They are also used to learn new features of the unknowable label data set. so, deep learning technology can build its data set by collecting unlabeled data sets.

XIII. CONCLUSIONS

The UAV shows its ability to be used and applied for different missions. Natural UAV communication links are subject to some threats that cause degradation in the performance of the UAV system. The adversary found that it is an easy task to listen to the transmission data traffic and execute their malicious actions. The jamming and spoofing attack was used with malicious intent to put the UAV under threat and cause catastrophic consequences. In this article, some attack defense and detection techniques that are used to detect these threats are addressed. However, these techniques are not suitable for application in UAV technology. Machine learning has emerged as a promising technology. Deep learning has the ability to learn and know new features from an unknown label data set, so it is an effective way to be used against jamming and spoofing attack even if they launched simultaneously.

REFERENCES

- [1] Slimane, H.O., Benouadah, S., Khoei, T.T. and Kaabouch, N., 2022, January. A Light Boosting-based ML Model for Detecting Deceptive Jamming Attacks on UAVs. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0328-0333). IEEE.
- [2] Greco, C., Pace, P., Basagni, S. and Fortino, G., 2021. Jamming detection at the edge of drone networks using Multi-layer Perceptrons and Decision Trees. *Applied Soft Computing*, 111, p.107806.
- [3] Manesh, M.R., Velashani, M.S., Ghribi, E. and Kaabouch, N., 2019, May. Performance comparison of machine learning algorithms in detecting jamming attacks on ADS-B devices. In 2019 IEEE International Conference on Electro Information Technology (EIT) (pp. 200-206). IEEE.
- [4] Dhomane, P. and Mathew, R., 2020. Counter-measures to spoofing and jamming of drone signals. Available at SSRN 3774955.
- [5] Ying, X., Mazer, J., Bernieri, G., Conti, M., Bushnell, L. and Poovendran, R., 2019, June. Detecting ADS-B spoofing attacks using deep neural networks. In 2019 IEEE conference on communications and network security (CNS) (pp. 187-195). IEEE.
- [6] F. alrefaei, "The Importance Of Security In Cyber-Physical System," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-3, doi: 10.1109/WF-IoT48130.2020.9221155.
- [7] Alrefaei, F., Alzahrani, A., Song, H. and Zohdy, M., 2020, September. Security of Cyber Physical Systems: Vulnerabilities, Attacks and Countermeasure. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.
- [8] Alrefaei, F., Alzahrani, A., Song, H., Zohdy, M. and Alrefaei, S., 2021, April. Cyber Physical Systems, a New Challenge and Security Issue for the Aviation. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-5). IEEE.
- [9] Alrefaei, F., Alzahrani, A., Song, H. and Zohdy, M., 2022. Aviation cybersecurity: a cyber-physical systems perspective. *Aviation Cybersecurity: Foundations, Principles, and Applications*, p.265.
- [10] Habler, E. and Shabtai, A., 2018. Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages. *Computers Security*, 78, pp.155-173.
- [11] Li, Y., Pawlak, J., Price, J., Al Shamaileh, K., Niyaz, Q., Paheding, S. and Devabhaktuni, V., 2022. Jamming Detection and Classification in OFDM-Based UAVs via Feature-and Spectrogram-Tailored Machine Learning. *IEEE Access*, 10, pp.16859-16870.
- [12] Wang, S., Wang, J., Su, C. and Ma, X., 2020, December. Intelligent Detection Algorithm Against UAVs' GPS Spoofing Attack. In 2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS) (pp. 382-389). IEEE.
- [13] Wang, J., Liu, Y. and Song, H., 2021. Counter-unmanned aircraft system (s)(C-UAS): state of the art, challenges, and future trends. *IEEE Aerospace and Electronic Systems Magazine*, 36(3), pp.4-29.
- [14] Wang, J., Juarez, N., Kohm, E., Liu, Y., Yuan, J. and Song, H., 2019, April. Integration of SDR and UAS for malicious Wi-Fi hotspots detection. In 2019 Integrated Communications, Navigation and Surveillance Conference (ICNS) (pp. 1-8). IEEE.
- [15] Wang, J., Liu, Y., Niu, S. and Song, H., 2020, December. 5G-enabled optimal bi-throughput for UAS swarm networking. In 2020 International Conference on Space-Air-Ground Computing (SAGC) (pp. 43-48). IEEE.
- [16] Xu, C., Zhang, K., Jiang, Y., Niu, S., Yang, T. and Song, H., 2021. Communication Aware UAV Swarm Surveillance Based on Hierarchical Architecture. *Drones*, 5(2), p.33.
- [17] Shafique, A., Mehmood, A. and Elhadef, M., 2021. Detecting signal spoofing attack in uavs using machine learning models. *IEEE Access*, 9, pp.93803-93815.
- [18] Liu, Y., Wang, J., Niu, S. and Song, H., 2021. ADS-B signals records for non-cryptographic identification and incremental learning. IEEE, Piscataway, NJ, USA, Data Set.
- [19] Wang, Z., Yu, Z., Liu, Y. and Song, H., 2021, August. Abnormal Data Detection of Unmanned Aerial Vehicles Based on Double Shortcuts ZB-ResNet. In 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC) (pp. 1-6). IEEE.
- [20] Eason, J., Xu, C. and Song, H., 2020, November. Software define radio in realizing the intruding uas group behavior prediction. In 2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC) (pp. 1-5). IEEE.
- [21] Wang, J., Liu, Y., Niu, S. and Song, H., 2021, December. Reinforcement Learning based Scheduling for Heterogeneous UAV Networking. In 2021 17th International Conference on Mobility, Sensing and Networking (MSN) (pp. 420-427). IEEE.
- [22] Romesburg, H., Wang, J., Jiang, Y., Wang, H. and Song, H., 2021, October. Software Defined Radio based Security Analysis For Unmanned Aircraft Systems. In 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC) (pp. 1-5). IEEE.
- [23] Asif, R., Hu, Y.F., Ali, M., Li, J.P. and Abdo, K., 2021, October. Signal Classification for Safety Critical Aeronautical Communications for Anti-Jamming using Artificial Intelligence. In 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) (pp. 1-6). IEEE.
- [24] Abdulaziz, A., Yaro, A.S., Adam, A.A., Kabir, M.T. and Salau, H.B., 2015. Optimum receiver for decoding automatic dependent surveillance broadcast (ADS-B) signals. *American Journal of Signal Processing*, 5(2), pp.23-31.

Highly Sensitive Hydrogen Gas Sensor Based on Fe₂O₃:ZnO Nanostructured Thin Film

Mikayel Aleksanyan

Artak Sayunts

Gevorg Shakhkhatuni

Zarine Simonyan

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
maleksanyan@ysu.am

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
sayuntsartak@ysu.am

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
gevshakhkhatuni@ysu.am

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
z.simonyan@ysu.am

Gohar Shahnazaryan

Vladimir Aroutiounian

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
Sgohar@ysu.am

Center of Semiconductor Devices and Nanotechnologies (Yerevan State University)
Yerevan, Armenia
kisahar@ysu.am

Abstract—This work aimed to manufacture a highly sensitive hydrogen gas sensor based on Fe₂O₃:ZnO thin film. The Fe₂O₃:ZnO based sputtering target was synthesized by a solid-state synthesis method. Gas-sensitive thin films were deposited onto alumina substrate by the high-frequency (RF) magnetron sputtering method. The thickness of the sensing film was measured and the morphology of the sensing layers was studied. The Fe₂O₃:ZnO sensor showed the highest hydrogen gas response at 100 °C and the operating temperature of 150 °C was selected as the optimum working temperature. The sensor showed sensitivity even to extremely low concentrations of hydrogen (100 ppm), where the response was 4.6. The sensing parameters of the fabricated sensor confirm that the Fe₂O₃:ZnO nanostructure is challenging for the detection of low concentrations of hydrogen.

Keywords—gas sensor, hydrogen, zinc oxide, thin film, sensitivity

I. INTRODUCTION

As know, the hydrocarbon-based energy sources on the Earth will be consumed in the next few decades, which is why intensive research is being done today to find alternative energy sources and to develop new technologies for this purpose. Among alternative energy sources, hydrogen is considered to be the most preferred due to its high calorific value, green emissions during combustion, and the presence of an unlimited amount on the Earth and in Space. Hydrogen has already been used as a source of energy in cars, airplanes, spacecraft, and in industry and technology, it is used as an essential gas for the synthesis of various materials and for carrying out various technological processes. Such widespread uses of hydrogen and their expansion in the future require the development of storage and transportation technologies for this gas. From this point of

view, hydrogen is not a preferred gas, as it is extremely flammable and explosive. The best way to avoid a hydrogen gas explosion is to use sensor systems. All this proves that the development of hydrogen sensors with high sensitivity, speed, low energy consumption, and advanced performance parameters are very urgent issues [1-7].

There are now hydrogen sensors on the market based on different physical principles, which do not meet modern requirements. The main problem is with the sensor performance parameters. The fact is that the existing sensors are not very fast, they work at high temperatures and they have a rather high detection limit and low sensitivity [8-13].

Nowadays, various nanostructured materials based on metal oxide semiconductors (MOS), such as SnO₂, In₂O₃, WO₃, ZnO, TiO₂, Fe₂O₃, CuO, and Ga₂O₃ are widely used in the manufacturing of resistive gas sensors sensitive to low concentration of hydrogen gas. As one of the most widely used oxides, hematite (Fe₂O₃) has n-type semiconducting properties with a bandgap of 2.1 eV under ambient conditions. It has the most thermodynamically stable phase among all iron oxides, low cost, nontoxicity, high chemical and temporal stability, and superior sensing performance towards reducing gases. Despite the above-mentioned advantages, pure iron oxide has fairly poor sensitivity to reducing gases and the introduction of dopants in pure Fe₂O₃ leads to the improvement of almost all performance parameters [14-20].

This work presents the design and development of a resistive sensor based on semiconductor Fe₂O₃:ZnO nanostructures with high sensitivity to low concentrations of hydrogen. The manufactured sensor can be successfully applied in a variety of

fields, as it has a low operating temperature, fast response/recovery behavior, high sensitivity and selectivity, and low gas detection limit.

II. EXPERIMENTAL

To synthesize the Fe₂O₃:ZnO sputtering target, Fe₂O₃ (99.9%) and ZnO (99.9%) nanopowders were purchased from Alfa Aesar. Fe₂O₃:ZnO ceramic target was synthesized by solid-phase reaction methods [21].

Using the Fe₂O₃:ZnO sputtering target thin sensing layer was deposited on Multi-Sensor-Platforms by VTC-600-2HD DC/RF Dual-Head High Vacuum Magnetron Plasma System. The Multi-Sensor-Platforms were purchased from TESLA BLATNA (Czech Republic). The factory-designed substrate consists of a temperature sensor (Pt 1000), a platinum heater, and interdigitated electrodes on the alumina substrate. The temperature sensor and heater were isolated with a glass layer. Fe₂O₃:ZnO layer was deposited onto the non-passivated platinum electrodes to convert the Multi-Sensor-Platform into a gas sensor (Fig. 1). The deposition time, input power, and substrate temperature during deposition were 20 min, 70 W, and 200 °C, respectively. In the final stage of the fabrication, palladium catalytic particles were deposited onto the sensing layer for sensitization of the active surface.

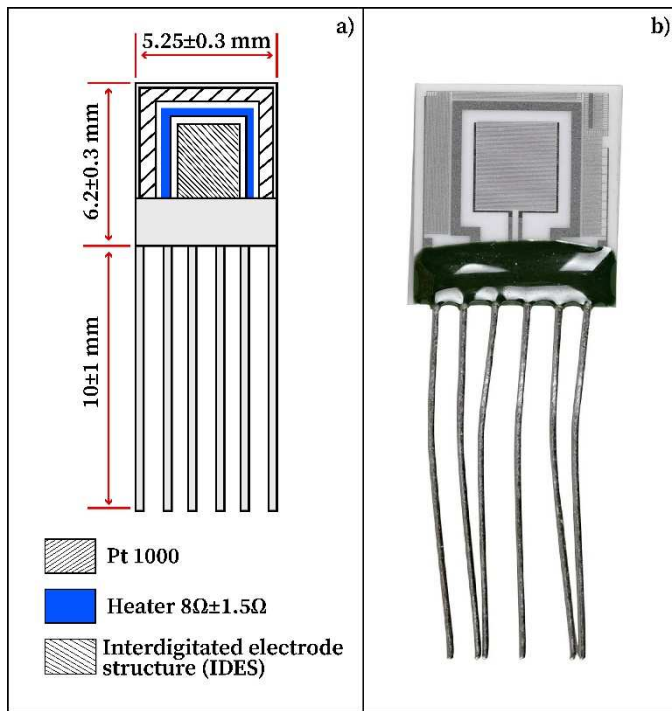


Fig. 1. The schematic diagram of the Multi-Sensor-Platform and the Multi-Sensor-Platform based sensor.

In the case of magnetron sputtering, the film thickness is controlled by the changing of sputtering time at a constant input power. One of the most important factors in terms of gas sensitivity is the thickness of the film, which was measured for 20 minutes of sputtering duration. The thickness of the film corresponding to this value was measured by the Alpha-Step D-300 (KLA Tencor) profiler. As shown in Fig. 2, the thickness of the Fe₂O₃:ZnO film was equal to 144 nm.

Morphological characteristics of the gas sensor have a great influence on sensor performance, the study of which is successfully carried out with the help of scanning electron microscopy (SEM). The SEM study of the Fe₂O₃:ZnO film was carried out by Vega 5130 MM (Tescan) (Fig. 3). The film has a granule structure with an average grain size of 150 nm.

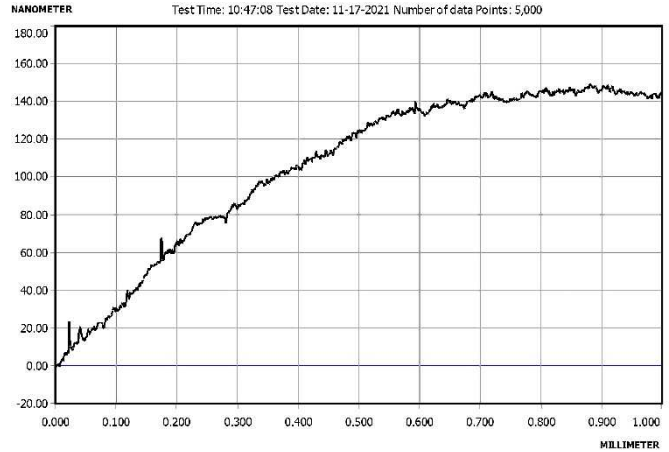


Fig. 2. The thickness measurement result of Fe₂O₃:ZnO film.

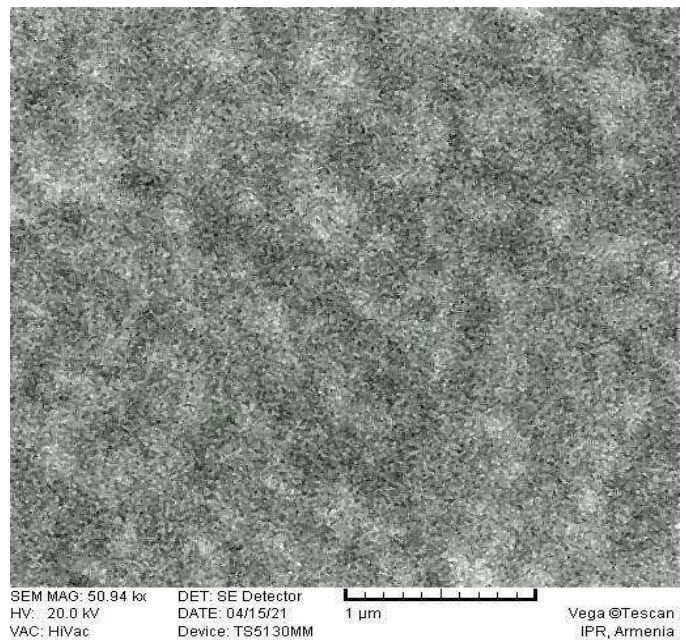


Fig. 3. SEM image of Fe₂O₃:ZnO film

The hydrogen sensing performance of the Fe₂O₃:ZnO sensor was tested by the laboratory-made computer-controlled gas testing system, which contains a test chamber, pressure sensor (Motorola-MPX5010DP), and a data acquisition system (PCLD-8115) [22]. For gas sensing measurements, the Fe₂O₃:ZnO sensor was installed in the gas chamber connecting

the six pins (two pins of a temperature sensor, two pins of the heater, and two pins of resistance measurement electrodes (Fig. 1) with the corresponding inputs on the sensor holder.

III. GAS SENSOR STUDY

The manufactured sensors are resistive type and the operation of this type of sensor is based on a change in the electrical resistance of a gas-sensitive layer under the influence of target gases. This is due to the exchange of charges between the molecules of the adsorbed target gases and the semiconductor film. So, the sensor response is defined as $S=R_a/R_g$, where R_a and R_g are the electrical resistances of the active layer in air and the presence of a target gas, respectively. Response and recovery time is defined as the time required to achieve 80-90% of the change in resistance from the corresponding steady-state value of each signal. The gas sensing characteristics of the $Fe_2O_3:ZnO$ sensors in the presence of hydrogen gas are shown below.

The sensing characteristics of the manufactured sensor were studied in the temperature range of 50–250 °C. The heater of the sensor platform allowed the temperature of the sensor to raise the temperature to 250 °C. All real-time measurements of the electrical resistance were carried out at 3 V DC voltage applied on the electrodes of the sensor.

The $Fe_2O_3:ZnO$ sensor demonstrated the maximal response value to 2000 ppm of hydrogen at 100 °C temperature where the resistance of the sensor changed more than 5000 times (Fig. 4).

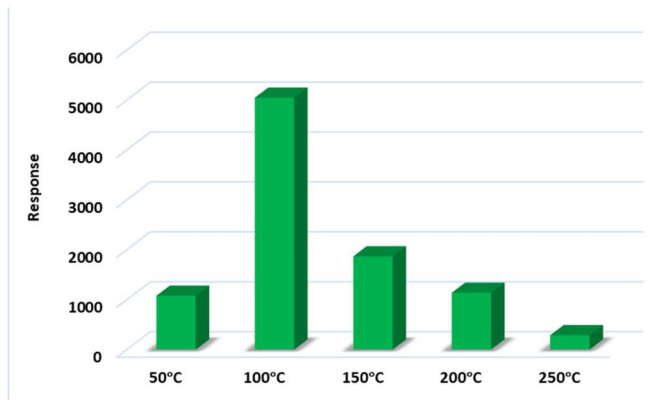


Fig. 4. Dependence of the response of the $Fe_2O_3:ZnO$ sensor on the operating temperature at 2000 ppm of hydrogen.

At 50 °C, 150 °C, and 200 °C, the sensor response was higher than 1000, while the response value was significantly lower (287) at 250 °C. The real-time resistance change of the $Fe_2O_3:ZnO$ sensor at different operating temperatures toward 2000 ppm hydrogen is revealed in Fig. 5. Despite high responses in the temperature range of 50-100 °C, the significantly lower response and recovery times, as well as the higher recovery levels of the base resistance were demonstrated at 150 °C. At higher temperatures (>200 °C) the sensor response decreased and the energy consumption increased. Consequently, the temperature of 150 °C was determined as the operating temperature for hydrogen detection. The response and

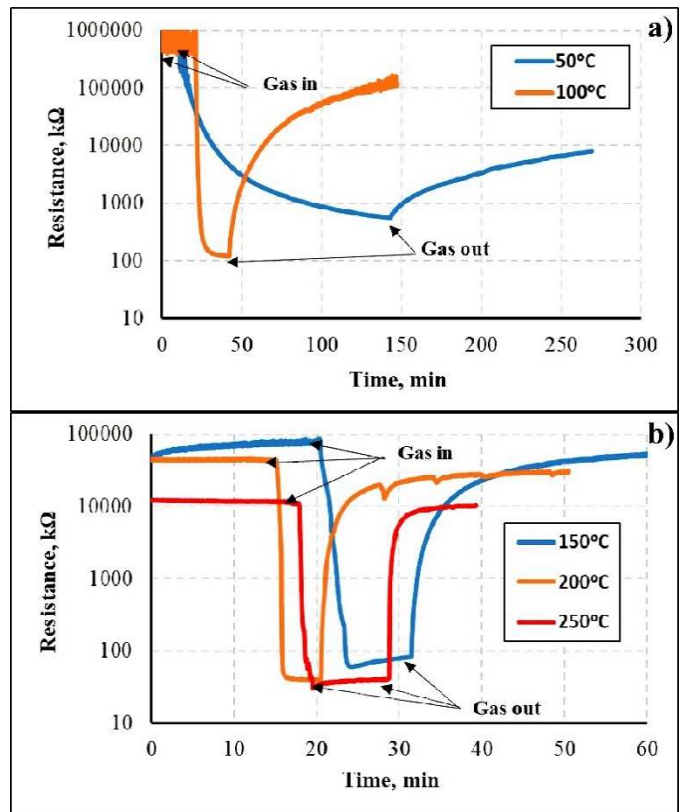


Fig. 5. The real-time resistance change of the $Fe_2O_3:ZnO$ sensor at different operating temperatures toward 2000 ppm hydrogen.

recovery times of the sensor toward 100 ppm hydrogen at operating temperature were found to be 9 min and 21 min, respectively.

The time-dependent response curves of the sensor in the presence of different concentrations of hydrogen and the dependence of the response of the sensor on hydrogen gas concentration at the operating temperature are presented in Fig. 6 and Fig. 7, respectively.

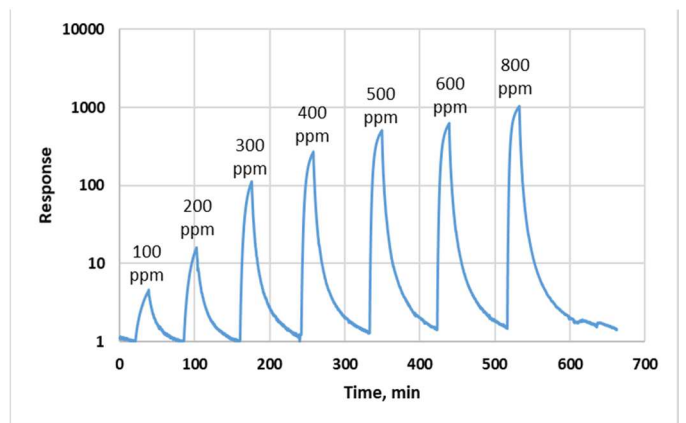


Fig. 6. The real-time response of the $Fe_2O_3:ZnO$ sensor at different concentrations of hydrogen gas at 150 °C operating temperature.

The sensor showed a clear and recoverable response in the hydrogen concentration range of 100-800 ppm. The sensor showed sensitivity even to extremely low concentrations of

hydrogen (100 ppm), where the response was 4.6. Toward the higher concentration (800 ppm), the response value exceeded 1042, which is a very promising result. The dependence of the response on hydrogen concentration has a linear characteristic which will allow not only to detect of H₂ traces but also to accurately measure the low concentrations of this gas. In order to check the selectivity, the gas sensing characteristics of the sensor were studied in the presence of acetone, ethanol, toluene, methane, and LPG. The sensor did not show any significant response to 400 ppm, 675 ppm, 700 ppm, 2000 ppm, 1000 ppm of acetone, ethanol, toluene, methane, and LPG, respectively, compared with the response to 300 ppm of hydrogen (Fig. 8), which confirmed the high selectivity of the sensor.

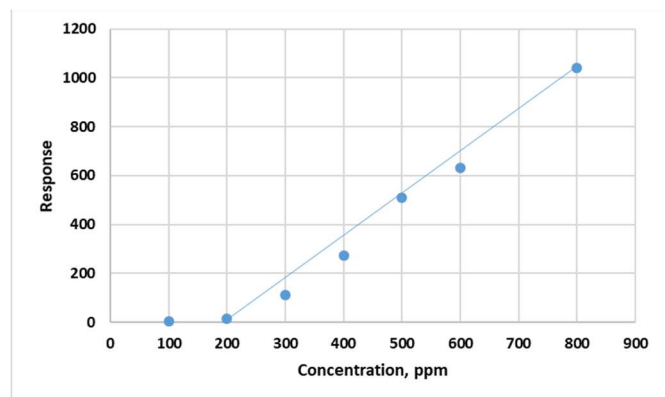


Fig. 7. Dependence of the response of the Fe₂O₃:ZnO sensor on hydrogen gas concentration at the 150 °C operating temperature.

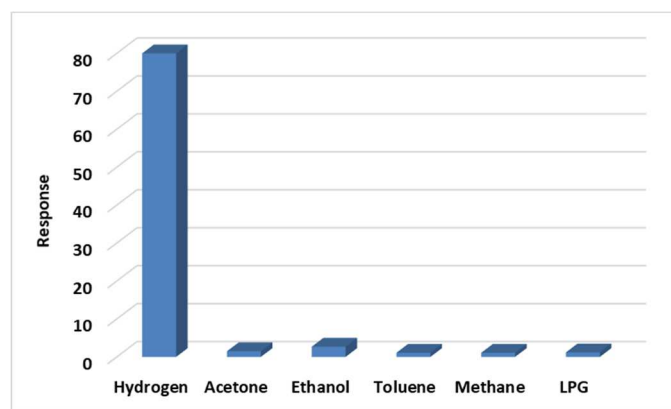


Fig. 8. Comparison of responses of the sensor to 300 ppm of hydrogen, 400 ppm of acetone, 675 ppm of ethanol, 700 ppm, 2000 ppm of methane, and 1000 ppm of LPG.

IV. CONCLUSIONS

A nanostructured film based on the Fe₂O₃:ZnO material was deposited on an alumina substrate by high-frequency magnetron sputtering and the active surface was sensitized by Pd catalytic particles. The sensor showed a clear and recoverable response in the hydrogen concentration range of 100-2000 ppm and the dependence of the response on hydrogen concentration has a linear characteristic. The high selectivity of the sensor toward other environmental gases was also confirmed.

ACKNOWLEDGMENT

The work was supported by the Science Committee of MESCS RA, in the frames of the research project № 21APP-2J001.

REFERENCES

- [1] C. Zhahg, X. Chen, X. Liu, C. Shen, Z. Huang, et al., "High sensitivity hydrogen sensor based on tilted fiber Bragg grating coated with PDMS/WO₃ film", *Int. J. Hydrogen Energy*, vol. 47, pp. 6415-6420, January 2022.
- [2] T. Wischmeyer, J. R. Stetter, W. J. Buttner, V. Patel, and D. Peaslee, "Characterization of a selective, zero power sensor for distributed sensing of hydrogen in energy applications", *Int. J. Hydrogen Energy*, vol. 46, pp. 31489-31500, September 2021.
- [3] S. Agarwal, S. Kumar, H. Agrawal, M. G. Moinuddin, M. Kumar, et al., "An efficient hydrogen gas sensor based on hierarchical Ag/ZnO hollow microstructures", *Sens. Actuators B Chem.*, vol. 349, 130510, November 2021.
- [4] G. Xin, C. Ji, S. Wang, H. Meng, K. Chang, et al., "Monitoring of hydrogen-fueled engine backfires using dual manifold absolute pressure sensors", *Int. J. Hydrogen Energy*, vol. 47, pp. 13134-13142, March 2022.
- [5] K. Chen, D. Yuan, and Y. Zhao, "Review of optical hydrogen sensors based on metal hydrides: Recent developments and challenges", *Opt. Laser Technol.*, vol. 137, 106808, May 2021.
- [6] H. Li, C.-H. Wu, Y.-C. Liu, S.-H. Yuan, Z.-X. Chiang, et al., "Mesoporous WO₃-TiO₂ heterojunction for a hydrogen gas sensor", *Sens. Actuators B Chem.*, vol. 341, 130035, August 2021.
- [7] Y. Zhao, Q. Liu, Y. Duan, Y. Zhang, Y. Cui, Y. Huang, D. Gao, L. Shi, J. Wang, and Q. Yi, "Hydrogen energy deployment in decarbonizing transportation sector using multi-supply-demand integrated scenario analysis with nonlinear programming — A Shanxi case study," *Int. J. Hydrog. Energy*, in press.
- [8] Y. Nishijima, K. Enomoto, S. Okazaki, T. Arakawa, A. Balčytis, et al., "Pulsed laser deposition of Pt-WO₃ of hydrogen sensors under atmospheric conditions", *Appl. Surf. Sci.*, vol. 534, 147568, December 2020.
- [9] N. Pradeep, G. T. Selvi, U. Venkatraman, Q. V. Le, S. K. Jeong, et al., "Development and investigation of the flexible hydrogen sensor based on ZnO-decorated Sb₂O₃ nanobelts", *Mater. Today Chem.*, vol. 22, 100576, December 2021.
- [10] Z. Cai, and S. Park, "Synthesis of Pd nanoparticle-decorated SnO₂ nanowires and determination of the optimum quantity of Pd nanoparticles for highly sensitive and selective hydrogen gas sensor", *Sens. Actuators B Chem.*, vol. 322, 128651, November 2020.
- [11] Q. Liu, J. Yao, Y. Wang, Y. Sun, and G. Ding, "Temperature dependent response/recovery characteristics of Pd/Ni thin film based hydrogen sensor", *Sens. Actuators B Chem.*, vol. 290, pp. 544-550, July 2019.
- [12] K. Chen, D. Yuan, and Y. Zhao, "Review of optical hydrogen sensors based on metal hydrides: Recent developments and challenges," *Opt. Laser Technol.*, vol. 137, p. 106808, May 2021.
- [13] F. Downes and C. M. Taylor, "Theoretical investigation of a multi-channel optical fiber surface plasmon resonance hydrogen sensor," *Opt. Commun.*, vol. 490, p. 126916, July 2021.
- [14] J.-H. Kim, A. Mirzaei, H. W. Kim, and S. S. Kim, "Improving the hydrogen sensing properties of SnO₂ nanowire-based conductometric sensors by Pd-decoration", *Sens. Actuators B Chem.*, vol. 285, pp. 358-367, April 2019.
- [15] Y. Luo, B. An, J. Bai, Y. Wang, X. Cheng, et al., "Ultrahigh-response hydrogen sensor based on PdO/NiO co-doped In₂O₃ nanotubes", *J. Colloid Interface Sci.*, vol. 599, pp. 533-542, October 2021.
- [16] S. Vallejos, I. Gràcia, T. Lednický, L. Vojkuvka, E. Figueras, et al., "Highly hydrogen sensitive micromachined sensors based on aerosol-assisted chemical vapor deposited ZnO rods", *Sens. Actuators B Chem.*, vol. 268, pp. 15-21, September 2018.
- [17] S. Saritas, M. Kundakci, O. Coban, S. Tuzemen, and M. Yildirim, "Ni: Fe₂O₃, Mg: Fe₂O₃ and Fe₂O₃ thin films gas sensor application", *Physica B: Condens. Matter*, vol. 541, pp. 14-18, July 2018.

- [18] V. Manikandan, "Real environment humidity-sensing ability of Nd-doped Fe₂O₃ sensor", *Sensing and Bio-Sensing Res.*, vol. 33, 100439, August 2021.
- [19] O. Alev, N. Sarıca, O. Özdemir, L. Ç. Arslan, S. Büyükköse, and Z. Z. Öztürk, "Cu-doped ZnO nanorods based QCM sensor for hazardous gases," *J. Alloys Compd.*, vol. 826, p. 154177, June 2020.
- [20] P. Patial and M. Deshwal, "A platinum-doped ZnO-based LPG sensor with high sensitivity," *Mater. Today: Proc.*, vol. 48 part 5, pp. 1201-1204, 2022.
- [21] M. Aleksanyan, A. Sayunts, H. Zakaryan, V. Aroutiounian, V. Arakelyan, et al., "Investigations of Sensors for Detection of Hydrogen Peroxide Vapors under the Influence of UV Illumination", *J. Contemp. Phys.*, vol. 55, pp. 205-212, September 2020.
- [22] M. Aleksanyan, A. Sayunts, G. Shahkhatuni, G. Shahnazaryan, and V. Aroutiounian, "Study of Gas Sensitivity of SnO₂ (Nb) Film in Liquefied Petroleum Gas", *J. Contemp. Phys.*, vol. 56, pp. 139-145, June 2021.

A Review on Trends in the Northern Virginia (NOVA) Housing Market and Understanding Home Characteristics for ML Models

Bethlehem Belaineh
George Mason University
bbelaine@gmu.edu

Omar Janjua
George Mason University
mjanjua@gmu.edu

Paul Karcic
George Mason University
pkarcic@gmu.edu

Ryan Thomas
George Mason University
rthoma43@gmu.edu

Abstract—This study aims to identify trends in the real estate market, as a lens to understand which home characteristics most are considered by home buyers in purchasing a home. This paper will uncover which important features (factors) need to be considered, how those factors affect the price of the home and gain an understanding for the geographic distribution of home types. The Northern Virginia (NOVA) housing market has witnessed various changes during the global pandemic, namely: increased migration of professionals in the region seeking employment opportunities with new companies such as Amazon, Facebook (Meta), and government-interfacing startups. The increased employment opportunities, along with additional construction activities has contributed directly to changes in mortgage and interest rates in the region. Accordingly, our predicted home prices in NOVA in 2023 are \$702K, \$700K, \$699K based on interest rates of 2%, 2.5%, and 3% respectively. According to our Zillow data analysis, the median home listing price as of March 2023 for the NOVA area was \$1,052,690. Comparatively, homes in Fairfax County are larger in square footage, and number of bedrooms and bathrooms, while Alexandria County has on average smaller sized homes.

Index Terms—Housing, Investment, Machine Learning, features

I. INTRODUCTION

Home prices have soared in the last decade, especially during the global pandemic. Within the

last year, we have observed home prices drastically shoot up while the national household income fell [1]. The current economic movements have made it difficult for first time homebuyers to secure their first home. Northern Virginia, due to its close proximity to Washington, D.C., a major metropolitan hub, has been affected more dramatically, which led to a price increase of 7% from October 2020 through 2021 (Kashino, 2020) [2]. In our analysis below, we will identify which factors contribute to an uneven real estate market, as well as look into the home characteristics that make certain listings more popular among buyers.

II. BACKGROUND

Purchasing a home is the biggest purchase in most Americans' lifetime. It allows them to not only have a home to live in, but also provides them with an opportunity to get a mortgage which will be a fixed cost over its lifespan, opposed to renting which continues to rise every year (Picchi, 2022) [3]. A paid off home is considered one of the foundations of the American dream, allowing homeowners to eventually retire without a monthly payment. Additionally, when the mortgage is finished, the homeowners are left with an asset that has greatly appreciated over the life of the loan.

With the rise of housing prices, most first home buyers are facing a massive barrier of entry as even the traditional 20% down payment can be in the hundreds of thousands of dollars. By being unable to secure a mortgage, first time homebuyers are stuck renting (Stahl, 2020) [4]. As rents increase in parallel to home prices, it becomes increasingly challenging to buy a home. Given this situation, it would take a long time to save enough for a down payment. Even when the mortgage is secured, the monthly mortgage would be much higher than if they had been able to purchase the property years ago. This causes Americans to have increasingly slower starts in building personal equity, making it harder to plan and save for retirement. By analyzing historical home prices data in Northern Virginia, we may be able to identify significant trends in the housing market and potentially increase future housing accessibility for first time homeowners.

New home buyers can also be assisted by gaining a better understanding of which important features to consider when buying a home. With such a big investment, it is important to think beyond immediate needs and to consider long-term goals and plans. It is important for new home buyers to understand which home attributes provide the most fiscal and livable value to the home. Location, lot size, number of bedrooms, and number of bathrooms are some of the most important attributes to consider (HOMEia) [5]. Homes tend to differ depending on their region and a new homebuyer can benefit by better understanding the geographic difference in homes in their area of interest.

III. ECONOMIC FACTORS TO BE CONSIDERED

In our analysis we described trends in the Northern Virginia (NOVA) housing market over the past two decades. According to the Survey of Consumer Finances by the Federal Reserve, 64.9% of American families owned their own primary residence in 2019 (“Changes in U.S. Family Finances from 2016 to 2019”). Hence for most people, real estate represents a significant portion of their wealth. As a family considers investing in real estate, a number of factors need to be considered: real estate prices, availability, and investment potential (Nguyen) [6].

Economists predict that real estate prices often follow the cycle of the economy. A major indicator

of economic activity is interest rates, as such we will be looking at the effect interest rates have historically on home price and identify a model for predicting future home valuations based on future interest rates. According to the economists Pažický and Falath, the US faces long-term elevated inflation rates over the coming decade as a result of the COVID pandemic (Pažický and Falath) [7]. This is in accordance with most sources we surveyed. The most frequented real estate mobile applications and websites predict increases in real estate prices in 2022 as rates were estimated to remain largely close to 11% to 2%. The following companies predicted an increase in real estate prices for 2023:

A. Projected Increases:

- Zillow predicted an increase of 11% in overall home prices across the U.S. [8]
- Fannie Mae predicts a 7.9% increase Freddie Mac predicts a 7% increase [9]
- Redfin Predicts A 3% Increase [10]
- Realtor.com Predicts A 2.9% Increase [11]
- Goldman Sachs predicts prices to increase by 16% (Walker) [12]

B. Projected Decreases:

- The Mortgage Bankers Association forecast model projects a 2.5% decrease in the median price of existing homes by the end of 2022 (“Northern Virginia Housing Market Forecast 2022 — Mashvisor”)

The housing market recorded a 40-year low in inventory this year—and this has put homebuyers at a greater risk of not securing their home. The ‘Bidding War’ has commenced, with a staggering 74% of homes receiving multiple competitive offers (“Housing Market Bidding War News - Redfin”) [10]. A topic of further study is understanding the unknown factors that contributed to such a range of predictions from the earlier section (Economic Factors). A topic of further study is to understand: why is there great level of uncertainty when it comes to the 2022 housing market?

Although in depth economic analysis is beyond the scope of the paper—we need to stress the importance of incorporating economics into gaining a holistic view of the real estate market for further study. From understanding inflation prices,

where the Federal Reserve predicted that inflation would remain relatively tame (at about 1.8%)—to now gaining perspective in knowing that the latest confirmed reading denoted a significant increase of inflation—6.2%, which is the highest rate of inflation since 1990 (Federal Reserve Data) [13]. Higher than expected inflation means that the federal reserve would raise the federal funds rate—a rate that has intentionally been kept zero during the pandemic to encourage more spending.

If the federal reserve raises rates, then it is only natural that the average 30-year fixed mortgage rate (which is currently at 2.98%) also rises (Federal Reserve Data) [13]. The increase in mortgage rates would directly influence higher monthly payments—which would disqualify buyers out of the market (Lambert) [14]. In addition to this context, the millennial workforce in most metropolitan cities are working from home—which has pushed the demand for suburban homes in greater numbers than previously recorded (Lambert) [11]. This would be a topic for further study, to analyze the impact of ‘Work from Home’ culture on the real estate market.

IV. LITERATURE REVIEW

James Conner published a U.S. Department of Housing and Urban Development (HUD) Housing Market Profile on Northern Virginia in June of 2021 which discusses economic activity, population changes, building activity, and sales and rental market conditions focusing on the most recent 24 months. The report covers 11 counties and 6 independent cities in the Northern Virginia Statistical Area. Northern Virginia sales market condition is characterized as “tight” with strong housing demands and high housing cost due to Northern Virginia’s relatively high incomes [15].

The Coronavirus pandemic has caused a strong increase in demand for home ownership in the region and the housing market had already been tightening prior to the onset of the pandemic. During the pandemic, prices for both existing homes and new homes increased while the number of existing homes sold rose 12 percent and the number of new homes sold decreased 12 percent. The pandemic slowed down construction efforts for new homes but construction has strengthened recently.

Professional and business services is the largest sector in Northern Virginia but experienced the smallest percentage decrease in jobs among all sectors throughout the pandemic. Overall, unemployment in the area has somewhat recovered from the pandemic. From May 2020 to May 2021, the economy in Northern Virginia outperformed the larger metropolitan area and the nation. The rate of seriously delinquent mortgages increased during the pandemic but were less than the Virginia and National rates. Recent building activities such as Amazon HQ2 in Arlington, Capital One in Falls Church, and Google expansion in Loudoun increase regional employment and may increase the housing market demand [15].

We identified that it’s imperative to look at the impact of monetary policy on housing prices in our evaluation. We will be looking at three data sets originated from the Federal Reserve, who set the interest rates for banks and lenders nationwide, as well as from the Federal Housing Finance Agency who are responsible for the supervision and regulation of the mortgage market, as well as secondary lenders such as Fannie Mae and Freddie Mac.

In addition to the recent economic changes and challenges the Northern Virginia housing market has experienced, the available inventory of homes in the region has dropped considerably. As of March 2022, there were 2,470 active listings in Northern Virginia in comparison to 3,029 active listings in March 2021. In combination to this 18% decrease, the average days on market has also decreased by 13% in comparison to March 2021 (Northern Virginia Housing Market Data). These changes emphasize the need for new home buyers to accurately assess the value of homes independent of economic conditions. This kind of assessment would allow new home buyers to better understand how much more or less they would potentially be spending on a home if listed at a different price.

A variety of research has been conducted on analyzing descriptive characteristics of homes to accurately assess the value of homes. Lian Pardoe proposes a multiple linear regression model analysis of home price data in his article “Modeling Home Prices Using Realtor Data”. Pardoe models’ sale price using a variety of property features as

predicting variables. He explains effective ways to modify and prepare the data for model creation such as converting certain numeric variables to ordinal variables, so the variables have a more realistic impact on the linear regression model. Pardoe discusses the use of statistical testing to determine which attributes are most correlated to the response variable before constructing the model. Pardoe also discusses the use of predictor effect plots which can be used to graphically show how a regression response variable changes with changes to the predictor variables. (Pardoe, 2017) [16].

V. DATASET

A. Northern Virginia Housing Data

We first explored a variety of online databases to find recent, relevant and reliable housing market data for the Northern Virginia area. The Northern Virginia Realtors (NVAR) [17] website provides historical month data broken down by monthly market stats spanning from 2000 - 2021 [18]. Northern Virginia counties include Arlington, Alexandria, Fairfax County, Fairfax City, Loudoun County, Falls Church, and Prince William. The attributes include those listed in yellow in Table ???. The original data downloaded was not formatted in a usable format for analysis. We transposed the data using excel to a more usable data format and the data was cleaned using R Studio (Historical Monthly Data) [19].

Originally, we decided to narrow our project scope to study the market from 2011-2021 and have a county level of granularity. We later decided to expand our dataset to include 2000-2021 to include more context to the history of the market. We also decided to aggregate counties together to have a holistic northern Virginia region dataset to allow the data to be merged with economic data. It is important to note that there is missing data for Prince William county for the months of 2000 through 2012 (Historical Monthly Data) [19].

Data for the Federal interest rates (blue) for each month from 2000-2021 was downloaded from the Federal Reserve database and merged with the NVAR housing data (Monthly Interest Rate Data). Data for the national average mortgage contract rate (green) and national average mortgage loan amount in thousands (green) was downloaded

from the Federal Housing Finance Agency through their National Mortgage database and merged with NVAR housing data (Table ??) (National Mortgage Database (NMDB) Aggregate Data) [20]. Data for Northern Virginia metropolitan area non-farm employees in thousands (red) was downloaded from the U.S. Bureau of Labor Statistics' BLS Data Finder and was merged with NVAR housing data (Table ??) (U.S. Bureau of Labor Statistics) [21].

B. Northern Virginia Zillow Home Listings

We also collected a dataset of active home listings for sale posted on Zillow in Arlington, Alexandria, Fairfax County, Fairfax City, Loudoun County, Falls Church, and Prince William County as of March, 23rd 2022 to explore differences in characteristics of homes in the different counties and see what factors are most predictive for a home's price (Table I). The dataset was collected using Zillow.com API in Python Jupyter Notebook. The code was executed using a Rapid API host and our Rapid API key. There was a total of 829 listings x 205 columns of features in the uncleaned original version.

Data cleaning was performed using Python. Each county's listings were collected independently then concatenated together. Eleven useful variables were extracted, missing data was either removed or managed effectively, and outliers were removed. Outlier listings were also removed based on outlier price of greater than \$4,334,000 and outlier lot size greater than 101,496 square feet. Duplicate listings were also removed by identifying duplicate street addresses. Data types were transformed for useful analysis. For example, timeOnZillow(days) values were originally formatted as "6 days". Lot Size values were either in square feet or acres units so acre values were converted to square feet by multiplying by a 43560 conversion factor. The final dataset is 396 rows x 11 columns.

VI. METHODOLOGY

The main purpose of the project is to explore the Northern Virginia (NOVA) real estate market from a time series perspective and a home characteristics perspective. The research questions of the project first section are the following:

TABLE I: Cleaned Northern Virginia Zillow Home Listings Data Variables Used

Variable Name	Description	Example
county	county	Alexandria
livingArea	living area square feet	1540
bedrooms	total number of bedrooms	2
bathrooms	total number of bathrooms	2
resoFacts.bathroomsFull	number of full bathrooms	2
resoFacts.bathroomsHalf	number of half bathrooms	0
resoFacts.hasGarage	whether the property has a garage or not (0 or 1)	0
resoFacts.garageSpaces	number of garage spaces	0
resoFacts.Stories	number of stories of the home	3
lotSize(sqft)	total size of lot in square feet	3920
resoFacts.isNewConstruction	whether or not the property is new construction or not (0 or 1)	0

TABLE II: Outlined Steps for Project Replicability

Step	Action	Resource
Step One	Collect relevant datasets	- Various offline online databases - Zillow Rapid API
Step Two	Data pre-processing and cleaning	- Excel - R - Python
Step Three	Compile necessary datasets	- House Pricing Data - Mortgage Data - Interest Rate Data
Step Four	Visualizations and Analysis	- Tableau - PowerBi - R
Step Five	Create a model incorporating a combination of the factors collected	- R or Weka

- What impact will increasing interest rates have on the values of NOVA home prices as a whole?
 - How does interest rates compare to other mortgage data such as the National Average Contract Mortgage Rate (NACMR) in regards to housing prices?
- What impact do other non-structure related attributes have on home prices?

Secondly, we will explore differences in home characteristics across the same counties by examining current Zillow home listings. The research questions of this project section are the following:

- Which counties have the most listings per capita and which counties have higher average prices?
- What is the distribution of characteristics on average across counties?
 - Example: Do listings in Fairfax County have larger living areas on average than all other counties?
- What home attributes have the most predicting power and influence on the price of home listing?

We will be exploring trends across different counties within NOVA, specifically historic real estate trends and active listings. Our paper will cover the following counties and cities:

- Arlington County
- Fairfax County
- Loudoun County
- Prince William County
- Alexandria City
- Fairfax City
- Falls Church City

We began with a collection of datasets relevant to our project including real estate data, economic data, and home listings data. Next, we conducted data preprocessing and cleaning methods such as normalization, data transformation, missing data management, and attribute selection. Necessary datasets were then merged or concatenated together. We then performed exploratory analysis of the data and created data visualizations which provide valuable insight into the trends and current state of the market. Finally, we created a model to predict future home prices based on projected interest rates. We also plan to develop a model to predict the price of homes based on descriptive characteristics of the property (Table #). Further research will be conducted to determine the best model type to employ for our specific application.

VII. PROPOSED METHOD FOR EVALUATION

It is interesting to note the major differences in listing prices from our historical analysis and our current market analysis. Our historical analysis of home prices showed an average home price in the region of \$707,990 in 2021 while our active listing analysis showed an average home price in the region of \$1,345,351 as of March 2023. This almost doubled figure difference presents challenges in evaluating our result. Although these averages were calculated for periods 2 years apart, It is unlikely the difference is realistically as dramatic. The large difference is likely due to the Zillow data having a right skewed distribution with a mean of \$1,345,351 and a median of \$1,052,690. We plan to further investigate this difference in the future.

We explored the statistical correlation between home listing attributes and home listing price to augment the regression model. As such we calculated the Pearson product-moment correlation

coefficient (PPMCC), to understand the correlation between interest rates and home prices. The standard error calculated from the data sample we took from the regression data.

VIII. RESULTS

A. Historical Analysis Northern Virginia Average Home Sold Price In Relation To Interest Rates

In our analysis we aimed to use interest rates as a predictor for home prices for 2023. We had already seen 2022 prices continue to rise with interest rates remaining low as predicted by major lenders such as Sallie Mae and Freddie Mac. Historically, interest rates have been associated as an indicator of housing prices (Sutton, 2017) [22]. As interest rates go down, house prices are expected to increase as homeowners will benefit from the lower interest on mortgages and will pay a higher premium on the purchase price for this trade off. Our datasets for mortgage rates only went up to 2018 for NOVA, thus we narrowed down the scope to just interest rates as they are very closely correlated regardless.

However, we found that over the last two decades home prices tended to steadily rise, regardless of interest rate (figure 1). During the Covid-19 pandemic, home prices skyrocketed. Although the latter could be attributed to interest rates, there were many factors which led to the increasing prices of homes such as the trillion dollar stimulus to facilitate economic growth during lockdowns, as well as the increase in work from home jobs. As these last two decades were atypical with the great recession of 2008 and the Covid-19 crisis, we assume that there may be a reversion to the mean with interest rates having a more direct impact on housing prices in NOVA again moving forward.

Our simple linear regression model showed that over the last 20 years, housing prices dropped \$24,850 per a 1% increase in federal interest rates (figure 2). With this calculation we aim to predict the trend in housing prices based on the federal reserve rate hike plan for 2023.

In March of 2022, the Federal Reserve announced their intention to hike interest rates by an average of .25% over the remaining 6 sessions of 2022, as well as three hikes in 2023 for an expected interest rate of 2%-3% in 2023 (Cox, 2023) [23]. Using a base of the average home price

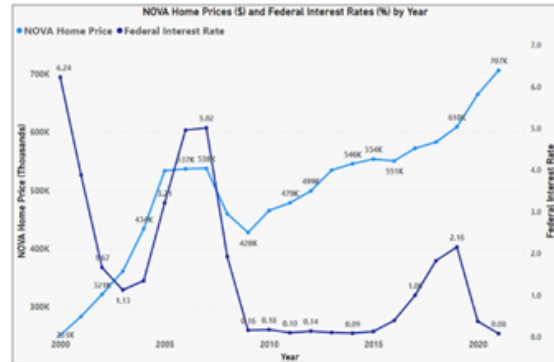


Fig. 1: Nova Home Prices and Federal Interest Rate By Year

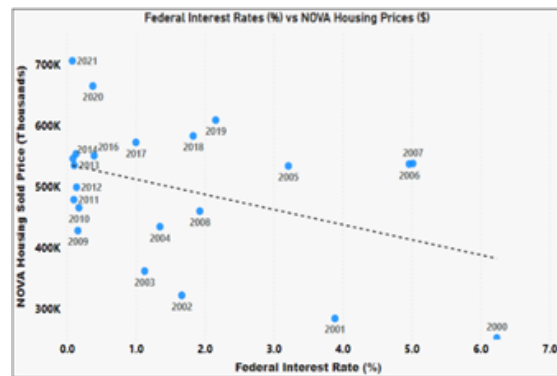


Fig. 2: Federal Interest Rates and Nova Home Prices

of \$707,990 in 2021 at an average interest rate of .08%, we can estimate three potential average home price values for NOVA in 2023 with our simple regression model based on a low, medium, and high interest estimate. (Table III.) Note this estimate does not account for error rates and it is assuming no other factors aside from interest rates affecting home prices from 2000-2021 in NOVA.

In our findings, we estimate home prices should drop significantly as interest rates increase in 2023.

This prediction is contingent on the behavior of interest rates impact on home prices reverting to historical norms aside from the abnormalities of the

TABLE III: Interest Rate predicted effect on NOVA Home prices in 2023

Interest Rate Effect on NOVA Home Prices in 2023				
Scenario	2021 Average	2023 Low	2023 Medium	2023 High
Interest Rate Estimate	0.08%	2%	2.5%	3%
Home Price Estimate	\$706,990	\$657,290	\$644,865	\$632,440

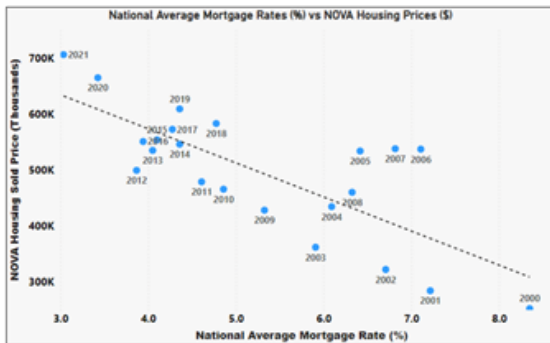


Fig. 3: NOVA Home Prices and National Average Mortgage Rate by Year

previous two decades. However, we do note that the relationship between interest rates and home prices is not the only factor to account for when predicting the price of home, hence we are implementing descriptive analysis in our findings.

B. Home Price Correlation With National Average Contract Mortgage Rate

House price has also held a strong correlation with National Average Contract Mortgage Rate (NACMR) over our collected data timeframe. The NACMR is an index derived from the Monthly Interest Rate Survey (MIRS) [24] which is conducted by the Federal Housing Finance Agency (FHFA) to describe mortgage loans through average interest rates, average loan maturity, average loan amount, and monthly average loan-to-value ratio (Monthly Interest Rate Survey, n.d.). However, MIRS is a lagging indicator, as the data is based on loan data from the two to three prior months, along with a month of processing for publishing by the FHFA (Monthly Interest Rate Survey, n.d.). As such, this metric may be less viable for future home predictions, whereas Federal Interest Rate changes may be announced in advance.

C. Exploring Active Northern Virginia Home Listings

We began by exploring the distribution of home listings across the NOVA region by county. The count was normalized by the 2022 population numbers (Population of Counties in Virginia (2022)). 43% of the listings in the dataset are from Fairfax County but also 46% of the population in the region

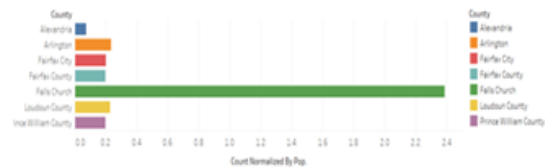


Fig. 4: Count of 2023 Home Listings Per County Normalized By 2022 Population

is in Fairfax county. Therefore, after normalizing the count of listings by population, Fall Church stands out as a county which is experiencing the most listings per capita (Figure #). Falls Church accounts for 6.8% of the listings but only contains 0.6% of the region’s population. This analysis indicates that Falls Church is experiencing the most real estate market activity per capita in the region.

The distribution of average listing home prices in NOVA counties varies and has a right skewed distribution. Due to the non-normal distribution, median was chosen over mean. The median home listing price in NOVA is \$1,052,690. Fairfax County, Arlington, and Falls Church median home listing prices are more than the overall median for the region while Alexandria, Fairfax City, Falls Church, Loudoun County, and Prince William median home listing prices are less than the overall median for the region (Figure #). Arlington and Fairfax county have large distribution of prices with standard deviations of \$824,270 and \$895,291 respectively. Prince William County and Fairfax City have small distribution of prices with standard deviations of \$215,037 and \$141,163 respectively. Loudoun County contains four outliers and Fairfax County has one (Figure 4).

46% of the listings in Falls Church are new constructions while only 1% of listings in Alexandria are new constructions. Homes in Fairfax County homes are on average the largest with an average of 4.97 bathrooms, 5.01 bedrooms, and 4,806 sq ft living area. Alexandria homes are on average the smallest with an average of 2.9 bathrooms, 3.6 bedrooms, and 2,146 sq ft living area. Loudoun County has the largest lot sizes on average at 26,694 square feet while Arlington has the smallest average lot sizes of 7,983 square feet (Figure 6).

Some characteristics of a home have more in-

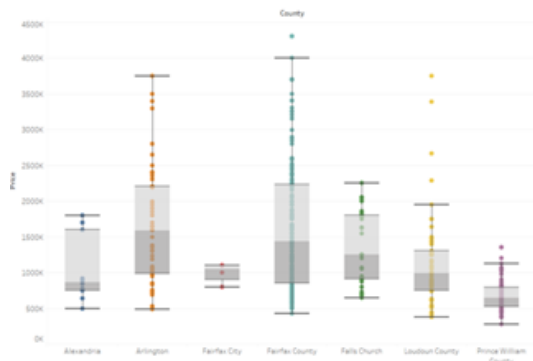


Fig. 5: Median Northern Virginia Zillow Home Prices By County

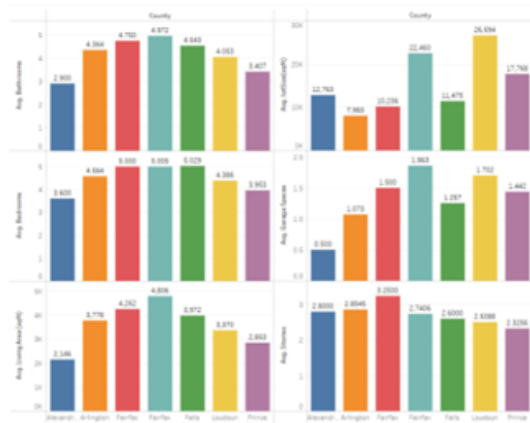


Fig. 6: Average Attribute Values For Home Listing In Each County

fluence on the price of the home than others. In this section we will explore this idea by performing correlation testing between home attributes and their listing prices on Zillow. The attribute with the strongest correlation to the price of a home is the number of bathrooms followed by the living area space, number of bedrooms, etc.. The correlation coefficient between the attributes and the price of the home were calculated using R studio code. Appropriate statistical testing methods were applied depending on the attribute data type. Ordinal data types were tested using the Spearman’s method while numeric and boolean data types were tested using the Pearson’s method [25]–[36].

TABLE IV: Attributes Correlation Coefficients to Price of Home

[HTML]EFEFEF Attribute	Statistical Method	Correlation Coefficient
Bathrooms	Spearman	0.805
Living Area	Pearson	0.718
Bedrooms	Spearman	0.646
Garage Spaces	Spearman	0.617
Lot Size (sqft)	Pearson	0.291
Has Garage	Pearson	0.267
New Construction	Pearson	0.162

REFERENCES

- [1] J. Dickler. Home prices are now rising much faster than incomes, studies show. (November 10, 2021). [Online]. Available: <https://www.cnbc.com/2021/11/10/home-prices-are-now-rising-much-faster-than-incomes-studies-show.html>
- [2] M. M. Kashino. Dc-area home prices hit new records in october. (November 11, 2021) Retrieved March 23, 2022. [Online]. Available: <https://www.washingtonian.com/2021/11/11/dc-area-home-prices-hit-new-records-in-october/>
- [3] A. Picchi. For most americans, owning a home is now a distant dream. (FEBRUARY 22, 2022) Retrieved March 21, 2022. [Online]. Available: <https://www.cbsnews.com/news/real-estate-home-prices-middle-class-affordability-2022-02-23/>
- [4] L. Stahl. Would-be home buyers may be forced to rent the american dream, rather than buy it. (March 20, 2022) Retrieved March 23, 2022. [Online]. Available: <https://www.cbsnews.com/news/rising-rent-prices-60-minutes-2022-03-20/>
- [5] G. Russell. 10 important features to consider when buying a house. (May 30, 2021) Retrieved April 15, 2022. [Online]. Available: <https://homeia.com/10-important-features-to-consider-when-buying-a-house/>
- [6] J. Nguyen. 4 key factors that drive the real estate market. (December 28, 2010). [Online]. Available: <https://www.investopedia.com/articles/mortgages-real-estate/11/factors-affecting-real-estate-market.asp>
- [7] M. Pažický and J. Falath. The big risk now for the us is not hyperinflation, but long-term elevated inflation rates. (January 7, 2022). [Online]. Available: <https://blogs.lse.ac.uk/usappblog/2022/01/07/the-big-risk-now-for-the-us-is-not-hyperinflation-but-long-term-elevated-inflation-rates/>
- [8] Zillow Research. Zillow’s hot housing takes for 2022. (8 December 8, 2021). [Online]. Available: <https://www.zillow.com/research/zillow-2022-housing-predictions-30394/>
- [9] Freddie Mac. Quarterly forecast: Strong housing market expected to persist notwithstanding rising mortgage rates and continued high home prices. (October 15, 2021) Accessed April 2, 2022. [Online]. Available: <https://www.freddiemac.com/research/forecast/20211015-quarterly-economic-forecast>
- [10] Redfin Real Estate News. Housing market news. (Accessed April 3, 2022). [Online]. Available: <https://www.redfin.com/news/bidding-wars/>

- [11] L. Lambert. Home price growth in 2022 to be the slowest in a decade, says realtor.com. (December 9, 2021). [Online]. Available: <https://fortune.com/2021/12/09/home-price-growth-2022-slowest-in-decade-realtor-com/>
- [12] Goldman Sachs. The housing shortage: Prices, rents, and deregulation (walker). (n.d.) Accessed April 2, 2022. [Online]. Available: <https://www.gspublishing.com/content/research/en/reports/2021/10/11/4e0070a-d902-4fbc-82e1-b53a053e06c1.html>
- [13] Board of governors of the Federal Reserve System. Federal reserve board - home. (n.d.). Retrieved March 23, 2022. [Online]. Available: <https://www.federalreserve.gov/>
- [14] L. Lambert. Home prices are set to soar in 2022, predicts zillow. (December 20, 2021). [Online]. Available: <https://fortune.com/2021/12/20/home-prices-set-to-soar-2022-predicts-zillow/>
- [15] J. Conner. Hud pd&r housing market profiles. (February 1, 2021). [Online]. Available: <https://www.huduser.gov/portal/periodicals/USHMC/reg/RichmondVA/HMP-April20.pdf>
- [16] I. Pardoe, "Modeling home prices using realtor data," *Journal of Statistics Education*, vol. 16, no. 2, 2008.
- [17] T. Clower. 2021 northern virginia association of realtors regional market forecast. (March 25, 2021). [Online]. Available: <https://www.nvar.com/realtors/news/real-estate-news/real-estate-news/real-estate-news/2021/03/25/2021-northern-virginia-association-of-realtors-regional-market-forecast>
- [18] R. Spensieri. 2021 was a record-breaking year for virginia's housing market. (January 21, 2022). [Online]. Available: <https://virginiarealtors.org/2022/01/20/2021-was-a-record-breaking-year-for-virginias-housing-market/>
- [19] Historical monthly data. (n.d.). Retrieved February 22, 2022. [Online]. Available: <https://www.nvar.com/realtors/news/market-statistics/historical-monthly-data>
- [20] Federal Housing Finance Agency. National mortgage database (nmdb) aggregate data. (n.d.). Retrieved March 23, 2022. [Online]. Available: <https://www.fhfa.gov/DataTools/Downloads/Pages/National-Mortgage-Database-Aggregate-Data.aspx>
- [21] U.S. Bureau of Labor Statistics. Bls data finder 1.1. (n.d.) Retrieved April 13, 2022. [Online]. Available: <https://beta.bls.gov/dataQuery/search>
- [22] G. D. Sutton, D. Mihaljek, and A. Subelyte, "Interest rates and house prices in the united states and around the world," 2017, retrieved April 12, 2022. [Online]. Available: <https://www.bis.org/publ/work665.htm>
- [23] J. Cox. Federal reserve approves first interest rate hike in more than three years, sees six more ahead. (March 16, 2022) Retrieved April 12, 2022. [Online]. Available: <https://www.cnbc.com/2022/03/16/federal-reserve-meeting.html>
- [24] Federal Housing Finance Agency. Monthly interest rate data. mirs transition index. (n.d.). Retrieved March 22, 2022. [Online]. Available: <https://www.fhfa.gov/DataTools/Downloads/Pages/Monthly-Interest-Rate-Data.aspx>
- [25] NVAHA: Northern Virginia Affordable Housing Alliance. Building northern virginia's future:policies to create a more affordable, equitable housing supply. (January 2019). [Online]. Available: https://nvaha.org/wp-content/uploads/NVAH001_1901_SupplyPapers-MAIN-FinalWeb-4.pdf
- [26] K. E. Case, "The central role of home prices in the current financial crisis: how will the market clear?" *Brookings Papers on Economic Activity*, pp. 161–193, 2008. [Online]. Available: https://www.brookings.edu/wp-content/uploads/2016/07/2008b_bpea_case.pdf
- [27] K. E. Case and J. H. Shiller. Rents rise most in 30 years, signaling more pain for americans. (March 10, 2022). [Online]. Available: <https://www.bloomberg.com/news/articles/2022-03-10/rents-rise-most-in-30-years-signaling-more-pain-for-americans>
- [28] P. Gelain and K. J. Lansing, "House prices, expectations, and time-varying fundamentals," *Journal of Empirical Finance*, vol. 29, pp. 3–25, 2014. [Online]. Available: <https://www.frbsf.org/wp-content/uploads/sites/4/wp2013-03.pdf>
- [29] G. Guilford. Will inflation fall? any pullback depends on these sectors. (March 7, 2022). [Online]. Available: <https://www.wsj.com/articles/inflation-high-forecast-economist-goodhart-cpi-11646837755>
- [30] T. Nicholas and A. Scherbina, "Real estate prices during the roaring twenties and the great depression," *Real Estate Economics*, vol. 41, no. 2, pp. 278–309, 2013. [Online]. Available: https://www.hbs.edu/ris/Publication%20Files/Anna_tom_59f6af5f-72f2-4a72-9ffa-c604d236cc98.pdf
- [31] N. Mansur. Northern virginia housing market forecast 2022. (January 10, 2022). [Online]. Available: <https://www.mashvisor.com/blog/northern-virginia-housing-market-forecast-2022/>
- [32] F. Mari. Will real estate ever be normal again? (November 12, 2021). [Online]. Available: <https://www.nytimes.com/2021/11/12/magazine/real-estate-pandemic.html>
- [33] Long & Foster - Real Estate Market Minute. Northern virginia housing market data. (March, 2022) Retrieved April 17, 2022. [Online]. Available: <https://marketminute.longandfoster.com/market-minute/va/northern-virginia.htm>
- [34] D. Olick. Mortgage rates are surging faster than expected, prompting economists to lower their home sales forecasts. (March 20, 2022). [Online]. Available: <https://www.cnbc.com/2022/03/22/mortgage-rates-are-surging-faster-than-expected-prompting-economists-to-lower-their-home-sales-forecasts.html>
- [35] W. Parker. Rent-control measures are back as home rents reach new highs. (March 13, 2022). [Online]. Available: <https://www.wsj.com/articles/rent-control-measures-are-back-as-home-rents-reach-new-highs-11647180001>
- [36] World Population Review. Population of counties in virginia (2022). (n.d.) Retrieved April 15, 2022. [Online]. Available: <https://worldpopulationreview.com/us-counties/states/va>

Preliminary Results on Analyzing Credit Card Fraud Detection

Abhishek Godavarthi
agodavar@gmu.edu

Raghad Almutairi
ralmuta@gmu.edu

Arthi Reddy Kotha
akotha2@gmu.edu

Abstract—Credit card use is not always the best way to use for payments, but the most demonstrable payment mode is through the credit card for both offline as well as for online payments, which can result in deficit of funds. As the online shopping is booming it helps in rendering the cashless payment modes. It can be used at shopping's, paying rent, paying utilities bill, internet bill, travel and transportation, entertainment, food. Using for all these things there is a chance of fraud transactions for a credit card, hence there is more risk. There are many types of fraudulent detections most of the banks and institutions are preferring fraud detection applications. It has become very hard to find out the fraud detections, After the transaction is done there is a chance of detecting fraudulent transactions in the manual business processing system. In real time the bunco transactions are done with real transactions, but it seems not to be sufficient for detecting [1]. Machine learning and data science both are playing a very important role in identifying the fraud detections. This study uses data science and machine learning for detecting the fraud detection to demonstrate various modellings. The problem enables the transactions of the previously done transaction data.

Index Terms—Credit card, Banking Services, Fraud detection, Cashless Payments

The dataset has been collected from a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. For statistical and visual analytics - Python has been used. The results would be visualized using various methodologies through these tools. The original dataset will be cleaned if necessary for accurate analysis of the data.

I. INTRODUCTION

Since 1950, when credit cards were first found and until this day, fraudsters have been competing to devise new ways to steal personal informa-

tion and reach people's credit cards and obtain money. With the popularity of e-commerce, most companies and institutions offer their products and services through websites that require consumers to create accounts and add personal information, including credit card information. Furthermore, unsecured websites may expose consumers to credit card fraud or identity fraud. Initially there was a two step verification method. Due to this users were effected of transaction fraud while shopping online. Consequently, companies and organizations with unsecured or less secured websites eventually lose their customers as customers will not trust sharing sensitive information with an unsecured site. Moreover, banks also play an essential role in keeping customers' personal information and transactions secured which is why banks can also lose customers if the level of security is not good enough. It shows the common terms and highlights the key statistics of credit card fraud detection. The proposals made in this paper in terms of savings and the time efficiency are the beneficial attributes. The techniques which are used in this paper reduces the credit card fraud but there is also a chance that it misguides the people to be misclassified as fraudulent. Sometimes there was some ethics in the banking and the complexity and efficiency obtaining the money. It deals with cases which involves the criminal cases which could be difficult in identifying sometimes.

Types of credit card fraud detections techniques, which helps to detect the transactions, the first one is fraud analysis which is used to analyze the fraud transactions of credit card and the second one is anomaly detection which is used to detect the anomalies based upon the previous data of transactions [2].

To detect credit card fraud using machine learning is to use and analyze data to investigate the habits and methods of fraudsters and build a model that helps detect and reduce fraudulent transactions [2]. To create a model to detect fraud, the data science team will need to collect the credit card users' data, such as the habit patterns, area, product types, amount, and spending, and then use the data and information to discover the fraudsters behavioral patterns [2]. The different techniques used by various companies for detecting the fraud transactions are Artificial Neural Networks, Hidden Markov Model, Naïve bayes, KNN classifier, Genetic Algorithm [2].

II. PROBLEM STATEMENT/OBJECTIVES

The increase in credit card fraud leads card users to lose money and companies to lose consumers. Credit card fraud directly affects consumer spending, which can also damage the economy. Every time consumer spending drops off, the prices drop, and the economic growth slows down [1].

In the United States, credit card fraud has significantly increased in the past few years as the number of reports went up from around 130K in 2017, to almost 400K report in 2020 [3]. Credit card fraud can happen in two different ways. The first way is stealing information or identity and opening a new bank account with someone's else's identity to be able to get their money while avoiding all responsibility. The second way is accessing an existing account. To succussed in credit card fraud, the fraudsters will have to access personal information such as full name, email address, ID number, credit card number and other sensitive information which can also leads to Identity theft.

Identity theft is another related issue as fraudsters steal people's personal information and use it to create a credit card with the stolen identity [4].

The goal of Identity thieves here is to steal money, and this can be a big issue as the number of people in debt will sharply increase causing more people to lose jobs, lose homes, and go to the prison which will also affect the economy as a result. The Covid-19's impact is one of the main reasons why Identity theft has become more common in the last two years. According to Daly (2021), about 1.4 million reports of identity theft were received in the

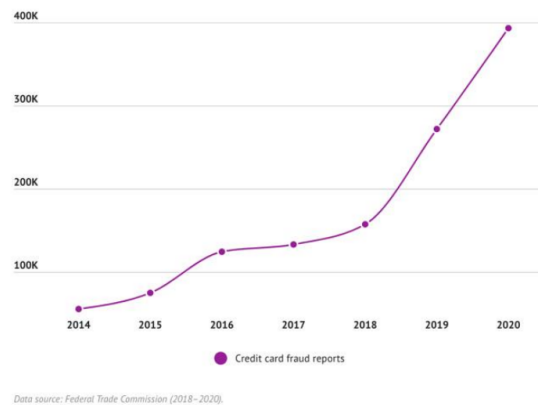


Fig. 1

United States in 2020, and a report from Aite Group stated that identity theft caused losses estimated at \$712.4 billion in 2020 [4]. In this research, we want to prevent this problem and reduce the losses as much as possible [3].

Another problem is credit card scams. These days scammers are creative and use many ways to scam people to steal their credit card information. Some scammers use phone calls to pretend to be bank employees, and others may send links to their victims, leading to a fake login page with the same original page design to steal people's login information [5]. Scammers use many ways that most people are not aware of, especially elderly and young people. Moreover, many hackers also use users' lack of awareness to steal and use their personal information.

Recognizing identity and credit card fraud transactions and identifying fraudsters' behavior is essential for every institution, including banks, to preserve the users' information and secure the institution's database and system. With the increase in the number of customers and technology development, creating a model that will help detect fraud transactions automatically and as accurately as possible is substantial.

- 1) How to process enormous data and how to build model that must be fast enough to respond to the scam in time.
- 2) Protecting the privacy of the user
- 3) How to deal with imbalanced data?
- 4) How to deal with misclassified Data?

- 5) Propose a conceptual framework from an almost anonymized real-world credit card dataset.

III. RATIONALE AND SIGNIFICANCE OF THE STUDY

Given the rapid digital transformation that businesses are witnessing, the need to improve credit card services and products' security is even more important and urgent. Such a need requires increasingly secure information systems in organizations [2]. Businesses need to ensure that their information technology infrastructure is upgraded with the latest that emerging and advanced technologies have to offer. To achieve this goal, a holistic understanding of the business is needed, as well as ongoing efforts to recognize any gaps and areas for improvement to safeguard the company's financial health. Because a substantial amount of financial dealings such as credit card processing are performed online, the transactional credit card data can help constantly learn from to detect anomalous activities.

IV. LITERATURE SURVEY

Numerous research projects have been published about credit card fraud detection since the 90s and until now. Basically, the fraud transactions are one of the illegal activities that are going on in today's world, which is also considered a crime nationally as well as internationally. As we discussed above it will check all the possible ways on who the fraud transaction is done and accordingly the steps are taken forward. Since 19th century we are facing problem in solving the probabilities or to handle it, but in 20th century they introduced a new algorithm and that is Bayesian algorithm which was used in the different ways to solve the issues which also called as frequentist statistics. There was also some deep learning proposal and topologies which was derived from the detections. The two types of random forests that are used for fraud transaction detection are normal and abnormal transactions. In the investigation there were several techniques used for fraud transactions and are also used for comparing the credit data. When presenting the crime data of the credit card there are certain issues faced by the data mining as well as non-data mining thefts [6].

The invoice bills and the payments are also issued to the investigator to investigate about the credit card detections, which is known as no cash mobile applications. There is a alarm raised when your transaction is confirmed as a fraudulent transaction or malicious [6].

Machine learning Based Approach to financial fraud detection is also done using the mobile application. Mobile payment fraud is identified in the growing issues of the credit card. Decision tree is also used for the fraudulent detection in the credit card system. It proposed the sampling process and the selection process for the feature use with the large amount of the transaction data which has high accuracy in the mobile payment. There are four modules that are been used and they are data collection, data preprocessing, feature extraction, evaluation model. Random forest selects the best feature for the data resulting that in the better model for the users [7]. With the advancements of the techniques with the machine learning it gives us the complete information and new techniques to learn. Credit card fraud detection using the machine learning is done by developing the new classifications and the regressions [6].

This is an exceptionally applicable issue that requests the consideration of networks, for example, AI and information science where the answer for this issue can be mechanized. This issue is especially difficult according to the point of view of learning, as it is portrayed by different factors like class unevenness. The quantity of legitimate exchanges far dwarf false ones. Additionally, the exchange designs frequently change their factual properties throughout the natural process of everything working out. Extortion goes about as the unlawful or criminal trickiness expected to bring about monetary or individual advantage. A purposeful demonstration is illegal, rule, or strategy with a mean to achieve unapproved monetary advantage. Various written works relating to irregularity or extortion recognition in this space have been distributed as of now and are accessible for public use [8].

Many models have been recommended for precise fraud detection. For instance, we can consider the brain network proposed by Ghosh and Reilly which is prepared on a large sample of named

Visa exchanges. These transactions contain an assortment of misrepresentation cases, for example, lost cards, stolen cards, application extortion, email misrepresentation and so on. Training on an assortment of information absolutely makes the model resistant to almost any sort of misrepresentation. Subsequently, the quality or assortment of data matters more than the amount or bulk. There were numerous different methodologies before and there are going to be numerous later and there is still a ton of scope for this field as the cheats keep on being unavoidable [8].

V. RELATED WORKS

Some research papers covered different methods for detecting financial fraud, and others included different learning techniques. These research papers were used to develop and improve fraud detection systems and algorithms. The previous work discussed two methods for machine learning: deep learning and traditional machine learning.

Deep learning is machine learning, but the main difference is the performance. Deep learning “uses a programmable neural network that enables machines to make accurate decisions without help from humans” (Grieve, 2020, para. 5) [9].

The deep learning method was studied in research by Thang, Tahir, Abdelrazek, and Babar in 2020 about credit card fraud detection [2]. In the study, deep learning has been used to solve complex problems. The research paper includes studying deep learning methods for credit card fraud detection issues and comparing this method’s performance with other machine learning algorithms on three different financial datasets. This research shows that deep learning methods had better performance than traditional machine learning models and that the results can be effectively used in the real world.

Another study by Zanin, Romance, Moral, and Criado in 2017 discussed how to detect credit card fraud through parenclitic network analysis. This study presented the first complex network classification algorithm to detect illegal cases in an actual card transaction dataset. This research shows how including features from the network data representation can improve the obtained result

by a standard neural network-based classification algorithm [10].

Modes of Fraud Detection usually used

The algorithms that are used for the model detections of credit card are as follows, they are Random Forest Classifier, Decision tree, Support Vector machine.

Random Forest Classifier: It is an inconsequential classifier among all the classifiers. It is an essential module of the classification. For the better evaluation and performance depends on several factors that includes depth, maximum bins, impurity [2].

Decision Tree: There is a test node which is represented with the root not and each node. Occurrences, Target Attribute, Attributes List are their elements used for planning tree which is called as leaf node [2].

Support Vector Machine: Based on the other algorithms this is best used for fraud detection. In this all the information or data is combined into one category and SVM is one of the form used to separate the data which was converted or separated by the different classes [2].

VI. DATASET REVIEW

The dataset has been collected from a research collaboration of Université Libre de Bruxelles on big data mining and fraud detection. This study uses a sample of data to detect fraudulent credit card transactions and prevent any credit card charges that take place without the card owner’s permission. Moreover, utilizing the data to build a model that is simple but fast enough to stop the fraud in time as well as report it to the relevant authorities. The dataset contains transactions made using credit cards by European cardholders in September 2013 and it shows that out of 284,807 transactions that were made in two days, there were 492 frauds and only 0.17% of the transactions were fraudulent. Unfortunately, the dataset does not provide more background about the data, but we will try to search about more data in other datasets the future. The credit card detection model we will be building will include Anomaly Detection which is the process of finding rare events, items, or actions that is considered suspicious to detect fraudsters. The anomaly detection will be applied by using machine learning

#	Column	Non-Null Count	Dtype
0	Time	284807 non-null	float64
1	V1	284807 non-null	float64
2	V2	284807 non-null	float64
3	V3	284807 non-null	float64
4	V4	284807 non-null	float64
5	V5	284807 non-null	float64
6	V6	284807 non-null	float64
7	V7	284807 non-null	float64
8	V8	284807 non-null	float64
9	V9	284807 non-null	float64
10	V10	284807 non-null	float64
11	V11	284807 non-null	float64
12	V12	284807 non-null	float64
13	V13	284807 non-null	float64
14	V14	284807 non-null	float64
15	V15	284807 non-null	float64
16	V16	284807 non-null	float64
17	V17	284807 non-null	float64
18	V18	284807 non-null	float64
19	V19	284807 non-null	float64
20	V20	284807 non-null	float64
21	V21	284807 non-null	float64
22	V22	284807 non-null	float64
23	V23	284807 non-null	float64
24	V24	284807 non-null	float64
25	V25	284807 non-null	float64
26	V26	284807 non-null	float64
27	V27	284807 non-null	float64
28	V28	284807 non-null	float64
29	Amount	284807 non-null	float64
30	Class	284807 non-null	int64

Fig. 2

where both normal and anomalous samples will be included to make the model predict future actions. We want the model to be trained by using reports and feedbacks to find the probability of fraud to give an alert [11].

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, they cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise [11].

VII. PROPOSED APPROACH

The data obtained from the dataset will be processed and analyzed properly, cleaned, and visual-

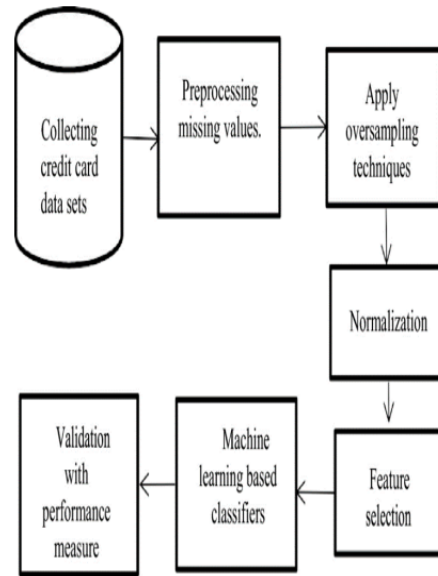


Fig. 3

ized. We will use the data later to try and build several machine learning models in order to find the best solution.

As shown in Figure 3, these are the steps to follow to accomplish the model.

After collecting, cleaning, and visualizing our data, the next step is to start building the model. Because the process of developing a data model may require several stages of editing, changing, or updating data over time during the development, we will be building several data model versions and comparing different versions using data modeling tools until we find the model that works best for us. The process will start by building a prototype and using PyCaret, which is an open-source machine learning library to facilitate the comparison process. First, after installing the PyCaret library, all the necessary files will be imported, the dataset will be loaded, and the PyCaret classifications will be set up [12]. Next, the comparison process begins as we start creating the best possible model, saving, loading, finalizing, and deploying the model.

A. Data Preprocessing

Data preprocessing involves transforming the raw data into the well-formed data sets so that the analysis can be performed in a better way. It

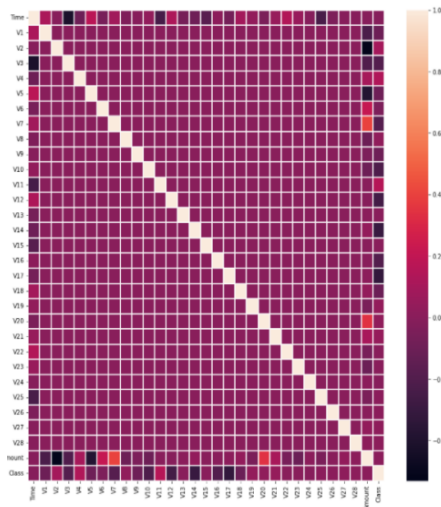


Fig. 4: Correlation Matrix

is used for both databases driven and rules-based applications normally it is a critical in some cases in the data processing when the dataset is large. Basically, the raw data which we have is normally incomplete and will be in an incontinent way.

Data preprocessing is done in steps that follows as formatting, cleaning, sampling. Formatting is used for putting the data in a way that it will be suitable for the work. The data and the formatting changes according to the needs of the people.

The most recommended or most used format is the CSV format. Data cleaning is the main and important procedure in the field of data science. It does a major part of the work. That is 80% of the work is done. Sampling is the technique for analyzing all the subsets with the whole large datasets, which gives us a better result for understanding and the pattern of the data in the integrated way.

Data is represented differently and are kept together with the help of data integration. When the data is large, or the dataset is huge then the databases can become very slow to work. To replace the data with the raw values within given intervals reduces the number of values by diving that in the given rage.

B. Correlation

There is no notable correlation between features V1- V28. There are certain correlations between

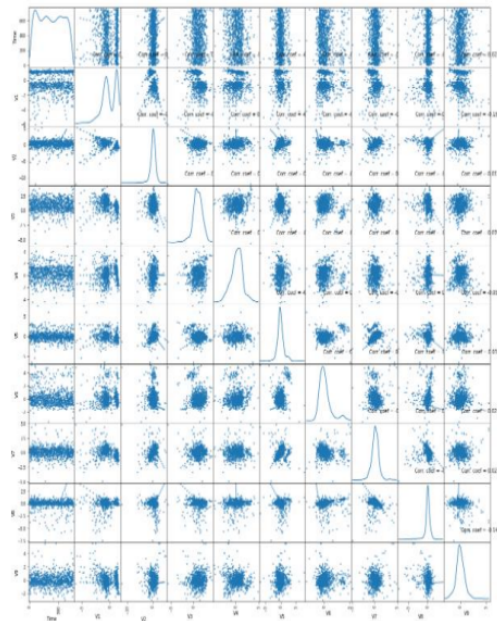


Fig. 5: Correlation - scatter and density plot

some of these features and Time (inverse correlation with V3). Features and Amount (direct correlation with V7 and V20, inverse correlation with V1 and V5).

There are no null values in the data.

C. Preliminary Results

Using the Random Forest method, a prototype that analyzes fraud identification was built. The Random Forest method is a good learning method that helps create an uncorrelated forest of trees with better prediction and more accurate results. The model's performance was evaluated on the test set. The results showed that the model scored a 0.88 f1-score, which is considered a good score for a good model that can detect fraud transactions. A PyCaret library was also used to compare the model with other versions to assure it is the best. Eventually, the results revealed that the first model we built will be able to detect fraud with a percentage of 90% accuracy.

citations [13], [14]

REFERENCES

[1] K. Amadeo. What you buy every day drives u.s. economic growth. the balance. (accessed February 23, 2022). [On-

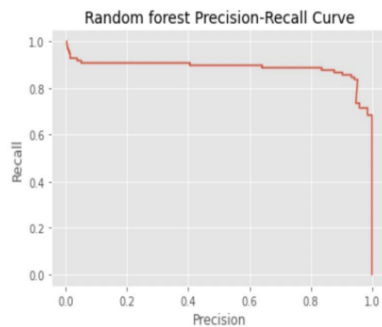


Fig. 6

Available: <https://www.geeksforgeeks.org/automating-the-machine-learning-pipeline-for-credit-card-fraud-detection/>

- [13] C. Oden. Design and implementation of a credit card fraud detection system. (accessed February 23, 2022)(n.d). [Online]. Available: <https://www.projecttopics.org/design-and-implementation-of-a-credit-card-fraud-detection-system.html>
- [14] techopedia. Data preprocessing. (accessed February 23, 2022). [Online]. Available: <https://www.techopedia.com/definition/14650/data-preprocessing>
- line]. Available: <https://www.thebalance.com/consumer-spending-definition-and-determinants-3305917>
- [2] T. T. Nguyen, H. Tahir, M. Abdelrazek, and A. Babar, "Deep learning methods for credit card fraud detection," *CoRR*, vol. abs/2012.03754, 2020. [Online]. Available: <https://arxiv.org/abs/2012.03754>
- [3] Saloni and M. Rout, "Analysis and comparison of credit card fraud detection using machine learning," in *Advances in Electronics, Communication and Computing*, P. K. Mallick, A. K. Bhoi, G.-S. Chae, and K. Kalita, Eds. Singapore: Springer Singapore, 2021, pp. 33–40.
- [4] L. Daly. Identity theft and credit card fraud statistics for 2021. (accessed February 23, 2022). [Online]. Available: <https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/>
- [5] B. Myers. Credit card fraud and scams: How to avoid both. (accessed February 23, 2022). [Online]. Available: <https://www.fool.com/the-ascent/credit-cards/scams-fraud-how-avoid/>
- [6] M. Thirunavukkarasu, A. Nimisha, and A. Jyothsna, "Credit card fraud detection through parenclitic network analysis," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 4, pp. 71–79, 2021.
- [7] T. Yiu. Understanding random forest. (accessed April 18, 2022). [Online]. Available: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
- [8] akanksha singh and A. Singh. Ieee-cis fraud detection. (accessed April 18, 2022). [Online]. Available: https://www.academia.edu/44187814/IEEE_CIS_
- [9] P. Grieve. Deep learning vs. machine learning: What's the difference? (accessed February 23, 2022). [Online]. Available: <https://www.fool.com/the-ascent/credit-cards/scams-fraud-how-avoid/>
- [10] M. Zanin, M. Romance, S. Moral, and R. Criado, "Credit card fraud detection through parenclitic network analysis," *CoRR*, vol. abs/1706.01953, 2017. [Online]. Available: <http://arxiv.org/abs/1706.01953>
- [11] GeeksforGeeks. ML — credit card fraud detection. (accessed February 23, 2022). [Online]. Available: <https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/>
- [12] @amankrsharma3. Automating the machine learning pipeline for credit card fraud detection. (accessed February 23, 2022). [Online].

Airbnb Data Analysis of Florida Real Estate

Geetna Penmasta
gpenmatsr@gmu.edu

Keerthana Vallamkonda
kvallamk@gmu.edu

AVenkata Sai Ravi Tej Adabala
vadabala@gmu.edu

Abstract—Since 2012, guests and hosts have used Airbnb to extend their travel options and provide more distinctive, personalized experiences. Airbnb is a one-of-a-kind service that is now known and used all over the world. The analysis of Airbnb’s millions of listings is critical to the company’s success. The millions of listings generate a lot of data, which can be analyzed and used for a variety of purposes, including security, business decisions, understanding customer and provider behavior and performance on the platform, guiding marketing initiatives, implementing innovative additional services, and so on.

The success of a firm is mainly determined by the loyalty of its clients. Data analysis aids in the understanding of the relationship between client loyalty, satisfaction, and image. This will be of great importance to both practitioners and scholars in the field of hotel management. The goal of this study is to discover the elements that lead to customer loyalty in the hotel business in Broward County, Florida.

The focus of this research paper is on the factors that influence the pricing of Airbnb rooms. By gathering data from several Airbnb resources, it is possible to determine how the Airbnb pricing is split throughout different neighborhood groups in Broward County, Florida, as well as the elements influencing the price, such as room type, rating, location, and service. To get structured data, associated data such as room service, covid policies, parking facilities, and complimentary breakfast is collected, cleaned, filtered, and modified.

Index Terms—Airbnb, prices, reviews, NLP, Data Analysis, Python, R

I. INTRODUCTION

Travel costs range from person to person, and everyone has distinct financial constraints. When planning a trip, one of the most important constraints to consider is lodging. We may look at the aspects that are crucial in increasing Airbnb’s income by performing big data analysis on its data. At first, revenue was

determined primarily by factors such as occupancy rate and average daily rates. Later, it was

discovered that these measures are insufficient for determining revenue production, and that more advanced measurements are required. As a result, to discover these measures, a thorough examination is required. Booking lead times only tell part of the picture when it comes to seasonality. You’ll need to look at the bigger picture of seasonality to properly master your vacation rental market using sophisticated Airbnb data research. Hosts can use the Demand Score to determine which days, weeks, and months should be priced higher and which should be priced lower by focusing on the Demand Score. Future Rates is the ideal place to go for hosts seeking for an all-in-one rental pricing solution to help them spot chances. Cancellation rules and minimum night stays are two aspects of property management that are sometimes disregarded. While some hosts choose to ignore these considerations, others strive to make the most of them. While aligning your listing’s settings with market norms is a wonderful method to improve your listing’s performance, you should also consider diverging from the conventions to make your listing stand out. Is a three-night minimum stay required by most hosts? Set yours for one or two nights to make yourself more accessible to more travelers.[12].

II. RELATED WORK

Airbnb is a website that connects people who wish to rent out their homes with people who are looking for lodging in specific areas. It presently includes over 5.6 million listings, encompassing over 100,000 cities and towns in over 200 countries throughout the world. Airbnb has revolutionized the travel industry by allowing homeowners and business owners to post their spare bedrooms, unoccupied hotel rooms, or entire houses for short-term rentals on the platform. In many urban areas, this disruption has resulted in housing shortages [1].

This analysis aimed to uncover intriguing insights that could be useful to a traveler, a host, an Airbnb decision maker, or even a D.C. housing regulator. Some of major insights include the cost of listings in relation to their proximity to the National Mall [1]. For those unfamiliar with Washington, DC, the National Mall is a grassy stretch of land bordered by many Smithsonian museums, the Capitol building, and the White House. Essentially, anyone staying in an Airbnb in DC has a good probability of seeing the National Mall at some time during their stay [1]. The data on Airbnb is updated on a regular basis. Some of the information was gathered from 2015 until the present day.

This project's datasets were collected in September 2019. Listings, Neighborhoods, Reviews, and Calendar data for the city are used. Each dataset comprised distinct data. statistics on all of D.C.'s Airbnb listings - listing ID, latitude, longitude, host, room type, price, number of reviews per listing, reviews per month, listings per host, neighborhood, availability for booking, and number of listings per host are all present in these listings.

III. BACKGROUND

The neighbourhood data comprises the names of 39 neighborhoods in Washington, D.C., as well as a JSON file containing the geographic coordinates of the neighborhood's outline. Over 358,000 reviews were written between January 2009 and September 2019 according to the Reviews statistics. We can cross-reference which listing they apply to because they all relate to a listing ID from the Listings dataset.

Finally, the Calendar dataset contains a yearly 'calendar' of pricing and listings for each day. Since DC is a popular tourist destination for visitors from all over the world as well as within the United States, they checked if the volume of visitors was related to some of the city's major events. Every year, 30,000 runners and their family and friends participate in the Marine Corps Marathon, which starts in Arlington, VA and passes through several of DC's landmarks. There are many other events, and organizations that attract visitors on a regular basis. They visualized the flow of tourists in relation to specific events in the Airbnb data. The inauguration of a new president always attracts

a large crowd. Supporters frequently organize a March on Washington for causes and movements that they want Congress to pay attention to and pass new laws for. Two of DC's professional sports teams have won national championships in the last two years [1]. Some of the findings include: What are the broad trends in the data? What is the relationship between Airbnb data and key events in the nation's capital during the previous 10 years? What can the data tell us about the impact of Airbnb on the housing market in Washington, D.C.?

IV. PROBLEM DESCRIPTION

Choosing a right location and making sure it is worth of, is very important while choosing our vacation spot (Airbnb). Any review a guest reads will reflect the actual experience another member of the Airbnb community had. Reviews include a rating system to quantify your performance as a host across both general and specific criteria. So, we are curious to find out this relationship between the reviews and prices of the bookings. What factors are making the people to choose Airbnb? Factors influencing the Airbnb prices and what are the quantifying parameters we should consider seriously while fixing the prices, can be detected by analyzing the data. Is it possible that these price values are depending only on the known factors like location, room type etc., or are there any hidden factors that are influencing these price values and therefore changing our decisions like cancellation policies that we could be missing? And how much weightage do these factors carry? Can they be ignored? Can they be taken seriously for the future success? How does the concept of Airbnb initially received and used by the people over the years (2022-2012) and what factors were improved to make it successfully?

Is it possible to predict the Airbnb price by using its features and how accurately can we build this model? Is it possible to do sentiment analysis of the reviews and find what keywords define low rated Airbnb's?

With these findings and insights, we can find where the potential value of Airbnb lies and what would be the possible changes in the Airbnb structure that can improve its value. Predicting Airbnb prices and identifying the most impactful features

of the price value can be helpful for the highlights of marketing and also provides a scope of improvement.

V. PROPOSED APPROACH

By gathering data from several Airbnb resources, it is possible to determine how the Airbnb pricing is split throughout different neighborhood groups in Broward County, Florida, as well as the elements influencing the price, such as room type, rating, location, and service. To get structured data, associated data such as room service, covid policies, parking facilities, and complimentary breakfast is collected, cleaned, filtered, and modified. This unstructured data must be transformed into structured data before it can be adequately analyzed, and insights extracted. The four iterative phases of deleting undesirable observations, addressing structural flaws, controlling undesired outliers, and handling missing data are used to collect and clean factors affecting the room pricing in Florida.

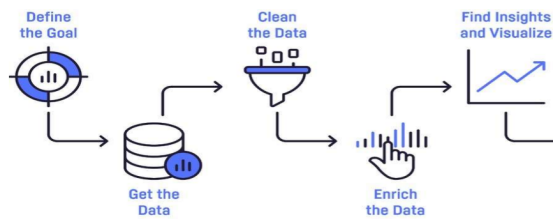


Fig. 1: Various Stages involved in the project

After cleaning the data, we can go on to exploratory data analysis with Python and RStudio. The data's linkages, patterns, and insights will be investigated and used to make decisions. To be useful, these relationships must be valid, not obvious, and human understood. Visualization technologies such as Power BI and Tableau are also used in exploratory data analysis to visually comprehend the correlations. Additionally, data modeling techniques such as regressions and classifications will be used to forecast the pricing of Airbnb in Broward County, Florida.

VI. DATA PREPROCESSING

Collected datasets needs to be prepared before we can apply any modeling techniques. Usually, raw data has inconsistent formatting, human errors,

and can also be incomplete. By preprocessing data, these issues can be resolved, and datasets can be analyzed more effectively and efficiently. Structured data is needed to apply any modeling technique. Data collected from different resources are merged and explored. Missing values and the data outliers were removed to avoid inaccurate and compromised results. The rows of the dataset were decreased by 1000 when the data cleaning techniques are applied to it. Used NumPy and pandas' libraries to clean the data. NumPy library is mainly used to explore the data and understand the domains of each attribute. Attributes such as cancellation policies and parking facilities of the Airbnb's were collected through various API's and open resources. The missing values have been filled with '*' values have been removed from the dataset. Attribute values greater than a threshold value have been considered as an outlier and have been removed from the dataset using panda's library in python. After making sure that the data quality attributes such as Accuracy, Consistency, Completeness, Validity, Timeliness are present, next step of analysis is proceeded.

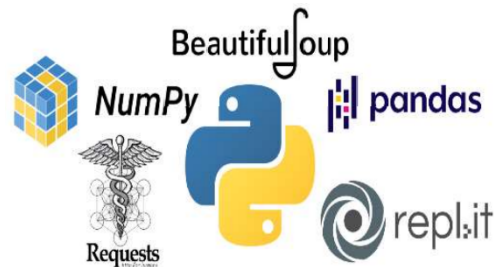


Fig. 2: Python libraries used for data preprocessing



Fig. 3: Data cleaning steps for achieving structured data

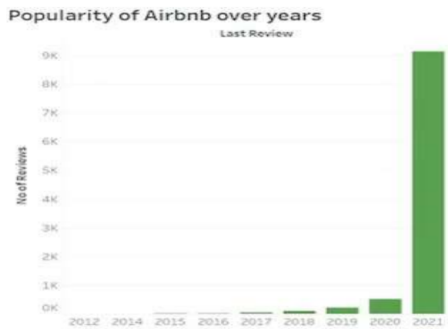


Fig. 4: Visualization 1: Number of Reviews per Year

VII. PRELIMINARY ANALYSIS RESULTS

Despite the fact that Airbnb was launched in 2008, it was only in 2012 that online vacation rental properties became popular. As a result, we've counted the number of reviews from 2012 to 2021. As can be seen from the graph, Airbnb's first two years were not particularly successful, with nearly no reviews in each of those years. People began to recognize and explore Airbnb in 2015, and the number of reviews gradually climbed throughout the years, reaching a peak of over 9000 reviews in the most recent year of 2021. Airbnb did not become popular overnight; in fact, it took nearly ten years from the time it was formed to gain popularity.

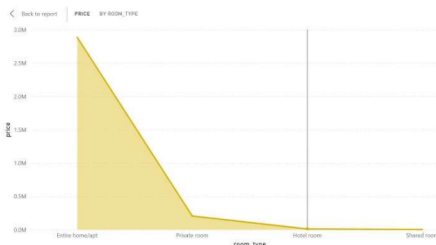


Fig. 5: Visualization 2: Room type Vs Price

Airbnb primarily offers four different types of lodging: entire home/apartment, private room, shared room, and hotel room. Shared and private room kinds are those in which the entire home, including bedrooms, or just the kitchen and patio are shared. The visualization shows that shared rooms are less popular among all four since they lack privacy. People prefer private rooms to hotel rooms, which may be due to the fact that hotel rooms are more expensive. Finally, in terms of



Fig. 6: Visualization 3: Neighborhood Vs Price range

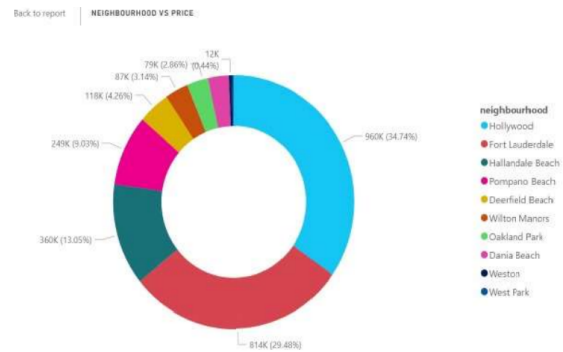


Fig. 7: Visualization 4: Price share of each Neighborhood

Airbnb rental homes, the entire room/apartment accommodation type is the most popular.

The Airbnb housings in the Florida region are the focus of this paper. The price ranges for each neighborhood in Florida where Airbnb homes are situated are shown in Visualization 3. The most expensive Airbnb properties are in the Hallandale Beach, Hollywood, and Lauderdale by the Sea neighborhoods. The most affordable areas are Pembroke Pines and Miramar.

The pricing share of each neighborhood in the Florida region is depicted in the doughnut chart above. The city of Hollywood accounted for over 38% of the total, with nearly 960K dollars, followed by Fort Lauderdale, which accounted for nearly 30% of the total, with 814K dollars. The areas of Dania Beach and Weston account for around 3% of

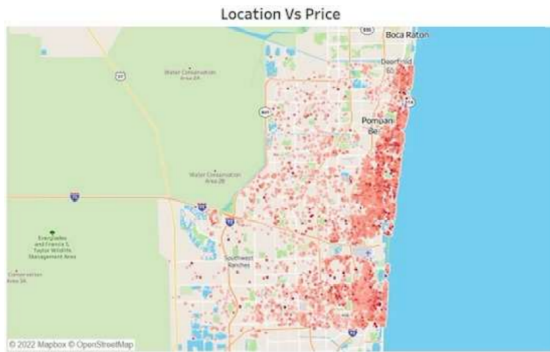


Fig. 8: Visualization 5: Location Vs Price



Fig. 9: Visualization6: Parking Facility Vs Cancellation Policy

the total proportion. Among Florida neighborhoods, West Park Zone has the lowest price share.

The above map chart depicts the pricing distribution of Airbnb rentals in various locations. The darker the hue, the higher the price, whereas the lighter the shade, the lower the price. From the map, we may deduce that coastal locations have denser hues and thus higher pricing, whilst off-the-coast places have lower prices. The increased prices in coastal counties may be owing to the tourist attractions, beaches, and varied activities accessible in these areas.

The price ranges for Airbnb listings with and without parking and cancellation policies are shown in the bar chart above. Rental properties with parking facilities are more expensive than those without. Customers are unconcerned about the cancellation policy while the parking facility is available. The cancellation policy has an impact on costs, but it isn't significant. Customers, on general, prefer

parking to cancellation policies.

VIII. FREQUENCY DISTRIBUTION ON REVIEWS USING NTLK

In this dataset we have reviews provided by the visitors regarding their stay in the Airbnb. All these reviews are used to use what are most of the user are talking about regarding their stay. The Natural Language Toolkit (NLTK) is a Python programming environment for working with human language data in statistical natural language processing (NLP). It includes tokenization, parsing, categorization, stemming, tagging, and semantic reasoning text processing packages. After filtering the reviews data and removing unnecessary words, NLTK used to first convert the sentences available in reviews to words as tokens. Then after removing the stop words, the frequency distribution is checked for the remaining words. From the below figure, we can see that most of the people who visited Airbnb in Florida are talking beach, pool, type of room they stayed in.

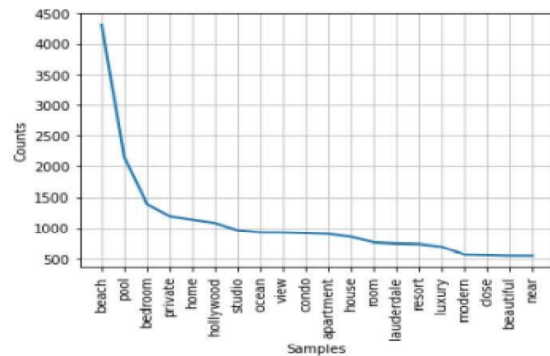


Fig. 10

citations [2]–[12].

REFERENCES

- [1] J. Owens. An analysis of airbnb in washington dc. (accessed May 14, 2020). [Online]. Available: <https://medium.com/@jessie.owens2/an-analysis-of-airbnb-in-washington-dc-8013cfef7379>
- [2] N. Lamba. Predicting airbnb prices in los angeles. (accessed March 20, 2022). [Online]. Available: <https://medium.com/analytics-vidhya/predicting-airbnb-prices-in-los-angeles-14758afc47e>
- [3] P. Huilgol. Accuracy vs. f1-score. (accessed March 20, 2022). [Online]. Available: <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2>

- [4] H. Jalali. Airbnb price prediction linear regression. (accessed April 29, 2019). [Online]. Available: <https://github.com/hemangjalali/AirBnB-Price-Prediction-Linear-Regression->
- [5] EHL Insights. What motivates travelers to book airbnb? (August 1, 2016). [Online]. Available: <https://hospitalityinsights.ehl.edu/travelers-airbnb-study>
- [6] Padlifter. The importance of good reviews. (January 1, 2020). [Online]. Available: <https://padlifter.com/free-tips-and-resources/reviews-and-credibility/the-importance-of-good-reviews-on-airbnb/>
- [7] D. Taylor. Power bi tutorial: What is power bi? why use? dax examples. (accessed March 20, 2022). [Online]. Available: <https://www.guru99.com/power-bi-tutorial.html>
- [8] HEVO. Top 10 best power bi dashboard examples in 2022. (accessed March 20, 2022). [Online]. Available: <https://hevodata.com/learn/top-10-best-power-bi-dashboard-examples-in-2021/>
- [9] Data Crunch. Best tableau dashboard examples for better dashboard layouts. (accessed March 20, 2022). [Online]. Available: <https://datacrunchcorp.com/tableau-dashboard-examples/>
- [10] K. Bernard. Tableau dashboard examples. (accessed March 20, 2022). [Online]. Available: <https://www.rigordatasolutions.com/post/tableau-dashboard-examples>
- [11] T. Shin. All machine learning models explained in 6 minutes. (accessed March 5, 2022). [Online]. Available: <https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a>
- [12] Linguamatics. What is text mining, text analytics and natural language processing? (accessed March 20, 2022). [Online]. Available: <https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing>

Taslina Akter
Dept. of EEE
Independent University, Bangladesh
Dhaka, Bangladesh
2222796@iub.edu.bd

Yeaser Sadman
Dept. of EEE,
Independent University, Bangladesh
Dhaka, Bangladesh
1522117@iub.edu.bd

Shatabdee Bala
Dept of CSE,
Gono Bishwabidyalay
Dhaka, Bangladesh
balashatabdee@gmail.com

Abstract— In this article, we generated some data as a result of new technologies, the internet, and linked items. Putting these facts into context and structuring them so that they may be perceived, understood, and reflected is critical. Humans had traditionally studied data. As the availability of data grows larger, humans are progressively turning to computerized technologies that can replicate them. Machine learning refers to technologies that can resolve issues by learning from both data and data modifications. Artificial intelligence (AI) has a significant influence on e-learning studies, and machine learning-based methodologies may be used to improve Technology Enhanced Learning Environments (TELEs). This publication provides an outline of relevant study outcomes in this domain. Initially, we'll go over some basic machine learning ideas. Then, we'll go through the significant latest research in the domain of e-learning that uses machine learning.

Keywords— *E-Learning, Technology Enhanced Learning Environments, Machine Learning, Artificial Intelligence*

I. INTRODUCTION

Nowadays we lead our lives on a digital track that expresses our performance, identifies our situation, and comes up with many other facts about what we imagine, what we get, etc. Most of the gadgets, machines, and whatever we use that create data because of both data capacity and societal digitization as an example- bring out details from anywhere like smartphones, videos, audios, etc It is vital to profit and makes sense of all obtained data [1].

Understanding events, modeling behaviors and making predictions are all feasible with data analysis. Humans used to evaluate data, write algorithms and then have the computer use those methods to solve issues. Humans now provide data that allows the computer to learn on its own without having to be expressly programmed. Representation, evaluation, and optimization- these three make a machine learning model. So, machine learning works in this concept [2].

In reality, individuals see the worth of information and its true capacity for wealth. AI approaches for dissecting convoluted information have arisen as a critical region in an assortment of logical review disciplines including clinical, interest business, industry, instruction, interpersonal, organization, financial aspects, money, etc.

Machine learning has a close relationship to many related fields including AI, data mining, statistics, and other listed shortly. Machine learning is a many-sided area that is connected to all these fields in various ways. Data mining uses statistics to bring out unseen facts from raw data. Deep learning is the key technology of both AI and machine learning.

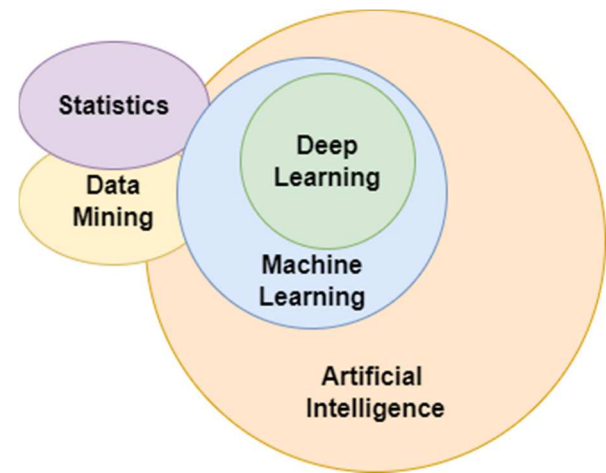


Figure 1: Machine Learning related fields

II. MACHINE LEARNING

The machine learning process is the first part of the ML process which involves gathering data from many roots and fine-tuning it, this data will be utilized for ML algorithms depending on issues such as prediction, classification, and other models that are accessible in the ML world. The machine learning process consists of seven stages which are described to come.

Data collection is the first step. It's a process where collecting and computing information from countless different sources. It's a vital task since it will define how effective the predictive model will be. But gathering information that is acquired is much of the time unstructured, has noise, or should be changed into various organizations to be useful for machine learning. So, the information needs to be cleaned and pre-handled [3].

Machine learning model building is next. It's working by gathering and summing up from preparing data, then applying that obtained data to new information it has never seen before to make forecasts and satisfy its motivation. To do so, we first do feature engineering in which we identify the most important data features and then strive to find the optimal machine learning algorithm for the situation at hand. It is critical to achieving the greatest potential outcomes [4].

The next door is training. In this stage, we progressively use a portion of our information to increase machine language prediction skills. After training, it's time to test and evaluate how it performs against the data that has yet to be seen. For evaluating performance used accuracy, precision, recall, etc various metrics. Sometimes it is feasible to go back and enhance training before retesting. The output of machine learning is the final phase. It might be a guess or an interface.

A. Machine Learning Paradigm

Machine learning helps to improve the e-learning framework by including the ability to place learners into those learning frameworks, as well as misleading and smart learning initiatives have been implemented [5].

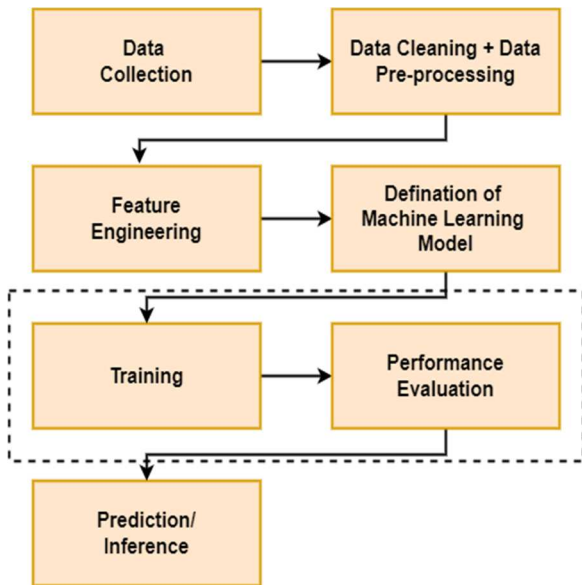


Figure 2: Machine Learning Process

Machine learning can be classified according to the approach used for the learning process. Four main categories identified-supervised, unsupervised, semi-supervised, and reinforcement learning.

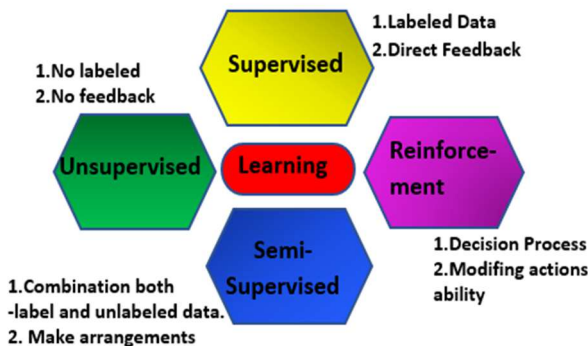


Figure 3: Machine learning Paradigms Categories [6]

In supervised learning, input information is called information prepared and has the specified spam/non-spam or storage costs at once. Place the model through the preparatory process it is obliged to produce predict and adjust if these are not met. The preparation methods are followed until the model achieves the desired precision on the preparing data. The concern, in this case, is grouping and relapse. Logistic Regression and the Back Propagation Neural network are used to calculate case values [7].

Conversely, Unsupervised learning is a machine learning technique that doesn't require the user to monitor the model. The model can work on itself to discover the patterns and data that were not before discovered. This case mainly handled unlabeled data.

However, semi-supervised learning which input data is a combination of labeled and unlabeled samples. There is a desired prediction issue, but the model takes into account the structure to compile the facts and produce expediencies. The challenges, in this case, are arrangement and regression [8].

At the end of the machine learning viewpoint when you know what you want but don't have the idea to gain it. The principle is to test solutions and then see which can achieve the desired result. Reinforcement learning issues can be formalized as agents which make good decisions and have good behavior [9 - 14]. This means modifying or acquiring new actions and abilities. So, this case doesn't require knowledge and control, it only needs able to be interact with the environment and collect information.

III. APPLICATIONS OF ML IN E-LEARNING

Now, in this 21st century people are learning and trying to utilize their skills at different levels, by researching and experimenting, Personalized and adaptive machine learning can change learning content or the mode of delivery on the fly and provide real-time feedback to learners. In machine learning, the computer learns from statistics to do with the given task this data will be utilized for ML algorithm depending on issues such as prediction, and classification. The next step is to create a machine learning model. It works by gathering and combining information gathered from data preparation, then applying that information to fresh information it has never seen before to produce projections and satisfy its goal [9]. We apply a portion of our data to gradually improve the prediction skill of machine language. After it has been trained, it is time to put it to the test and see how it performs against previously unseen data. Various criteria including accuracy, precision, and recall were utilized to evaluate performance these parts are the main vital parts of the application Before retesting, sometimes it is possible to go back and improve training. The final phase is the machine learning output. And we have to focus on the output and have a look carefully at what we have given the machine to do besides what he gives us as an output, it will help us to summarise the application and its success rate.

A. Sentiment Analysis

Sentiment analysis can be used to predict learner satisfaction by identifying multiple complex emotions researchers intend to determine the polarity of learners' attitudes by positive and negative sentiments using forum messages in MOCs. It supervised machine learning algorithms. logistic regression, support vector machine, decision tree, random forest, and nave Bayes have been employed more frequently in contributions linked to prediction in MOOCs. The random forest approach was shown to be the most reliable. It is critical to comprehend the function of emotion in MOOC students learning experiences, on the other hand, regulation of achievement emotions according to [10 - 14] may help to promote learner engagement with a supervised machine learning model based on SVM to categorize achievement emotions automatically. SVM was chosen because it outperforms nave Bayes, logistic regression, and the decision tree in terms of performance.

B. Student Behavior

The application of machine learning in forecasting student conduct was the subject of an interesting literature study. Two research objectives have been identified: dropout prediction and student classification.

1) Student Classification

Personalities, histories, knowledge, abilities, and preferences undoubtedly play an important role in the machine learning process. The recommender system helps each learner find the most appropriate content. Learner profiling and classification are essential not only for personalizing learning but also for identifying abandonment causes and verify of other purposes. In table 1 we just came out with some recent machine learning-based student classification studies.

Table 1. Student Classification [10]

ML Algorithms	Classification Goal	Results
k-means, SVM, Naive Bayes	Classification of engaged and disengaged faces of students with dyslexia	Accuracy: 97–97.8%
Backpropagation, Support Vector Machine (SVM), Gradient-Boosting Classifier (GBC)	classification of student performance	Accuracy: BP = 87.78%, GBC= 82.44%
Decision tree, Logistic Regression, k-nn, SVM, Random Forest	Classification of successful and unsuccessful students	KNN gives the higher accuracy = 85%
K-modes, Clustering, Naive Bayes	Classification of learner’s learning style	Accuracy: 89%

2) Dropout Prediction

To access interacting behavior traces leftover TELE a variety of machine learning approaches were used. Logistic regression(LR) has been the most commonly utilized technique to predict student dropout in the MOOC setting according to [1] who focuses on learners and clickstream data percentage.SVM and decision tree ranked 2nd and 3rd respectively. The neural language processing technique is ranked 3rd.

C. Self-regulated Learning

The majority of TELE has little external instructor monitoring, learners are expected to decide on their activities Individuals with good self-regulated learning (SRL) skills are

defined as the ability to plan, manage, and control one's learning process, it helps them to learn faster and better in the instance MOOC aims learners to self-evaluate the quality of the process of their work, it also helps them to make objectives, plans and prove them the option to read notes, logs, test or learning material to prepare for testing among other things [7]. As it is one of the e-learning platforms enabling SRL techniques despite all these advantages several researchers believe that it is still viral to improve student SRL using a machine learning technique.

IV. CONCLUSION

To improve the experience of learning, e-learning experts have invested a lot of time evaluating learners' data using machine learning approaches. This appears to be prudent, given that the student is the most integral part of the e-learning world. However, to the best of our knowledge, no research has been performed on using learning data to measure the quality of the content to enhance it.

Therefore, we will concentrate our future efforts on evaluating e-learning content using machine learning. The fundamental purpose is to encourage course creators in the educational conceptual framework based on using machine learning and a variety of parameters, including previous learner encounters.

REFERENCES

- [1] N. R. Khan, M. Rabbi, K. Al Zabir, K. Dewri, S. A. Sultana and K. J. Lippert, "Internet of Things-Based Educational Paradigm for Best Learning Outcomes," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-8, doi: 10.1109/ACCAI53970.2022.9752569.
- [2] M. N. Rahman Khan, S. Yesmin, M. Aktar, K. B. Quader Chowdhury, K. Labeeb and M. Z. Abedin, "Techniques for Multi-Omics Data Incorporating Machine Learning and System Genomics," 2021 6th International Conference on Communication and Electronics Systems (ICES), 2021, pp. 1524-1528, doi: 10.1109/ICES51350.2021.9489222.
- [3] Khan, Mohammad Nasfikur R and Lippert, Kari J. "A Framework for a Virtual Reality-based Medical Support System." In The Intelligent Systems and Machine learning for Industry: Advancements, Challenges, and Applications, edited by Kishor Kumar Reddy C, 8 – 28. CRC Press, Taylor & Francis Group, 2022.
- [4] Khan, Mohammad Nasfikur R and Lippert, Kari J. "Immersive Technologies in Healthcare Education." In The Intelligent Systems and Machine learning for Industry: Advancements, Challenges, and Applications, edited by Kishor Kumar Reddy C, 111 – 132. CRC Press, Taylor & Francis Group, 2022.
- [5] M. N. R. Khan, A. K. E. H. Mashuk, W. F. Durdana, M. Alam, R. Roy and M. A. Razzak, "Doctor Who? - A Customizable Android Application for Integrated Health Care," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCNT), 2019, pp. 1-6, doi: 10.1109/ICCNT45670.2019.8944501.
- [6] M. N. R. Khan, F. B. Shahin, F. I. Sunny, M. R. Khan, A. K. E. Haque Mashuk and K. A. Al Mamun, "An Innovative and Augmentative Android Application for Enhancing Mediated Communication of Verbally Disabled People," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCNT), 2019, pp. 1-5, doi: 10.1109/ICCNT45670.2019.8944655.
- [7] M. N. R. Khan, W. F. Durdana, R. Roy, G. Poddar, S. Ferdous and A. K. E. H. Mashuk, "Health Guardian - A Subsidiary Android Application for Maintaining Sound Health," 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 2018, pp. 1406-1409, doi: 10.1109/ICRIEECE44171.2018.9008485.
- [8] M. N. R. Khan, H. H. Sonet, F. Yasmin, S. Yesmin, F. Sarker and K. A. Mamun, "'Bolte Chai' — An Android application for verbally challenged children," 2017 4th International Conference on Advances

- in Electrical Engineering (ICAEE), 2017, pp. 541-545, doi: 10.1109/ICAEE.2017.8255415.
- [9] M. N. R. Khan, M. N. H. Pias, K. Habib, M. Hossain, F. Sarker and K. A. Mamun, "Bolte Chai: An augmentative and alternative communication device for enhancing communication for nonverbal children," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), 2016, pp. 1-4, doi: 10.1109/MEDITEC.2016.7835391.
- [10] M. N. R. Khan, M. M. T. Iqbal, S. Yesmin, A. K. E. H. Mashuk, F. B. Shahin and M. A. Razzak, "Development of an Automatic Detection of Pressure Distortion and Alarm System of Endotracheal Tube," 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2018, pp. 476-479, DOI: 10.1109/IECBES.2018.8626676.
- [11] H. Haque, K. Labeeb, R. B. Riha and M. N. R. Khan, "IoT Based Water Quality Monitoring System By Using Zigbee Protocol," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 619-622, DOI: 10.1109/ESCI50559.2021.9397031.
- [12] K. Labeeb, K. B. Q. Chowdhury, R. B. Riha, M. Z. Abedin, S. Yesmin and M. N. R. Khan, "Pre-Processing Data In Weather Monitoring Application By Using Big Data Quality Framework," 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2020, pp. 284-287, DOI: 10.1109/WIECON-ECE52138.2020.9397990.
- [13] K. Lippert, M.N. R. Khan, M. M. Rabbi, A. Dutta, R. Cloutier, "A Framework of Metaverse for Systems Engineering", 2021 IEEE International Conference on Signal Processing, Information, Communication and Systems, December 3 – 4, 2021.
- [14] Khan M.N.R., Shakir A.K., Nadi S.S., Abedin M.Z. (2022) An Android Application for University-Based Academic Solution for Crisis Situation. In: Shakya S., Balas V.E., Kamolphiwong S., Du KL. (eds) Sentimental Analysis and Deep Learning. Advances in Intelligent Systems and Computing, vol 1408. Springer, Singapore. https://doi.org/10.1007/978-981-16-5157-1_51.



IEMTRONICS

International Conference

Toronto, Canada