

Rajesh Kumar · Ajit Kumar Verma ·
Tarun K. Sharma · Om Prakash Verma ·
Sanjay Sharma *Editors*

Soft Computing: Theories and Applications

Proceedings of SoCTA 2022

Lecture Notes in Networks and Systems

Volume 627

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA,
School of Electrical and Computer Engineering—FEEC, University of
Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering,
Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of
Illinois at Chicago, Chicago, USA

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of
Alberta, Alberta, Canada

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering,
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,
Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Rajesh Kumar · Ajit Kumar Verma ·
Tarun K. Sharma · Om Prakash Verma ·
Sanjay Sharma
Editors

Soft Computing: Theories and Applications

Proceedings of SoCTA 2022



Springer

Editors

Rajesh Kumar
Department of Electrical Engineering
Malaviya National Institute of Technology
Jaipur, Rajasthan, India

Tarun K. Sharma
Department of Computer Science
Shobhit University
Gangoh, India

Sanjay Sharma
University Institute of Technology
Himachal Pradesh University
Shimla, Himachal Pradesh, India

Ajit Kumar Verma
Western Norway University of Applied
Sciences
Bergen, Norway

Om Prakash Verma
Department of Instrumentation and Control
Engineering
Dr. B. R. Ambedkar National Institute
of Technology
Jalandhar, Punjab, India

ISSN 2367-3370

Lecture Notes in Networks and Systems

ISBN 978-981-19-9857-7

<https://doi.org/10.1007/978-981-19-9858-4>

ISSN 2367-3389 (electronic)

ISBN 978-981-19-9858-4 (eBook)

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore

Preface

This book stimulated discussions on various emerging trends, innovations, practices and applications in the field of soft computing ranging image and signal processing, network security, supply chain management, computational biology and bioinformatics, human resource management, finance and economics, Internet of things (IoT), AI for smart city, AI for healthcare system, machine vision, remote sensing and GIS, aircraft sensor management, multidisciplinary aerospace design to name a few. This book that we wish to bring forth with great pleasure is an encapsulation of research papers, presented during the three-day International Conference on Seventh International Conference on Soft Computing: Theories and Applications (SoCTA 2022) organized at the University Institute of Technology (UIT), Himachal Pradesh University Shimla, Himachal Pradesh, India, in hybrid mode in technical association with Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Shobhit Deemed University, Meerut, and supported by STEM Research Society. We hope that the effort will be found informative and interesting to those who are keen to learn about technologies that address the challenges of the exponentially growing information in the core and allied fields of soft computing. This book will be beneficial to the young scholars and researchers working in the allied domain. We are thankful to the authors of the research papers for their valuable contribution to the conference and for bringing forth significant research and literature across the field of soft computing. Offering valuable insights into soft computing for the young scholars, researchers, academician and industrialists, this book will inspire further research in this dynamic field.

We express special thanks to reviewers, Springer Nature and team for their valuable support in the publication of the proceedings. With great fervor, we wish to bring together researchers and practitioners in the field of soft computing year after year to explore new avenues in the field.

We are looking forward to the Eighth International Conference on Soft Computing: Theories and Applications (SoCTA 2023), which is scheduled on

December 21–23, 2023, at the Indian Institute of Information Technology (IIIT) Una, Himachal Pradesh, India, with a special focus on *Ethics in Artificial Intelligence*.

Jaipur, India
Bergen, Norway
Gangoh, India
Jalandhar, India
Shimla, India

Rajesh Kumar
Ajit Kumar Verma
Tarun K. Sharma
Om Prakash Verma
Sanjay Sharma

Contents

A CNN-Based Approach for Facial Emotion Detection	1
D. Sahana, K. S. Varsha, Snigdha Sen, and R. Priyanka	
Implementation of Artificial Neural Network for Demanufacturing Operation in the Rail Industry	11
Humbulani Simon Phuluwa, Ilesanmi Daniyan, and Khumbulani Mpofu	
Diversified Recommendation Generation Using Graph Convolution Neural Network	25
Naina Yadav	
Brain Tumor Detection with GLCM Feature Extraction and Hybrid Classification Approach	37
Shardeep Kaur Sooch and Nitika Kapoor	
Optimization of an Inventory Model with Selling Price and Stock Sensitive Demand Along with Trade Credit Policy	47
Mamta Kumari, Pankaj Narang, and Pijus Kanti De	
A New Family of Generalized Euler-Genocchi Polynomials Associated with Hermite Polynomials	59
Azhar Iqbal and Waseem A. Khan	
6G-Enabled Internet of Medical Things	75
Sumit Singh Dhanda, Tarun Kumar Sharma, Brahmjit Singh, Poonam Jindal, and Deepak Panwar	
An Inductively Degenerated LNA for ISM Applications: Design and Performance Comparison	85
P. K. Verma, Himanshu Katiyar, Prashant Pandey, Vikas, and D. K. Tripathi	
Prostate Cancer Risk Analysis Using Artificial Neural Network	99
Anjali Patel, Subhankar Jana, and Juthika Mahanta	

On Two Bivariate Kinds of (p, q)-Euler Polynomials	109
Atul K. Singh, Idrees A. Khan, Nidal Abu-Libdeh, and Waseem A. Khan	
Different Stages of Watermelon Diseases Detection Using Optimized CNN	121
Samah Alhazmi	
Machine Learning: An Analytical Approach for Pattern Detection in Diabetes	135
Ritu Chauhan, Anika Goel, Harleen Kaur, and Bhavya Alankar	
A Dynamic Weighted Federated Learning for Android Malware Classification	147
Ayushi Chaudhuri, Arijit Nandi, and Buddhadeb Pradhan	
Entropy Measure for the Linguistic q-Rung Orthopair Fuzzy Set	161
Neelam, Kamal Kumar, and Reeta Bhardwaj	
Empirical Analysis of Unsupervised Link Prediction Algorithms in Weighted Networks	173
Ajay Kumar, Shashank Sheshar Singh, and Shivansh Mishra	
Detection of Fraudulent Credit Card Transactions Using Deep Neural Network	185
Kotireddy Yazna Sai, Repalle Venkata Bhavana, and Natarajan Sudha	
Recommendation System and Its Techniques in Machine Learning: A Survey	197
Neeru Banwala, Gurpreet Singh, Jaspreet Singh, Vishwajeet Shankar Goswami, and Aashima Bagnia	
Approximation of Signal Belongs to $W'(L^p, \xi(t))$ Class by Generalized Nörlund-Cesáro Product Means	207
Smita Sonker and Paramjeet Sangwan	
Dignet: A Deep Learning-Based Efficient Digit Recognition System	219
Debashish Mondal, Narinder Kumar, and Ravinder Kaur	
Voice Command Automation System (VCAS) for Controlling Electrical Devices Using Arduino	231
Maliha Rahman, Abdullah Al Farabe, Md. Rayhan Al Islam, Moshiur Rahman, Md. Rezyuan, and Ghalib Ashraf	
Turing Machines Behind the Horizon: Modeling Black Hole Interiors as Transfinite Limited Turing Machines	243
Ajay Agarwal	
A Review on Deep Learning-Enabled Healthcare Prediction Technique: An Emerging Digital Governance Approach	253
D. Anand, Venkateswarlu Tata, Jitendra Kumar Samriya, and Mohit Kumar	

Evaluation of Deep Learning Technique on Working Model of Self-driving Car—A Review	265
Somin Sangwan, Gurpreet Singh, Aashima Bagnia, and Vishwajeet Shankar Goswami	
Analysis of the Distractions in Youth Due to Social Media and the Effects on Their Concentration Abilities	279
S. Prajwal, N. Aditi, Dharithri B. Sharma, S. Syed Afreeth, K. Ashwini, and Srirupa Guha	
The Intervention of Technology in Education Under Isolation: Intuitions from Covid	293
Stephen Owusu Afriyie, Joseph Akwasi Nkyi, Gertrude Amoakohene, Mohammed Musah, and Peter Yao Lartey	
Analysis of Bao-Zhou-Chen-Liu's Hybrid Chaotic System	303
Meenakshi Agarwal, Arvind, and Ram Ratan	
Identification of Skin Lesion with Adaptive Tasmanian Devil Optimization-Based Transfer Learning	317
Vineet Kumar Dubey and Vandana Dixit Kaushik	
Synchronization of MLS Chaotic System Using Sliding Mode Control Technique	335
Pallav and Himesh Handa	
Performance Analysis of User Behavior Pattern Mining Using Web Log Database for User Identification	347
Gokulapriya R. and Ganesh Kumar R.	
An Imperfect Production System for Non-instantaneous Deteriorating Goods with Preservation Technology Under Cap-and-Trade Policy	357
Pankaj Narang, Mamta Kumari, and Pijus Kanti De	
Human Activity Recognition Using a Hybrid Dilated CNN and GRU ...	371
Preeti Gupta and Satish Chand	
A New Framework for Disease Prediction: Using Dimensionality Reduction and Feature Selection	381
Shreya Sahu, Pranesh Das, and A. Binu Jose	
Review of Metaheuristic Techniques for Feature Selection	397
Sanat Jain, Ashish Jain, and Mahesh Jangid	
New Type of Degenerate Changhee–Genocchi Polynomials of the Second Kind	411
Azhar Iqbal, Waseem A. Khan, and Mohd Nadeem	
Temperature Aware Bi-partitioning Multi-level Logic Synthesis	423
Apangshu Das, Vivek Kumar Singh, and Sambhu Nath Pradhan	

Determination Human Behavior Prediction Supported by Cognitive Computing-Based Neural Network	431
Jyoti Parashar, Virendra Singh Kushwah, and Munishwar Rai	
Single IC-Based Third-Order Sinusoidal Oscillator	443
Gurumurthy Komanapalli, Pandey Rajeshwari, and Pandey Neeta	
Internet of Things (IoT)-Based Waste Dumping System	453
Birinderjit Singh Kalyan	
Constructing a Smart School Based on the Internet of Things Using RFID Technology	463
Soha Alhelaly	
Effect of Confining Walls on Settling Permeable Rigid Isolated Semi-torus Particle Applying Immersed Boundary Method (IBM)	473
Pooja Yadav, Sudeshna Ghosh, Amit Sharma, and Rekha Panghal	
Hybridized Shuffled Frog Leaping Algorithm for Solving Facility Location Problem for Maternal Healthcare	481
Ankit Chouksey, A. K. Agrawal, and Arkaprava Ray	
Key Observation to Prevent IP Spoofing in DDoS Attack on Cloud Environment	493
T. Sunitha, V. Vijayashanthi, M. Navaneethakrishnan, T. A. Mohanaprakash, S. Ashwin, T. R. Harish, and Emmanuel A. Stanes	
Performance Enhancement of Magnetic Levitation System Using GWO-ABC Tuned High-Dimensional Robust Controller	507
Shirish Adam and Prashant Gaidhane	
Approximation of Function Belonging to $\text{Lip}(\xi(t), p)$ Class by Using Borel's Mean	519
Smita Sonker and Rozy Jindal	
Combining Genetic Algorithm and Support Vector Machine for Classification of Cancer on Microarray Data	527
Tanja Plagemann, Rolf Dornberger, and Thomas Hanne	
Design and Simulation of a Novel Digital Technology for Assembly Operations: A Case Study of Railcar Bogie Application	539
Ilesanmi Daniyan, Khumbulani Mpofu, Lanre Daniyan, Felix Ale, and Nokulunga Zamahlubi Dlamini	
Solving Fixed Charge Transportation Problem with Interval Parameters Using Generalized Reduced Gradient Method	551
Subhayan Das and Subhra Das	
Automated Solar PV Array Cleaning Based on Aerial Computer Vision Framework	563
Shreya Nallapaneni, Kairavi Shah, and Harsh S. Dhiman	

An Ensemble Framework for Glaucoma Classification Using Fundus Images	573
Achirangshu Patra, Arijit Nandi, Mayaluri Zefree Lazarus, and Satyabrata Lenka	
A Note on Laguerre-Based Appell-Type Daehee Polynomials and Numbers	589
Waseem A. Khan, Azhar Iqbal, and Mohd Nadeem	
Type-II Fuzzy Kernel-Based Multi-layer Extreme Learning Machine ...	601
Avatharam Ganivada and Sayima Mukhtar	
Energy Trading in Smart Grids Using Game Theoretic Approach	611
Anshul Agarwal	
Design of Disturbance Observer-Based Dynamic Sliding Mode Control	623
S. S. Nerkar and B. M. Patre	
Analysis on the Financial Performance of Technology Companies in Malaysia with VIKOR Model	637
Weng Siew Lam, Kah Fai Liew, Weng Hoe Lam, and Mohd Abidin Bin Bakar	
IoT-Based Online Condition Monitoring and Fault Analysis of Bearings of a Rotating Machinery	645
Sudarsan Sahoo, Chokka Upendra, Krishnananda Sahu, Nabajit Bharali, and Suresh Nuthalapati	
Influence of Time Delay on Predator-Prey Model Having Herd Behaviour and Hunting Cooperation	655
Shivam, Teekam Singh, and Mukesh Kumar	
An Improved Jaya Algorithm (IJAYA) for Optimization	665
Sonal Deshwal, Pravesh Kumar, and Sandeep Mogha	
An Optimized Approach for Emotion Detection-Based Music Recommendation System	675
Manoj K. Sabnis and Bhavesh Bhatia	
A Method to Solve Fractional Transportation Problems with Rough Interval Parameters	689
Shivani and Deepika Rani	
Applications of IoT and Various Attacks on IoT	705
Sumeet Dhillon, Nishchol Mishra, and Devendra Kumar Shakya	
Real-Time Implementation of Laguerre Neural Network-Based Adaptive Control of DC-DC Converter	721
Sasank Das Gangula, Tousif Khan Nizami, U. Ramanjaneya Reddy, and Priyanka Singh	

5G New Radio Physical Downlink Shared Channel Throughput Analysis with Different Numerology and Modulation Schemes	733
Rajesh Kumar, Deepak Sinwar, and Vijander Singh	
Automated Text Summarization Using Transformers	743
Yogesh Kumar, Ashish Jangir, Bhavya Meena, and Isha Pathak Tripathi	
Minimizing Building Energy Waste by Detecting and Addressing HVAC Issues	755
Anshul Agarwal	
Knowledge Representation and Information Retrieval from Ontologies	765
Azra Bashir, Renuka Nagpal, Deepti Mehrotra, and Manju Bala	
Design and Analysis of GIGA Fiber like Connectivity of 5G Technology Using 60 GHz Band	777
Bhanu P. Singh and Anand Agarwal	
Coronavirus Herd Immunity Optimization-Based Control of DC-DC Boost Converter	787
Manoj Sai Pendem, Tousif Khan Nizami, Priyanka Singh, and Mohamed Shaik Honnurvali	
Improvisation in Opinion Mining Using Negation Detection and Negation Handling Techniques: A Survey	799
Kartika Makkar, Pardeep Kumar, and Monika Poriye	
Localization of the Closed-Loop Differential Drive Mobile Robot Using Wheel Odometry	809
Gurpreet Singh and Vijay Kumar	
Analysis of Effectiveness of Online Classes During COVID	819
Disha Sriram, Lavanya Sanjay, Neha Nayak, Sathwik Sathish, Ashwini Kodipalli, and P. N. Anil	
Suspicious Crime Identification and Detection Based on Social Media Crime Analysis Using Machine Learning Algorithms	831
C. Jayapratha, H. Salome Hema Chitra, and R. Mahalakshmi Priya	
Deep Learning-Based Similar Languages' POS Tagging: Experiments on Bhojpuri, Maithili, and Magahi	845
Rajesh Kumar Mundotiya, Praveen Gatla, Nikita Kanwar, and Anil Kumar Singh	
CS-Jaya: Hybridization of Cuckoo and Jaya Algorithm	857
Megha Varshney, Pravesh Kumar, and Tarun Kumar Sharma	
Plant Leaf Disease Detection Using ResNet	867
Amit Kumar, Manish Kumar Priyanshu, Rani Singh, and Snigdha Sen	

Contents	xiii
Automatic Infographic Builder Using Natural Language Statements	879
Chetali Neema and Anuradha Purohit	
A Novel Type-2 Fuzzy Programming Approach for Solving	
Multiobjective Programming Problems	889
Animesh Biswas, Debjani Chakraborty, Bappaditya Ghosh, and Arnab Kumar De	
Person Detection Using YOLOv3	903
Bhawana Tyagi, Swati Nigam, and Rajiv Singh	
Incident Reporting of Forest Fire with Azure Cognitive Services	
and Twitter API	913
Rakesh Kumar, Meenu Gupta, Dhruv Kinger, and Sayanto Roy	
Role of Telemedicine in Healthcare Sector for Betterment of Smart	
City	925
Prashant Sahatiya and Dheeraj Kumar Singh	
A Comparative Study on Abstractive Text Summarization	
Techniques Using Deep Learning (ATS-DL)	937
S. Adhithyan, A. R. Nirupama, S. Sri Akshya, S. Swamynathan, and K. Girthana	
Author Index	949

Editors and Contributors

About the Editors

Dr. Rajesh Kumar received the Bachelor of Technology in Engineering degree with honours in Electrical Engineering from the Department of Electrical Engineering, National Institute of Technology, Kurukshetra, India, in 1994, Master of Engineering with honours in Power Engineering from the Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India, in 1997 and Ph.D. degree in Intelligent Systems from Department of Electrical Engineering, Malaviya National Institute of Technology (MREC, University of Rajasthan), India, in 2005. He is currently Professor at Department of Electrical Engineering and Adjunct Professor at Centre of Energy and Environment at Malaviya National Institute of Technology, Jaipur, India. He has been Research Fellow (A) at the Department of Electrical and Computer Engineering at National University of Singapore from 2009–2011. He has been Reader from 2005–2009, Senior Lecturer from 2000–2005 and Lecturer from 1995–2000 at Department of Electrical Engineering, Malaviya National Institute of Technology. He is Founder of ZINE student innovative group. His background is in the fields of computational intelligence, artificial intelligence, intelligent systems, power and energy management, robotics, bioinformatics, smart grid and computer vision.

Ajit Kumar Verma is Professor in the Faculty of Engineering and Natural Sciences, Western Norway University of Applied Sciences, Haugesund, Norway (since March 2012) and has been Professor (since February 2001) with the Department of Electrical Engineering at IIT Bombay with a research focus in Reliability and Safety Engineering. He was Director of the International Institute of Information Technology, Pune, on lien from IIT Bombay, from August 2009–September 2010. He is also Guest Professor at Lulea University of Technology, Sweden. He has supervised/co-supervised 38 Ph.D.s and 95 Masters theses in the area of software reliability, reliable computing, power systems reliability (PSR), reliability centred maintenance (RCM) and probabilistic safety/risk assessment (PSA). He has executed

various research projects to promote industry interface and has been course co-ordinator for industry CEPs like reliability engineering, six sigma, software inspections, competency tracking system and software reliability for IT industries. He is Springer Book Series Editor of Asset Analytics-Performance and Safety Management, Fire safety Engineering and Management and Reliable and Sustainable Electric Power and Energy Systems Management and has jointly edited books titled *Reliability and Risk Evaluation of Wind Integrated Power Systems* (Springer), *Reliability Modelling Analysis of Smart Power Systems* (Springer) and *Current Trends in Reliability, Availability, Maintainability and Safety* (Springer) and is Author of books titled *Fuzzy Reliability Engineering-Concepts and Applications* (Alpha Science International—Narosa), *Reliability and Safety Engineering* (Springer) and *Dependability of Networked Computer Based Systems* (Springer), *Risk Management of Non-Renewable Energy Systems* (Springer) and *Optimal Maintenance of Large Engineering Systems* (Narosa). He has over 250 publications in various journals (over 100 papers) and conferences. He is Senior Member of IEEE and Life Fellow of IETE. He has been Editor-in-Chief of *OPSEARCH* published by Springer (January 2008–January 2011) as well as Founder Editor-in-Chief of *International Journal of Systems Assurance Engineering and Management* (IJSADM) published by Springer and Editor-in-Chief of *Journal of Life Cycle Reliability and Safety Engineering*. He is also on the editorial board of several international journals. He was Guest Editor of Special Issue on “Reliable Computing” of *International Journal of Automation and Computing* (October 2007), May 2010, *IJSADM* (June 2010) (with Prof. Roy Billington and Prof. Rajesh Karki) and *IJSADM* (March 2011) (with Prof. Lotfi A. Zadeh and Prof. Ashok Deshpande) and *IEEE Transactions on Reliability*, March 2011 among others. He was also nominated as Chairman of the “Special Interest Group on System Assurance Engineering and Management” of Berkeley Initiative in Soft Computing” by Prof. Lotfi A. Zadeh.

Dr. Tarun K. Sharma holds Ph.D. in Soft Computing from the Department of Applied Science and Engineering of IIT Roorkee. Since 1 April 2022, he is associated with Shobhit Deemed University, Meerut, as Professor and Head of CSE and Dean—School of Engineering and Technology. Earlier, he worked with Shobhit University Gangoh as Professor and Head of CSE and Dean—School of Engineering and Technology and Amity University Rajasthan as Associate Professor and Head—Department of Computer Science and Engineering/IT as well as Alternate Director—Outcome. He has supervised 4 Ph.D.s, 14 M.Tech. dissertations, several M.C.A. and B.Tech. projects. He has over 80 research publications in his credit. He has been to Amity Institute of Higher Education Mauritius on deputation. He has availed grants from Microsoft Research India, CSIR, New Delhi and DST New Delhi to visit Australia, Singapore and Malaysia, respectively. He is Founding Member of International Conference on Soft Computing: Theories and Applications (SoCTA Series) and Congress on Advances in Materials Science and Engineering (CAMSE). He has edited seven volumes of Conference Proceedings published by AISC series of Springer (SCOPUS) Publication and four edited books with Springer Nature.

Dr. Om Prakash Verma is presently associated with Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, India, since January 2018 as Assistant Professor in the Department of Instrumentation and Control Engineering. He has almost 11 years of teaching experience. He did his Ph.D. from IIT Roorkee, M.Tech. from Dr. B. R. Ambedkar NIT Jalandhar and B.E. from Dr. B. R. Ambedkar University, Agra. He is presently working on ISRO Sponsored Project as PI. He has edited a book on *Soft Computing: Theories and Applications* and has been Potential Reviewer of several International Journals of high repute. He has published more than 30 research papers in SCI/Scopus/ESI indexed Journals. He has published recently published a paper in Renewable and Sustainable Energy Reviews, (IF: 12.110). He has guided four M.Tech. students and supervising six Ph.D. students.

Dr. Sanjay Sharma received B.Tech. degree in Electrical Engineering from Himachal Pradesh University, Shimla, India, in 2007 and M.Tech. degree in Power System from National Institute of Technology, Hamirpur, India, in 2010. He has completed her Ph.D. degree in Department of Electrical Engineering from Punjab Engineering College Deemed to be University, Chandigarh, India, in December 2019. Presently, he is working as Assistant Professor in University Institute of Technology, Himachal Pradesh University, Shimla Himachal Pradesh, India. He has good research experience in various areas of Electrical Engineering. He worked on a Project “Development and validation of technology for production of high energy density from rice straw and Agri-biomasses (Funding agency PSA, GOI and Sweden). He published 16 research papers in reputed international and national journals, and 18 research papers in international and national conferences. His area of research includes power system, renewable energy, network planning, microgrid, optimization, GIS and machine learning. He has published a number of research papers in various journals and conferences. Presently, his one book on renewable energy has been accepted for Publication in Willey Publication likely to be in market in December 2021. He also granted with one patent and one in progress. He has six-year teaching experience of NITs and government institute. He has delivered expert lecture in various colleges/universities in India. He is Reviewer of various conferences and international journals including Elsevier, Springer, IEEE and Taylor & Francis.

Contributors

Nidal Abu-Libdeh Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

Shirish Adam Government College of Engineering, Jalgaon, India

S. Adhithyan College of Engineering, Anna University, Chennai, Tamil Nadu, India

N. Aditi Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

S. Syed Afreeth Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Stephen Owusu Afriyie Ghana Communication Technology University, I.T. Business, Accra-North, Ghana

Ajay Agarwal DIT University, Dehradun, India

Anand Agarwal ECE, Indian Institute of Information Technology, Kota, India

Anshul Agarwal Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT) Nagpur, Nagpur, Maharashtra, India

Meenakshi Agarwal Department of Mathematics, University of Delhi, Delhi, India

A. K. Agrawal Indian Institute of Technology (BHU) Varanasi, Uttar Pradesh, India

Abdullah Al Farabe Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Md. Rayhan Al Islam Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Bhavya Alankar Department of Computer Science and Engineering, Jamia Hamdard, Delhi, India

Felix Ale Department of Engineering and Space Systems, National Space Research & Development Agency, Abuja, Nigeria

Samah Alhazmi College of Computing and Informatics, Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia

Soha Alhelaly College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia

Gertrude Amoakohene Ghana Communication Technology University, I.T. Business, Accra-North, Ghana

D. Anand Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

P. N. Anil Department of Mathematics, Global Academy of Technology, Bangalore, India

Arvind Department of Mathematics, Hansraj College, University of Delhi, Delhi, India

Ghalib Ashraf Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

S. Ashwin Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, India

K. Ashwini Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Aashima Bagnia Department of Mathematics, Chandigarh University, Mohali, India

Mohd Abidin Bin Bakar Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Manju Bala Indraprastha College for Women, Delhi University, New Delhi, India

Neeru Banwala Data Science, Department of Mathematics, Chandigarh University, Mohali, India

Azra Bashir Amity University, Noida, India

Nabajit Bharali NIT Silchar, Silchar, Assam, India

Reeta Bhardwaj Department of Mathematics, Amity School of Applied Sciences, Amity University Haryana, Gurugram, India

Bhavesh Bhatia VESIT, Collector Colony, Chembur, Mumbai, India

A. Binu Jose Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India

Animesh Biswas University of Kalyani, Kalyani, India

Debjani Chakraborty Indian Institute of Technology Kharagpur, Kharagpur, India

Satish Chand Jawaharlal Nehru University, New Delhi, India

Ayushi Chaudhuri Department of Computer Science and Engineering, Vellore Institute of Technology (VIT), Bhopal, India

Ritu Chauhan Center for Computational Biology and Bioinformatics, Amity University, Noida, Uttar Pradesh, India

H. Salome Hema Chitra Department of CS, Sri Meenakshi Govt. Arts College for Women's, Madurai, Tamil Nadu, India

Ankit Chouksey Indian Institute of Technology (BHU) Varanasi, Uttar Pradesh, India

Ilesanmi Daniyan Department of Industrial Engineering, Tshwane University of Technology, Pretoria, South Africa

Lanre Daniyan Department of Instrumentation, University of Nigeria, Nsukka, Nigeria

Apangshu Das Department of Electronics and Communication Engineering, NIT, Agartala, India

Pranesh Das Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India

Subhayan Das School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha, India

Subhra Das Amity School of Engineering and Technology, Amity University Haryana, Gurugram, India

Arnab Kumar De Government College of Engineering and Textile Technology, Serampore, India

Pijus Kanti De Department of Mathematics, National Institute of Technology Silchar, Silchar, Assam, India

Sonal Deshwal Department of Mathematics, Chandigarh University, Mohali, Punjab, India

Sumit Singh Dhandha Department of Electronics and Communication, National Institute of Technology, Kurukshetra, Haryana, India

Sumeet Dhillon Computer Science and Engineering Department, SATI, Vidisha, India

Harsh S. Dhiman Department of Electrical Engineering, Adani Institute of Infrastructure Engineering, Ahmedabad, India;

Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune, India

Nokulunga Zamahlubi Dlamini Department of Industrial Engineering, Tshwane University of Technology, Pretoria, South Africa

Rolf Dornberger University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland

Vineet Kumar Dubey Harcourt Butler Technological University, Kanpur, India

Prashant Gaidhane Government College of Engineering, Jalgaon, India

Ganesh Kumar R. Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, Karnataka, India

Sasank Das Gangula School of Engineering and Sciences, SRM University AP, Guntur, Andhra Pradesh, India

Avatharam Ganivada School of Computer and Information Sciences, University of Hyderabad, Hyderabad, Telangana, India

Praveen Gatla Department of Linguistics, Faculty of Arts, BHU, Varanasi, India

Bappaditya Ghosh University of Kalyani, Kalyani, India

Sudeshna Ghosh Amity School of Applied Sciences, Amity University, Haryana, India

K. Girthana College of Engineering, Anna University, Chennai, Tamil Nadu, India

Anika Goel Center for Computational Biology and Bioinformatics, Amity University, Noida, Uttar Pradesh, India

Gokulapriya R. Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore, Karnataka, India

Vishwajeet Shankar Goswami Department of Mathematics, Chandigarh University, Mohali, India

Srirupa Guha National Institute of Technology Durgapur, Durgapur, India

Meenu Gupta Department of Computer Science, Chandigarh University, Punjab, India

Preeti Gupta Jawaharlal Nehru University, New Delhi, India

Himesh Handa National Institute of Technology, Hamirpur, Himachal Pradesh, India

Thomas Hanne University of Applied Sciences and Arts Northwestern Switzerland, Basel, Olten, Switzerland

T. R. Harish Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, India

Mohamed Shaik Honnurvali A'Sharqiyah University, Ibra, Sultanate of Oman

Azhar Iqbal Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

Ashish Jain Department of Information Technology, Manipal University Jaipur, Jaipur, India

Sanat Jain Department of Information Technology, Manipal University Jaipur, Jaipur, India

Subhankar Jana NIT Silchar, Silchar, Assam, India

Mahesh Jangid Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India

Ashish Jangir Indian Institute of Information Technology, Kota, Rajasthan, India

C. Jayapratha Department of MCA, Karpaga Vinayaga College of Engg. Tech, Madurantakam, Tamil Nadu, India

Poonam Jindal Department of Electronics and Communication, National Institute of Technology, Kurukshetra, Haryana, India

Rozy Jindal Department of Mathematics, National Institute of Technology Kurukshetra, Thanesar, Haryana, India

Birinderjit Singh Kalyan University Institute of Engineering, Chandigarh University, Mohali, India

Nikita Kanwar Computer Science and Engineering Department, PRATAP Institute of Technology and Science, Sikar, India

Nitika Kapoor Chandigarh University, Gharuan, Mohali, India

Himanshu Katiyar Electronics Engineering Department, Rajkiya Engineering College, Sonbhadra, UP, India

Harleen Kaur Department of Computer Science and Engineering, Jamia Hamdard, Delhi, India

Ravinder Kaur Department of CSE, Chandigarh University, Mohali, India

Vandana Dixit Kaushik Harcourt Butler Technological University, Kanpur, India

Idrees A. Khan Department of Mathematics and Statistics, Faculty of Science, Integral University, Lucknow, India

Waseem A. Khan Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

Dhruv Kinger Department of Computer Science, Chandigarh University, Punjab, India

Ashwini Kodipalli Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Gurumurthy Komanapalli School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

Ajay Kumar UPES, Dehradun, India

Amit Kumar Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

Kamal Kumar Department of Mathematics, Amity School of Applied Sciences, Amity University Haryana, Gurugram, India

Mohit Kumar Dr B. R. Ambedkar NIT Jalandhar, Jalandhar, India

Mukesh Kumar Department of Mathematics, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India

Narinder Kumar Department of CSE, Chandigarh University, Mohali, India

Pardeep Kumar Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India

Pravesh Kumar Department of Mathematics, Rajkiya Engineering College Bijnor, Bijnor, India;

Rajkiya Engineering College Bijnor (AKTU Lucknow), Lucknow, India

Rajesh Kumar Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India

Rakesh Kumar Department of Computer Science, Chandigarh University, Punjab, India

Vijay Kumar Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Yogesh Kumar Indian Institute of Information Technology, Kota, Rajasthan, India

Mamta Kumari Department of Mathematics, National Institute of Technology Silchar, Silchar, Assam, India

Virendra Singh Kushwah VIT Bhopal University, Kothrikalan, Sehore, Madhya Pradesh, India

Weng Hoe Lam Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Weng Siew Lam Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Peter Yao Lartey School of Accounting, Universidade Federal de Uberlândia, Uberlândia, MG, Brazil

Mayaluri Zefree Lazarus Department of Electrical and Electronics Engineering, C.V. Raman Global University, Bhubaneswar, Odisha, India

Satyabrata Lenka Department of Electrical and Electronics Engineering, C.V. Raman Global University, Bhubaneswar, Odisha, India

Kah Fai Liew Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

Juthika Mahanta NIT Silchar, Silchar, Assam, India

Kartika Makkar Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India

Bhavya Meena Indian Institute of Information Technology, Kota, Rajasthan, India

Deepti Mehrotra Amity University, Noida, India

Nishchol Mishra School of Information Technology, RGPV Bhopal, Bhopal, India

Shivansh Mishra IIT BHU Varanasi, Varanasi, India

Sandeep Mogha Department of Mathematics, Chandigarh University, Mohali, Punjab, India

T. A. Mohanaprabakash Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India

Debashish Mondal Department of CSE, Chandigarh University, Mohali, India

Khumbulani Mpofu Department of Industrial Engineering, Tshwane University of Technology, Pretoria, South Africa

Sayima Mukhtar School of Computer and Information Sciences, University of Hyderabad, Hyderabad, Telangana, India

Rajesh Kumar Mundotiya University of Petroleum and Energy Studies, Dehradun, India

Mohammed Musah Ghana Communication Technology University, I.T. Business, Accra-North, Ghana

Mohd Nadeem Department of Natural and Applied Sciences, Glocal University, Saharanpur, Uttar Pradesh, India

Renuka Nagpal Amity University, Noida, India

Shreya Nallapaneni Department of Electrical Engineering, Adani Institute of Infrastructure Engineering, Ahmedabad, India

Arijit Nandi Department of Computer Science, Universitat Politècnica de Catalunya (Barcelona Tech), Barcelona, Spain;
Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain

Pankaj Narang Department of Mathematics, National Institute of Technology Silchar, Silchar, Assam, India

M. Navaneethakrishan Department of Computer Science and Engineering, St. Joseph College of Engineering, Chennai, India

Neha Nayak Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Neelam Department of Mathematics, Amity School of Applied Sciences, Amity University Haryana, Gurugram, India

Chetali Neema Department of Computer Engineering, Shri G. S. Institute of Technology and Science, Indore, Madhya Pradesh, India

Pandey Neeta Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

S. S. Nerkar Department of Instrumentation Engineering, Government College of Engineering, Jalgaon, Maharashtra, India

Swati Nigam Department of Computer Science, Banasthali Vidyapith, Banasthali, Rajasthan, India

A. R. Nirupama College of Engineering, Anna University, Chennai, Tamil Nadu, India

Tousif Khan Nizami School of Engineering and Sciences, SRM University AP, Guntur, Andhra Pradesh, India

Joseph Akwasi Nkyi Ghana Communication Technology University, I.T. Business, Accra-North, Ghana

Suresh Nuthalapati Technische University at Dresden, Dresden, Germany

Pallav National Institute of Technology, Hamirpur, Himachal Pradesh, India

Prashant Pandey Electronics Engineering Department, Rajkiya Engineering College, Sonbhadra, UP, India

Rekha Panghal Amity School of Applied Sciences, Amity University, Haryana, India

Deepak Panwar Manipal University Jaipur, Jaipur, India

Jyoti Parashar Dr. Akhilesh Das Gupta Institute of Technology Management, New Delhi, India

Anjali Patel NIT Silchar, Silchar, Assam, India

Achirangshu Patra Department of Mechatronics Engineering, C.V. Raman Global University, Bhubaneswar, Odisha, India

B. M. Patre Department of Instrumentation Engineering, S.G.G.S. Institute of Engineering and Technology, Nanded, Maharashtra, India

Manoj Sai Pendem SRM University-AP, Guntur, Andhra Pradesh, India

Humbulani Simon Phuluwa Department of Industrial Engineering, Tshwane University of Technology, Pretoria, South Africa

Tanja Plagemann University of Applied Sciences and Arts Northwestern Switzerland, Basel, Olten, Switzerland

Monika Poriye Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India

Buddhadeb Pradhan Department of Computer Science and Engineering, University of Engineering and Management, Kolkata, India

Sambhu Nath Pradhan Department of Electronics and Communication Engineering, NIT, Agartala, India

S. Prajwal Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

R. Mahalakshmi Priya Department of CS, Mangayarkarasi College of Arts and Science, Madurai, Tamil Nadu, India

R. Priyanka Department of ECE, Global Academy of Technology, Bengaluru, Karnataka, India

Manish Kumar Priyanshu Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

Anuradha Purohit Department of Computer Engineering, Shri G. S. Institute of Technology and Science, Indore, Madhya Pradesh, India

Maliha Rahman Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Moshiur Rahman Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Munishwar Rai Maharishi Markandeshwar Deemed-to-be University, Ambala, India

Pandey Rajeshwari Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

U. Ramanjaneya Reddy School of Engineering and Sciences, SRM University AP, Guntur, Andhra Pradesh, India

Deepika Rani Department of Mathematics, Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, India

Ram Ratan Defence Research and Development Organization, Delhi, India

Arkaprava Ray Jadavpur University Kolkata, West Bengal, India

Md. Rezyuan Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Sayanto Roy Department of Computer Science, Chandigarh University, Punjab, India

Manoj K. Sabnis VESIT, Collector Colony, Chembur, Mumbai, India

D. Sahana Department of ECE, Global Academy of Technology, Bengaluru, Karnataka, India

Prashant Sahatiya Department of Information Technology, Parul University, Vadodara, Gujarat, India

Sudarsan Sahoo NIT Silchar, Silchar, Assam, India

Krishnananda Sahu NIT Silchar, Silchar, Assam, India

Shreya Sahu Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India

Jitendra Kumar Samriya Dr B. R. Ambedkar NIT Jalandhar, Jalandhar, India

Paramjeet Sangwan National Institute of Technology Kurukshetra, Kurukshetra, India

Somin Sangwan Data Science, Department of Mathematics, Chandigarh University, Mohali, India

Lavanya Sanjay Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Sathwik Sathish Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Snigdha Sen Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

Kairavi Shah Department of Electrical Engineering, Adani Institute of Infrastructure Engineering, Ahmedabad, India

Devendra Kumar Shakya Department of Electronics Engineering, SATI, Vidisha, India

Amit Sharma Amity School of Applied Sciences, Amity University, Haryana, India

Dharithri B. Sharma Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Tarun Kumar Sharma Shobhit Institute of Engineering and Technology (Deemed to be) University, Meerut, India

Shivam Department of Mathematics, Graphic Era (Deemed to be) University, Dehradun, Uttarakhand, India

Shivani Department of Mathematics, Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, India

Anil Kumar Singh Department of Computer Science and Engineering, IIT (BHU), Varanasi, India

Atul K. Singh Department of Mathematics and Statistics, Faculty of Science, Integral University, Lucknow, India

Bhanu P. Singh ECE, Lakshmi Narian College of Technology, Bhopal, Madhya Pradesh, India

Brahmjit Singh Department of Electronics and Communication, National Institute of Technology, Kurukshetra, Haryana, India

Dheeraj Kumar Singh Department of Information Technology, Parul University, Vadodara, Gujarat, India

Gurpreet Singh Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;

Department of Computer Science and Engineering, Chandigarh University, Mohali, India

Jaspreet Singh Department of Computer Science and Engineering, Chandigarh University, Mohali, India

Priyanka Singh School of Engineering and Sciences, SRM University AP, Guntur, Andhra Pradesh, India

Rajiv Singh Department of Computer Science, Banasthali Vidyapith, Banasthali, Rajasthan, India

Rani Singh Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

Shashank Sheshar Singh Thapar University, Patiala, Punjab, India

Teekam Singh Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India

Vijander Singh Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India

Vivek Kumar Singh Department of Electronics and Communication Engineering, NIT, Agartala, India

Deepak Sinwar Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India

Smita Sonker Department of Mathematics, National Institute of Technology Kurukshetra, Thanesar, Haryana, India

Shardeep Kaur Sooch Chandigarh University, Gharuan, Mohali, India

S. Sri Akshya College of Engineering, Anna University, Chennai, Tamil Nadu, India

Disha Sriram Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

Emmanuel A. Stanes Department of Computer Science and Engineering, Pani-malar Institute of Technology, Chennai, India

Natarajan Sudha SASTRA Deemed University, Thanjavur, India

T. Sunitha Department of Computer Science and Engineering, P. B. College of Engineering, Chennai, India

S. Swamynathan College of Engineering, Anna University, Chennai, Tamil Nadu, India

Venkateswarlu Tata Guntur Engineering College, JNTUK, Kakinada, India

D. K. Tripathi Electronics Engineering Department, Rajkiya Engineering College, Sonbhadra, UP, India

Isha Pathak Tripathi Indian Institute of Information Technology, Kota, Rajasthan, India

Bhawana Tyagi Department of Computer Science, Banasthali Vidyapith, Banasthali, Rajasthan, India

Chokka Upendra NIT Silchar, Silchar, Assam, India

K. S. Varsha Department of ECE, Global Academy of Technology, Bengaluru, Karnataka, India

Megha Varshney Rajkiya Engineering College Bijnor (AKTU Lucknow), Lucknow, India

Repalle Venkata Bhavana SASTRA Deemed University, Thanjavur, India

P. K. Verma Electronics Engineering Department, Rajkiya Engineering College, Sonbhadra, UP, India

V. Vijayashanthi Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College (Autonomous), Chennai, India

Vikas Institute of Electronics Engineering, National Tsing Hua University, Hsinchu, Taiwan

Naina Yadav School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India

Pooja Yadav Amity School of Applied Sciences, Amity University, Haryana, India

Kotireddy Yazna Sai SASTRA Deemed University, Thanjavur, India

A CNN-Based Approach for Facial Emotion Detection



D. Sahana, K. S. Varsha, Snigdha Sen, and R. Priyanka

Abstract One of the most versatile ways in which individuals express their state of mind is through facial expressions. The advancement of deep learning-based technologies helped us to detect human emotion from images that can be used for understanding human feelings as well. The image can be static or can be captured through a web camera in real time. The precise analysis of human facial expressions is necessary for a better understanding of human behaviour. With the recent progress in deep learning, Convolution Neural Network (CNN) with its enhanced complex architecture is capable of emotion detection in a much better and more efficient way. In this paper, we experiment and demonstrate how to build a CNN predictor model using TensorFlow that can predict the emotion from images of human facial expressions with satisfactory accuracy. Additionally, we also develop an application that asks for image input from the user and predicts the emotion from the given input image. Through this experiment, we are successful in demonstrating how CNN is an appropriate model for this task. Our work is beneficial in many applications such as lie detectors and student assessments to detect facial expressions very accurately.

Keywords Convolution neural network · Deep learning · Facial emotion recognition · Static image · Webcam

1 Introduction

Emotions often arbitrate and ease interactions among human beings. A person's state of mind is conceivable and determined by various modes such as speech patterns, gestures, and many other complex methods. However, the easier and more practical method is to examine facial expressions. Machine learning has shown tremendous

D. Sahana (✉) · K. S. Varsha · R. Priyanka

Department of ECE, Global Academy of Technology, Bengaluru, Karnataka, India

e-mail: dayanandsahana@gmail.com

S. Sen

Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

e-mail: snigdha.sen@gat.ac.in

performance in various areas like data analysis, image analysis, IoT, health care, astronomical data analysis [1–4], etc. Machine learning is a part of artificial intelligence that helps machines to understand patterns via a given dataset. Deep training is a subdivision of machine learning which is a very powerful technology capable of automatically extracting features from images or videos and handling complex data with high dimensions without human intervention. There are various deep learning algorithms like recurrent neural networks (RNNs), generative adversarial networks (GANs), multilayer perceptron (MLPs) [5], etc., to serve various purposes. One such deep learning algorithm is the convolution neural network (CNN) which is the type of artificial neural network, generally used to recognize and classify images or objects. Indeed, this algorithm is good for uprooting the characteristic of the image and is apt for image analysis subjects like image classification. CNN architecture consists of a heap of different surfaces that modifies input capacity into an output volume through a differential function. These multiple layers are made of artificial nodes in which each node gets weighted input data. The data is then passed into an activation function to output the result. CNN is proven to be an efficient recognition algorithm because of its striking features like simple structure, fewer training parameters, and adaptability. In this work, the TensorFlow framework is used to create a highly flexible CNN architecture to process pixel data and perform the task. When compared to the other deep learning algorithms, CNN has the advantage of requiring little pre-processing input. In this paper, we examine and explore how emotion recognition from facial expressions can be done using the CNN model which has numerous applications in real-life situations such as identification of driver's drowsiness, in medical research like autism therapy, and in mobile phones to automate clicking selfie. As deep learning and machine learning have shown remarkable performance in various domains [6–11] we believe using CNN is a suitable choice for this task.

The article is assembled into five sections as follows:

The literature survey has been described in Sect. 2. The dataset description goes on in Sect. 3. Our presented framework is demonstrated in Sect. 4. Section 5 discusses the experimental set-up and result of our trained model. Finally, we conclude with future work on facial emotion recognition.

2 Literature Survey

Many researchers have already published an enormous amount of information on the FER field. For example, in the late twentieth century, the value of FER was identified by Charles Darwin in the book ‘The Expression of Emotions in Man and Animals’. He mainly described emotions. Corneanu et al. fundamentally classified emotion recognition by multimodal approaches. He concentrated on methods and parameters used for emotion recognition. He focused on the classification of FER on the principles of parameterization and facial expression [12]. The first FER model was created by Matsugu et al. He operated it with the CNN model which generated robust identification and unconventional of the subject [13]. Concerning the

Matsugu CNN model, the Fasel model consists of two CNN which were used to recognize facial expressions and face identity recognition. He performed on 5600 inert pictures of ten topics and achieved an accuracy of 97.6% [14]. Anil and others developed the terse survey of the techniques; this describes emotion recognition and exact value on several databases. A succinct differentiation was made between 2 and 3D methods [15]. Mohammed et al. newly discovered an algorithm that was a combination of extreme learning machine, bilateral two-dimensional principal component analysis, and curvelet-based algorithm. He succeeded in getting a very high rate by curvelet features which gave him vast replicas, and therefore, face emotion recognition was obtained [16]. The local directional number pattern was discussed by Rivera and others. The techniques which had the competence of surpassing the rules were distinct from several systems [17]. Shan and others recommended the local binary pattern because characteristics were delivered at a fast rate with contrast to the Gabor wavelets.

Then focussing on SVM, he mainly focused on algorithms like linear discriminant analysis and template matching [18]. Yu and Zhang suggested a method that had three states of the art face sensors continued with different deep CNN models. Merging of CNN was taken place by reducing the hinge loss and log-likelihood loss [19]. Using deep neural network, emotion recognition was performed based on three architectures by Enrique Correa et al. His first architecture contained three complexity layers with two completely linked layers. Enrique Correa et al. operated second architecture with three connected layers than using two linked layers which made paced up the operation. He used three separate layers for the third architecture as a max-pooling layer, convolutional layer, and local contrast normalization; as time passed, they upgraded the third max-pooling layer to minimize the limit factor. They obtained precision for architectures of about 63%, 53%, and 63% considerably. The research shows having a minimum mesh size decreases the process of an authentic network more than anticipated values. Therefore, it concluded that the second architecture was not determined as the other two architectures [20]. For FER, Kahou and others operated on a hybrid CNN-RNN architecture. The combination of CNN and RNN was used to create a hybrid model. As the study shows, they constructed three CNN types— 3×3 frame size, 5×5 filter size with three layers, and 9×9 filter size. They merged and operated feature level and decision level as a result they obtained remarkable upgrades. The cluster of CNN, RNN, and mean of per structure deploy categorized methods was used to excel this architecture [21].

3 Description of Dataset Used

Here, we have upskilled the model by using a dataset called FER2013 which is publicly available on Kaggle. The FER2013 is a foresighted dataset, and it is frequently used in ICML and manifestos. This is one of the most difficult databases with the human point accuracy of 65% and maximal operating publish work performing efficiency of 75.2%. The dataset consists of 35,887 images that

are assigned to 48×48 pixels in monochrome. Yet FER2013 is not an established index, as it includes pictures of the seven facial interpretations with a circulation of Anger 4,953, Disgust 547, Fear 5,121, Happy 8,989, Sad 6,077, Surprise 4,002, and Neutral 6,198 [22]. The priming activity is accomplished with a fixed framework which will generate an upskill miniature that is used as a forecast criterion and is stored in a folder with the appendix. The details tutoring procedure design entering training data and attestation facts. The instruction details are refined using the CNN method on create attribute descent that will be approximate to information affirmation. Interpretation conclusion will outcome skilled exemplary architectonics to reach the extremity era rate. The CNN designs residual modules and amalgamation in a complexity layer.

4 Our Proposed Framework

Our proposed workflow is represented in Fig. 1. In this work, we explore how facial emotion recognition can be done in two ways such as.

- (I) Facial emotion recognition from a static image
- (II) Real-time facial emotion recognition through webcam.

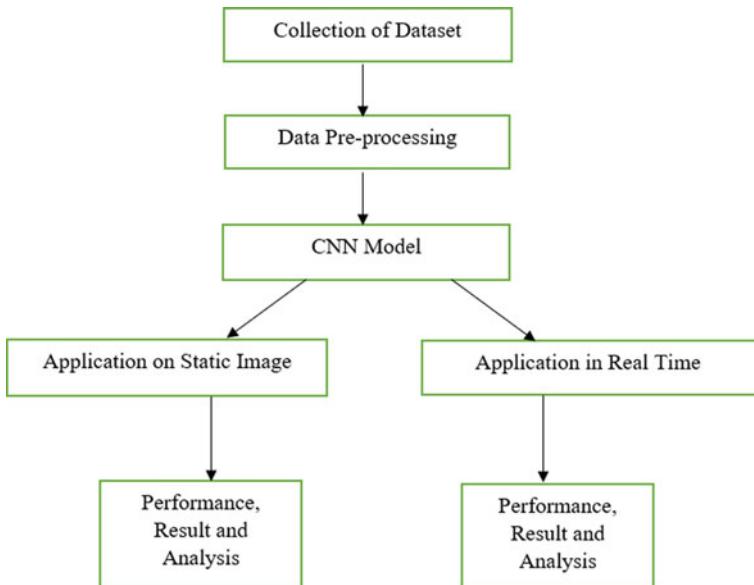


Fig. 1 Proposed workflow

5 Experimental Set-Up and Result

The whole project has been carried out using Python in Jupyter Notebook. It is a web application for making and sharing computational archives. It offers a basic, smoothed-out, and report-driven insight for executing machine learning algorithms.

I. Facial Emotion Recognition from a static image

A program is developed by using an open-source Python library called facial emotion recognizer (FER) for sentiment analysis. The primary function of any sentiment analysis is to isolate the polarity of the input (textual content, speech, facial features, and many more) and understand whether the primary sentiment presented is positive, negative, or neutral. The FER() constructor is set by giving it a multi-task cascaded convolution network which is a type of neural network to identify faces and facial expressions. When the multi-task cascade convolution network is initialized to ‘True’, the model detects a face, and when it is initialized to ‘False’, the function makes use of the OpenCV Haar cascade classifier. On program execution, it first asks for user-defined image input, and once the user gives the input of his choice, it specifies different emotions along with intensity levels in the output. The emotions are categorized into seven categories, namely ‘Fear’, ‘Neutral’, ‘Disgust’, ‘Sad’, ‘Happy’, ‘Anger’, and ‘Surprise’. Each emotion is calculated, the result is placed on a scale of 0 to 1, and finally, the dominant emotion with the highest score is displayed. In Fig. 2, we have shown the sample screenshot of user-defined image input. In Fig. 3, we have shown the representation of scores of various emotions on a scale of 0–1 and dominant emotion with the highest score displayed at the bottom.

II. Real-time Facial Emotion Recognition

A CNN model with various convolutional filters working and examining the entire feature matrix is built using functional API to carry out the dimensionality reduction. Dimensionality reduction is done to convert the high-dimensional dataset into the lesser-dimensional dataset. A set of deep neural networks is created to analyse

```
from fer import FER
import matplotlib.pyplot as plt
%matplotlib inline

test_image_one = plt.imread(input("Hi user, put an image"))
emo_detector = FER(mtcnn=True)
# Capture all the emotions on the image
captured_emotions = emo_detector.detect_emotions(test_image_one)
# Print all captured emotions with the image
print(captured_emotions)
plt.imshow(test_image_one)

# Use the top Emotion() function to call for the dominant emotion in the image
dominant_emotion, emotion_score = emo_detector.top_emotion(test_image_one)
print(dominant_emotion, emotion_score)

Hi user, put an image
```

Fig. 2 Sample screenshot of user-defined image input

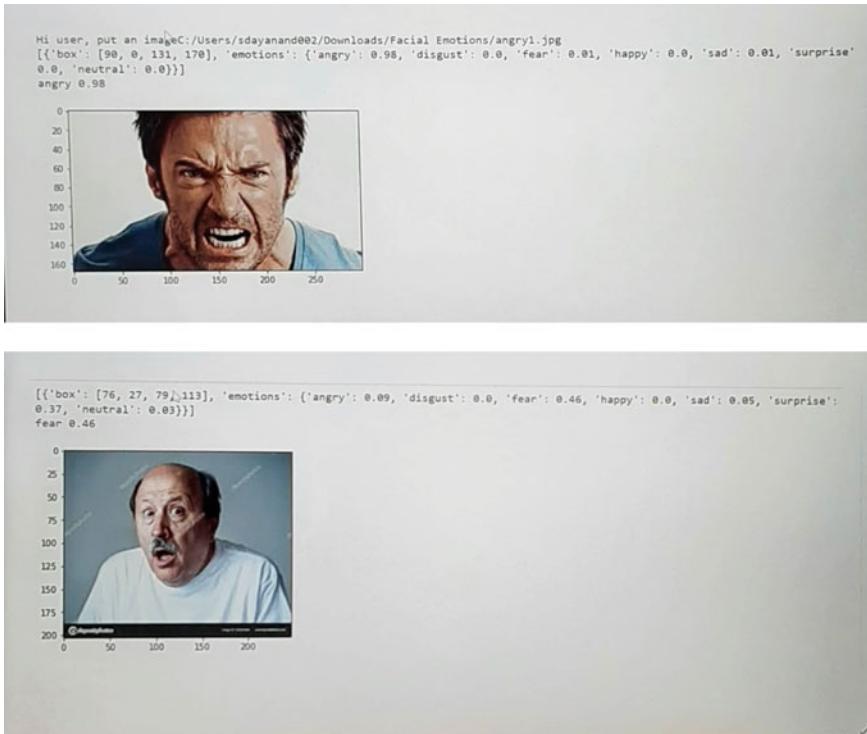


Fig. 3 Representation of scores of various emotions on a scale of 0–1 and the predominant emotion with the highest score is shown at the bottom

visual imagery. Each concurrent layer of the neural network is connected to the input neurons. Artificial neurons or nodes in CNN's accept the image pixels as input in the form of arrays. Since CNNs are feedforward networks, the progression of data takes place only in one direction, from their inputs to their outputs.

Figure 4 contains the snapshot of the loading dataset, and Fig. 5 depicts the training model with the batch size of 64 and 100 epochs.

The model is trained by using the FER2013 dataset in which emotions are classified into seven categories: 0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, and 6 = Neutral [23].

We have used 100 epochs with a batch size of 64 to train our model.

Inside the model, blocks are created using the Conv2D layer, Batch-Normalization, Max-Pooling2D, Dropout, and Flatten which are stacked together. Batch normalization is used to enhance the strength and performance of neural networks by allowing inputs with unit variance and zero means. Pooling decreases the dimensionality of characteristics while holding on to the most important data like the image hidden-layer output matrix. Dropout minimizes overfitting and voids the contribution of a few neurons towards the succeeding layer by arbitrarily not

	emotion	pixels	Usage
0	0 70 80 82 72 58 58 60 63 54 58 60 48 89 115 121...		Training
1	0 151 150 147 155 148 133 111 140 170 174 182 15...		Training
2	2 231 212 156 164 174 138 161 173 182 200 106 38...		Training
3	4 24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 1...		Training
4	6 4 0 0 0 0 0 0 0 0 0 0 3 15 23 28 48 50 58 84...		Training

Fig. 4 Loading dataset

```

Epoch 00097: val_loss did not improve from 0.94382
Epoch 98/100
448/448 - 17s - loss: 0.3218 - accuracy: 0.8816 - val_loss: 1.3699 - val_accuracy: 0.6877

Epoch 00098: val_loss did not improve from 0.94382
Epoch 99/100
448/448 - 17s - loss: 0.3126 - accuracy: 0.8828 - val_loss: 1.3140 - val_accuracy: 0.6773

Epoch 00099: val_loss did not improve from 0.94382
Epoch 100/100
448/448 - 17s - loss: 0.3112 - accuracy: 0.8848 - val_loss: 1.3259 - val_accuracy: 0.6801

Epoch 00100: val_loss did not improve from 0.94382

```

Fig. 5 Training the model using 100 epochs with the batch size of 64

renovating the weights of some nodes. Finally, Flatten is used to transform multi-dimensional input arrays into a single-dimensional long continuous linear vector to classify the image. In the end, we use the Dense layer to get the output from the preceding layer and provide the output to the next layer. We have used a Python library called Facial Emotion Recognizer to identify faces and predict emotions. With the usage of Haar cascade, the position of faces is detected and cropped. We have used OpenCV to read frames and process the image. Image augmentation is done to enhance the overall performance and capability of the model to generalize. Finally, with the execution of code, the webcam automatically turns on and is capable of predicting emotion by looking into the facial expression of a person in front of the camera or from the human face images located on the webcam which is popularly known as real-time facial emotion recognition. From Figs. 6 and 7, it is very evident that our proposed CNN model is capable of predicting human emotions from various facial expressions.

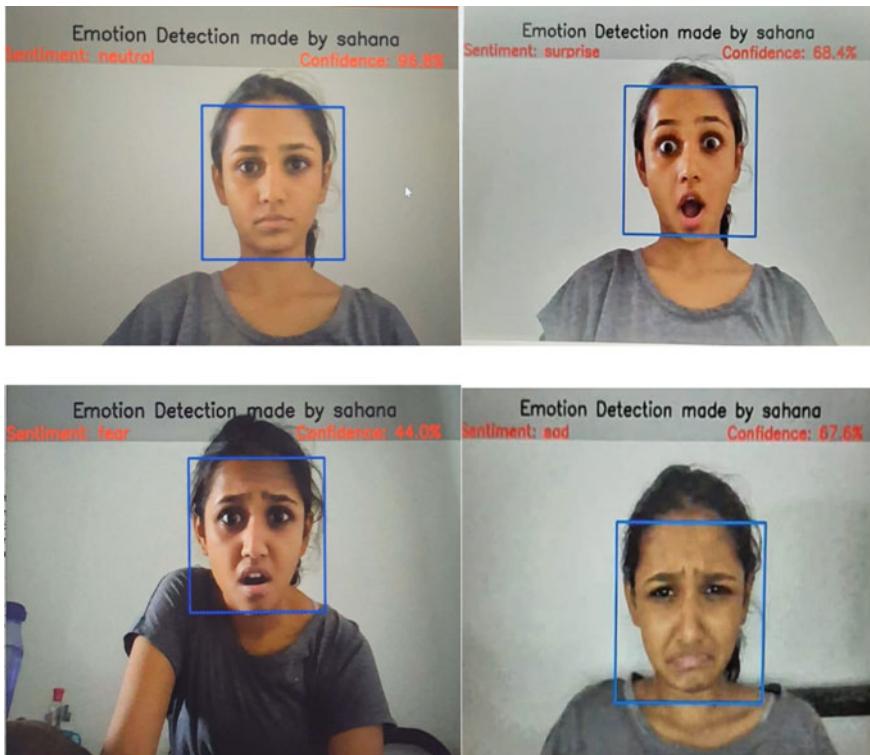


Fig. 6 Output window containing different human facial expressions along with emotion prediction and confidence

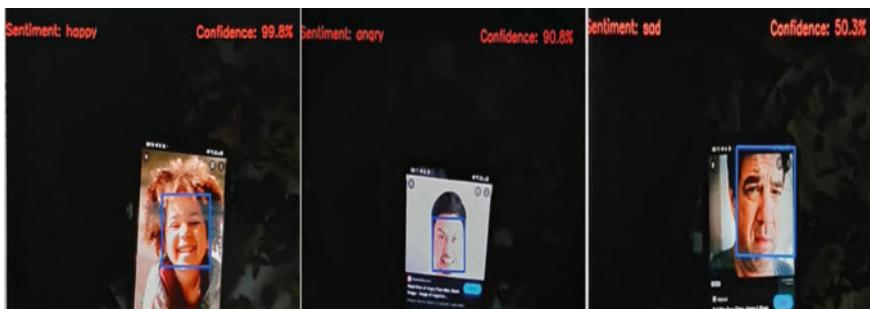


Fig. 7 Depiction of emotion and confidence predicted when images are shown to the webcam from a mobile phone

6 Conclusion

Facial expressions depict a non-verbal communication method that is vital in interpersonal relations. The results revealed above show how CNN is capable of understanding facial characteristics and detecting facial emotion. This technology can be implemented in many real-life situations such as lie detectors, mood-based learning for students, and detection of masked individuals [24]. Facial emotion recognition can be widely used in areas such as diagnosing mental illness and detecting social or physiological interactions between people. Furthermore, the study shows that facial emotion recognition could provide society with better regard and contribute to human–robot interface (HRI) interactions in the upcoming days. As a part of our future work, we plan to explore this technology in the medical field, specifically in the psychology domain, to find out emotional states based on the observation of optical and audial non-verbal signs or gestural signs. Non-verbal signs or gestural signs include voice, postural, facial, and signs displayed by a person.

References

1. Snigdha S et al (2022) Astronomical big data processing using machine learning: a comprehensive review. *Experiment Astron* 1–43. <https://doi.org/10.1007/s10686-021-09827-4>
2. Sandeep VY, Sen S, Santosh K (2021) Analysing and processing of astronomical images using deep learning techniques. In: 2021 IEEE international conference on electronics, computing and communication technologies (CONNECT). IEEE. <https://doi.org/10.1109/CONECCT52877.2021.9622583>
3. Sen S et al (2021) Implementation of neural network regression model for faster redshift analysis on cloud-based spark platform. In: International conference on industrial, engineering and other applications of applied intelligent systems. Springer, Cham. https://doi.org/10.1007/978-3-030-79463-7_50
4. Monisha R, Sen S, Davangeri RU, Sri Lakshmi KS, Dey S (2022) An approach toward design and implementation of distributed framework for astronomical big data processing. In: Intelligent systems. Springer, Singapore, pp 267–275. https://doi.org/10.1007/978-981-19-0901-6_26
5. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm>
6. Sen S et al (2021) Analysis, visualization and prediction of COVID-19 pandemic spread using machine learning. In: Innovations in computer science and engineering. Springer, Singapore, pp 597–603
7. Sen S, Singh KP, Chakraborty P (2023) Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network. *New Astron* 99:101959
8. Sen, S, Amrita I (2022) A transfer learning based approach for lung inflammation detection. In: Advanced techniques for IoT applications: proceedings of EAIT 2020. Springer, Singapore
9. Mayank K, Sen S, Chakraborty P (2022) Implementation of cascade learning using apache spark. In: 2022 IEEE international conference on electronics, computing and communication technologies (CONECCT). IEEE
10. Khasnis NS, Sen S, Khasnis SS (2021) A machine learning approach for sentiment analysis to nurture mental health amidst COVID-19. In: Proceedings of the international conference on data science, machine learning and artificial intelligence
11. Pankaj, Sen S, Chakraborty P (2022) A novel classification-based approach for quicker prediction of redshift using apache spark. In: 2022 International conference on data science, agents &

- artificial intelligence (ICDSAAI). Chennai, India, pp 1–6. <https://doi.org/10.1109/ICDSAAI5433.2022.10028971>
- 12. Corneanu CA, Simón MO, Cohn JF, Guerrero SE (2016) Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans Pattern Anal Mach Intell* 38:1548–1568. <https://doi.org/10.1109/TPAMI.2016.2515606>
 - 13. Matsugu M, Mori K, Mitari Y, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw* 16:555–559. [https://doi.org/10.1016/S0893-6080\(03\)00115-1](https://doi.org/10.1016/S0893-6080(03)00115-1)
 - 14. Fasel B (2002) Robust face analysis using convolutional neural networks. In: Proceedings of the 16th international conference on pattern recognition; Quebec City, QC, Canada, pp 40–43
 - 15. Anil J, Suresh LP (2016) Literature survey on face and face expression recognition. In: Proceedings of the 2016 international conference on circuit, power and computing technologies (ICCPCT); Nagercoil, India, pp 1–6
 - 16. Mohammed AA, Minhas R, Wu QJ, Sid-Ahmed MA (2011) Human face recognition based on multidimensional PCA and extreme learning machine. *Patt Recognit* 44:2588–2597. <https://doi.org/10.1016/j.patcog.2011.03.013>
 - 17. Rivera AR, Castillo JR, Chae OO (2013) Local directional number pattern for face analysis: face and expression recognition. *IEEE Trans Image Process* 22:1740–1752. <https://doi.org/10.1109/TIP.2012.2235848>
 - 18. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27:803–816. <https://doi.org/10.1016/j.imavis.2008.08.005>
 - 19. Yu Z, Zhang C (2015) Image-based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. Seattle, WA, USA. New York, NY, USA: ACM, pp 435–442
 - 20. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülcühre Ç, Memisevic R, Vincent P, Courville A, Bengio Y, Ferrari RC et al (2013) Combining modality specific deep neural networks for emotion recognition in the video. In: Proceedings of the 15th ACM on international conference on multimodal interaction. Sydney, Australia, New York, NY, USA: ACM, pp 543–550
 - 21. Ebrahimi Kahou S, Michalski V, Konda K, Memisevic R, Pal C (2015) ICMI '15, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM; New York, NY, USA. Recurrent Neural Networks for Emotion Recognition in Video, pp 467–474
 - 22. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH et al (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing. Springer, pp 117–124
 - 23. <https://www.analyticsvidhya.com/blog/2021/11/facial-emotion-detection-using-cnn/>
 - 24. Kumar S, Yadav D, Gupta H et al (2022) Towards smart surveillance as an aftereffect of COVID-19 outbreak for recognition of face masked individuals using YOLOv3 algorithm. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-021-11560-1>

Implementation of Artificial Neural Network for Demanufacturing Operation in the Rail Industry



Humbulani Simon Phuluwa, Ilesanmi Daniyan , and Khumbulani Mpofu 

Abstract End-of-life (EoL) component recovery is gaining popularity in a variety of industries around the world. The dominance of linear manufacturing in a variety of industries has given birth to new ideas about EoL component recirculation. End-of-life railcar component degrading treatment is expected to have an impact on disposal and resource sustainability. As a result, the goal of this study is to create a predictive model for the scheduling of demanufacturing operations activities using an artificial neural network (ANN). During the development of the predictive tool, an ANN tool with a prediction mechanism was set up to focus on the completion times of demanufacturing options. The MATLAB2018a program and the transfer function “tansig” were used to train the ANN using the backpropagation and Levenberg–Marquardt methods. Data was trained to visualize the predictability of the demanufacturing operation, and a correlation coefficient of 1 was obtained. The data set was observed to fall along the line of best fit after repeated training of the input data. This indicated that the developed approach was highly efficient, with a strong ability to predict the completion timeframes of demanufacturing operation options.

Keywords ANN · Demanufacturing · EoL · Predictive model

1 Introduction

The rail industry has advanced rapidly and becomes more sophisticated over the last 160 years. Railcar advancements have contributed significantly to the growth of the rail industry. As a result, numerous industries’ treatment of end-of-life (EoL) railcar components has degraded and been improperly applied, causing environmental problems. Demanufacturing operations and the development of a predictive model for optimal recovery are two potential methods for reducing the amount of EoL railcar components disposed of in landfills. The degradation of end-of-life railcar

H. S. Phuluwa · I. Daniyan  · K. Mpofu

Department of Industrial Engineering, Tshwane University of Technology, Pretoria 0001,
South Africa

components is expected to have an impact on their disposal and resource sustainability. The benefits of the circular economy were demonstrated in diverse books, and they addressed issues posed by linear production. However, other industries, such as the rail manufacturing industry, are falling behind in developing plans to properly execute the EoL component recovery strategy.

Products that have reached their end-of-life (EoL) are frequently abandoned to landfills. The continued degradation of the environment and lack of room in landfills could be viewed as a challenge to creating a climate-safe world. According to Merkisz-Guranowska et al. [1], the disposal of a passenger railcar is equivalent to the disposal of 36–42 passenger automobiles in terms of waste weight. The author argued that because railway vehicles are made up of a variety of materials such as ferrous and non-ferrous metals, elastomers, polymers, glass, fluids, modified organic natural materials, compounds, and electronics, the EoL rolling stock is a valuable resource whose recycling provides measurable economic benefits. The goal of the study was to provide a scientific approach to the rail industry that is based on modern methods and technology. The study also aimed to create a roadmap for effective EoL recovery in the rail industry and other manufacturing sectors.

The paper is structured as follows. Methodology identifies methods and tools used in the study. The development of demanufacturing operation activities schedules is to critically understand the schedules patterns of recovery of EoL. The development predictive modeling is to strengthen the possibility of eliminating uncertainty of demanufacturing operation activities. The conclusion identifies the study impact toward stakeholders.

2 Methodology

To derive an understanding of demanufacturing operations activities in the recovering of EoL components, various literature and rail industries unpublished documents were collected and documented. This paper is arranged by various categories.

The key categories involved the following:

1. Formulation of the problem
2. Data gathering
 - Rail industry case study
 - Literature
3. Development of demanufacturing operation activities schedules and predictive model
4. Predictive modeling analysis for the recovering of EoL components.

2.1 Formulation of the Problem

The purpose of this work was to develop a demanufacturing operation schedule model to aid rail industry management and operators in recovering railcar components at their EoL. This method was used to aid in the identification and classification of demanufacturing operations into recovery schedule completion times. The publication raised a question to unpack the contribution of the study to the body of knowledge due to the adoption of artificial intelligence in demanufacturing operations. The following research topic was posed:

- Can computational approaches help EoL components recover more quickly?

The data collection method was conceptualized to determine data gathering instruments in order to answer the study topic.

2.2 Data Gathering Process

Different issues informed the data collection procedure, including knowledge on demanufacturing operations, the scheduling process, and railcar components. To achieve its main goal, the method relied on a large number of documents. The data sources used in this investigation are highlighted in Fig. 1.

Figure 1 depicts the primary sources that contributed to the study. The sources are scientifically linked with the goal of achieving the study goal. To define the problem, the data sources relevant to this study, as depicted in Fig. 1, were exhausted. The materials, methods, and observations on the rail industry workshops were thoroughly reviewed. The research question was developed, and the variables were identified.

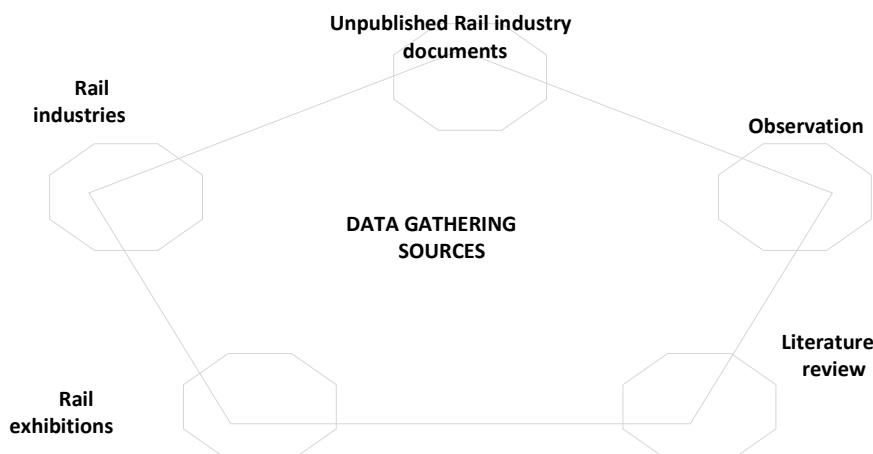


Fig. 1 Data source for the study

The PERT design was created, and the results were used to help develop the ANN model. The results of the ANN model were examined, and conclusions were drawn. Figure 1 sources were investigated in greater depth.

2.2.1 Rail Industries

The rail business has a wide range of structures and functions. The study, on the other hand, focused on rail users and people who maintain railcars. This was done to assist rail users in determining the best recovery procedures for EoL components. The main focus on rail users was due to a large number of components that reach end-of-life status regularly and may be readily disposed of in landfills. This research was based on the rail users in South Africa.

2.2.2 Unpublished Rail Industry Documents

To minimize material being duplicated for public consumption, the data gathered came from processes that had not yet been disclosed. Demanufacturing operations and scheduling times are included in the document. The goal of gathering this information was to gain a better understanding of the demanufacturing processes and the need for a speedy turnaround on recovering end-of-life rail components.

2.2.3 Observation

The observation was used to learn how the recovery processes were carried out in order to achieve effective EoL component recovery strategies. The focus of this research was on the actual demanufacturing process, with the goal of finding components that are frequently dumped into landfills or open spaces. In-depth observations on the scheduling of demanufacturing operations were analyzed.

2.2.4 Rail Exhibitions

The rail exhibits were utilized to establish direct contact with rail manufacturers from throughout the world. This data collection procedure was created to learn more about product designs and life spans. The purpose of this data was to better understand the train manufacturing process.

2.2.5 Literature

In many aspects, the recovery of EoL components themes is numerous. The diversity in EoL component recovery necessitates a thorough investigation from operating to

disposal to landfills. The phrase “end-of-life” is used in this study to describe when a product no longer meets the needs of its last user [2]. According to Paterson et al. [3], a product that has reached the end of its useful life (EoL) can be remanufactured, reconditioned, mended, or reused, or it can be recycled to make a new product. Recovering valuable components from end-of-life (EoL) product is regarded as a means to extend remaining useful life and reduce production cost of used components in the context of remanufacturing [4]. There are many different definitions of EoL in the literature, but one thing that they all have in common is the lack of function after its life span.

To recover EoL components, a new paradigm in demanufacturing design and execution of recovery operations is critical. Material recycling, part reuse, shredding, and landfill choices are all parts of the demanufacturing process; hence, the attempts to salvage any remaining economic value in the EoL product through disassembly promote material recycling over disposal [5]. Steel and aluminum, according to Mistry et al. [6], have good strength, formability, and weldability features, as well as a low cost, making them a versatile option. However, the trains feature a variety of steel and aluminum components that require careful planning for recovery. According to Muvunzi et al. [7], when replacing a broken spare part, considerable lead times may be encountered. Due to tooling expenses, some of the parts are very expensive to recover using traditional methods.

The proliferation of product designs, the difficulty of obtaining product feature and material composition information, and the lack of integration of collecting and demanufacturing processing have all hampered recycling automation for selective disassembly to date [8]. According to Karakayali et al. [9], by removing the components that can be reused, demanufacturing reclaims the economic and environmental value contained in end-of-life or end-of-use (i.e., used) products (e.g., as-is, remanufactured, or recycled).

To support effective and efficient recovery operations, the intelligence decision is required. Artificial intelligence is defined as the study of having computers do things that require intelligence in humans, such as inference based on knowledge, reasoning with uncertainty, various forms of perception, learning, and applications to issue prediction, classification, and optimization [10]. According to Mohamed Na et al. [11], an artificial neural network (ANN) worked as an approach to streamline input–output relationships to obtain an estimated anticipated function. The authors further state that a well-trained artificial neural network is used as an optimization model for engineering applications. However, various algorithms influence the effectiveness of ANN. Some of the algorithms that literature identified are as follows: Levenberg–Marquardt algorithm, gradient descent algorithm, and conjugate gradient algorithm. According to Lv et al. [12], weight and bias variables can be modified in the training utilizing the Levenberg–Marquardt (LM) method, and backpropagation is used to generate the Jacobian matrix of the performance function with respect to the weight and bias variables. The gradient-based method is the most straightforward training algorithm for feeding multilayer perceptron networks [13]. Hence, the conjugate gradient methods do not necessarily produce the fastest convergence due to adjusting of weights in the steepest descent direction [14]. According to Bilski

et al. [15], the LM algorithm can adjust the speed of training based on the shape of the error function. Based on the shape of the error function, the LM algorithm can alter the training speed. In the training of data, the LM capacity positioned some beneficial characteristics. An evolutionary computation can be created by combining an artificial neural network (ANN) and a genetic algorithm to solve an optimization model [16]. Other optimization models show how both economic and environmental impacts help in the recovery of the EoL products. Shokohyar et al. [17] built a decision-making methodology that concurrently optimizes the service period and recovery possibilities. However, there are no demanufacturing operation models in the optimization models for trade-off considerations between efficient EoL recovery methods in terms of demanufacturing operation time and cost implications.

ANN can reduce pattern information when applied to new data, tolerate noisy inputs, and give reliable and realistic estimates [18, 19]. More training data enhances the classification model, whereas more test data helps to precisely estimate error [20]. Machine learning algorithms for regression modeling of tiny datasets (less than ten observations per predictor variable) are still scarce [21]. According to Bagshaw [22], effective project management necessitates optimizing the project's duration in order to reduce the overall project time and expense.

In comparison with Project Evaluation Review Technique (PERT) or Critical Path Method (CPM) charts, Petri Nets are more powerful models [23]. The author showed Petri Nets display similarities with PERT/CPM on scheduling limitations and can be easily converted from a PERT/CPM network. However, PERT has demonstrated capability in dealing with time horizon is unknown. The recovery of EoL components requires proper planning on the operation time horizons for the completion of the project. PERT is event oriented, probabilistic, and time focused, and it applies to projects with an undetermined time horizon [22].

2.3 *Development of Demanufacturing Operation Activities Schedules*

In this scenario, the demanufacturing processing operations were divided into the days required to complete the EoL recovery activities. The rail industry was used as a case study in the research to create a table of generic demanufacturing operation processing processes. The tasks involved in recovering EoL components through demanufacturing operations are depicted in Table 1. To define the critical path of demanufacturing activities, a precedence diagram was used to critically analyze work flow of EoL components.

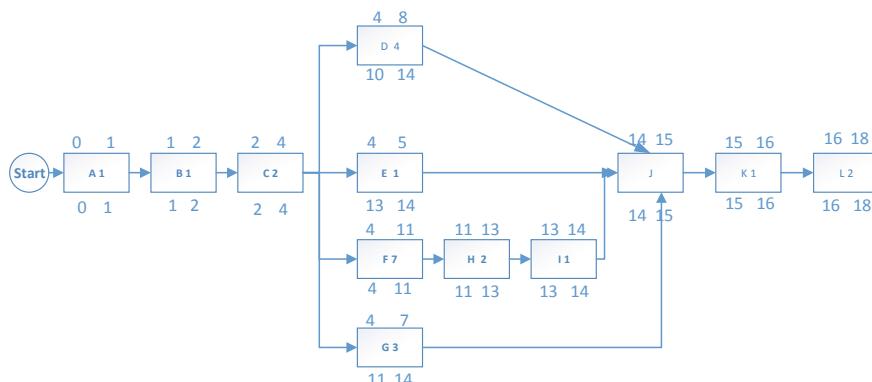
Figure 2 shows the precedence diagram of demanufacturing processes for railcar bogie.

A-B-C-F-H-I-J-K-L is the critical path. D, E, and G are the non-critical paths. This analysis' float is as follows:

- EST—earliest starting time, LST—latest starting time, EFT—earliest finish time

Table 1 EoL recovery activities for railcar bogie

Activities	Precedence	Description	(Days)
A	–	The railcar bogie is positioned correctly for disassembly	1
B	A	Inspection	1
C	B	Disassembly process	2
D	C	Maintenance	4
E	C	Reuse	1
F	C	Remanufacturing	7
G	C	Recycling	3
H	F	Assembly	2
I	H	Final assembly	1
J	D, E, G, I	Final inspection	1
K	J	Activities of evaluation and testing	1
L	K	Completing all required paperwork	2

**Fig. 2** Precedence diagram of demanufacturing processes for railcar bogie

- LFT—latest finish time, CT—completion time, DD—due dates

The critical path is essential to the demanufacturing process and is central to it. Any disruption in the continuity of those routes, on the other hand, will affect the efficacy of demanufacturing operations. Demanufacturing operations are delayed because of the path breakup, resulting in fines for the demanufacturer. As a result, the variance in this path must be reduced to zero to avoid being late.

This means that the demanufacturing operation's pathways are continuous, with no breaks in between as highlighted in Table 2. Over the years, manufacturers and recyclers have relied on human expertise to determine the completion dates for the disassembly process, but with the increasing complexity of the railcar bogie, this is becoming increasingly unreliable. The use of an artificial intelligence (AI) system,

Table 2 Float analysis for demanufacturing operation activities

Activities	LST	EST	Floats	EFT a	LFT b	Duration (m)	Expected time (E)	Variance (V)
A	0	0	0	1	1	1	1	0
B	1	1	0	2	2	1	1.33	0
C	2	2	0	4	4	2	2.66	0
D	4	10	6	8	14	4	6.33	1
E	4	13	9	5	14	1	3.83	2.25
F	4	4	0	11	11	7	8.33	0
G	4	11	7	7	14	3	5.5	1.36
H	11	11	0	13	13	2	5.66	0
I	13	13	0	14	14	1	5.33	0
J	14	14	0	15	15	1	5.66	0
K	15	15	0	16	16	1	6.00	0
L	16	16	0	18	18	2	7.33	0

on the other hand, will provide a scientific basis for making decisions about due dates and completion times. The decision was made to use ANN because it is data driven and can consistently train and use previous data to predict future events, which is consistent with the data revolution and the concept of Industry 4. 0.

The values of the expected time and variance were computed in Table 2 using Eqs. 1 and 2, respectively.

$$E = \frac{a + 4m + b}{6} \quad (1)$$

$$V = \left(\frac{b - a}{6} \right)^2 \quad (2)$$

According to the training function graph in Fig. 3, the ANN performed 153 iterations out of 300 before meeting the performance goal of parameter correlation after 53 s. The input and output were used for each loop. There were 48 inputs and 48 outputs for each loop. The iteration is carried out by the software using the entered algorithm. When the performance goal is not met, the software automatically adjusts the weights and biases and repeats the iteration until the goal is met. The algorithm is run 300 times to determine whether the performance requirements will be met. The performance was achieved after 153 iterations. However, if the performance goal was not fulfilled, the training function graph will change. The ANN using the transfer function “tansig” used the backpropagation method and Levenberg–Marquardt algorithms to train the input parameter in a supervised learning environment using the MATLAB 2018 program. Before the associated response in the form of the output was obtained, the input was passed through four layers. Figure 3 presents the training

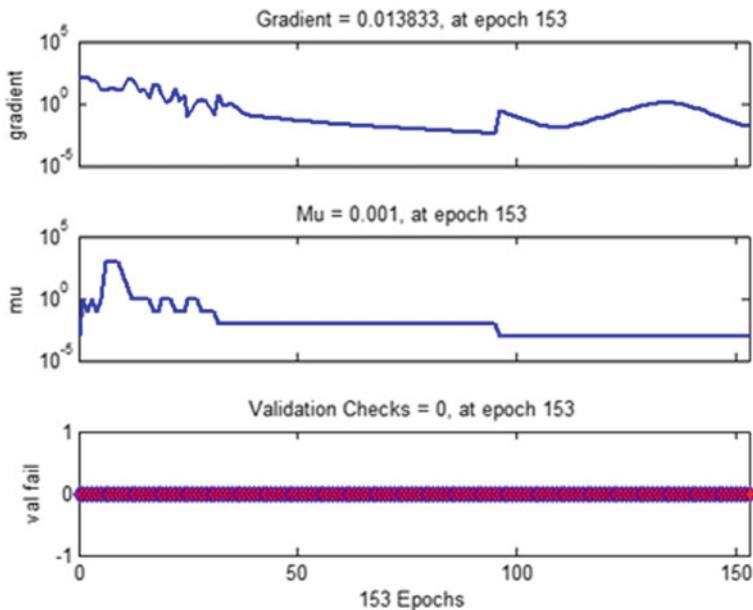


Fig. 3 Training gain plot

gain utilized to train the network. This is determined by the input correlation matrix's highest eigen value, and the value is impacted by the convergence error.

After 153 iterations, the values of training gain Mu and gradient were 0.001 and 0.013833, respectively (Fig. 3). The ANN is given four sets of variables as inputs, each with a corresponding output. For the first set of iterations, the first target-output demonstrates a lack of good degree of agreement between the data points and the line of best fit.

3 Results and Discussion

The training and mean square error (MSE) plots are shown in Fig. 4. When the training line crosses the goal horizontal line as depicted, the performance target is considered to be met. This indicates that the ANN has been properly trained and correlated with the input data in order to make predictions. For the 153 iterations, the MSE was 0.003. The MSE's insignificant value shows that the input variables are highly correlated.

Equations 3–6 express the gradient equation for the predictive line for EFT, LFT, CT, and DD, respectively.

$$y = (1)T + (0.00014) \quad (3)$$

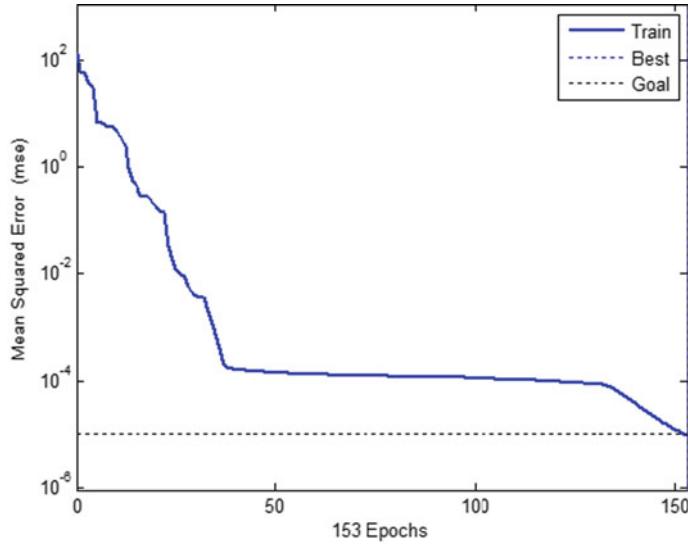


Fig. 4 Training and MSE plot

$$y = (1)T + (0.00012) \quad (4)$$

$$y = (1)T + (0.00019) \quad (5)$$

$$y = (1)T + (0.00005) \quad (6)$$

The target-output for the EFT, LFT, CD, and DD prediction shows the degree of agreement between the data points and the line of best fit. The correlation coefficient was calculated to be 1 which further validates the agreement between the target and the output from the ANN model [24, 25].

In a MATLAB2018b environment, the values of latest start time (LST), earliest start time, processing time, and float were entered as input parameters and trained. The ANN's output targets were the values of earliest finish time (EFT), latest finish time (LFT), completion time (CT), and due dates (DD). Time is crucial in positively influencing demanufacturing operations. To demonstrate the significance of the demanufacturing operation completion status, a comparative analysis of actual and ANN predictive results was performed. Table 3 shows various comparisons of datasets used in this study.

Table 3 shows the agreement between real values and ANN projected values for demanufacturer/rail industry decision-making in the recovering of EoL components demanufacturing operation in the rail industry. The correlation in the four data sets demonstrates that ANN can help the rail industry or other industries make informed decisions when recovering EoL components using known data. The results also show

Table 3 Real vs ANN predicted results

Earliest finish time		Latest finish time		Completion dates		Due dates	
Real	ANN predicted	Real	ANN predicted	Real	ANN predicted	Real	ANN predicted
1	1.0001	1	1.0001	1	0.9999	1	0.9999
2	2.0003	2	2.0002	2	2.0000	1.33	1.3299
4	4.0003	4	4.0001	3	3.000	2.66	2.6600
8	7.9998	14	13.9998	4	3.9998	6.33	6.3303
5	4.9999	14	13.9998	5	5.0000	3.83	3.8302
11	11.0000	11	11.0002	6	5.9991	8.33	8.3301
7	6.9998	14	13.9999	8	8.000	5.5	5.5002
13	13.0001	13	12.9998	10	9.9996	5.65	5.6601
14	13.9953	14	14.0029	13	13.000	5.33	5.3330
15	15.0120	15	14.9924	18	17.9998	5.66	5.6525
16	15.9911	16	16.0057	22	22.000	6.0	6.0053
18	18.0001	18	17.9992	29	28.9955	7.33	7.3294

that with a small amount of data, the expectation can be more accurate; this study used 48 data sets. Time and cost are the best ways to explain a proposed practical application of the ANN predictive model.

If the demanufacturing operation is in good condition, it means that the recovery of EoL components was completed on time and within budget. Nonetheless, when the demanufacturing operation condition is negative, it means that the recovery of EoL components will take longer than expected and will incur higher recovery costs. It is important to note that in a circular economy, the rapid responsiveness of the EoL component's recirculation process is critical. The predictive tool developed is intended to reduce incoherent scheduling and uncertainty caused by the demanufacturing operation.

4 Conclusion

PERT and ANN-based models were used in this study to evaluate and predict demanufacturing operation schedules for recovering EoL rail components. Training of data was conducted to visualize the predictability of the demanufacturing operation, and a correlation coefficient of one was obtained. The objective achieved after adequate iterative training of the input data was observed to fall along the line of best fit. This indicated that the developed approach is highly efficient with a good prediction ability for the demanufacturing operation options completion times. These findings indicate that the proposed ANN model on demanufacturing operation schedules is a viable alternative to traditional estimation models with acceptable accuracy.

In essence, demanufacturing is a recovering operation that aims to economically recover materials and components from end-of-life products, with a focus on the extent of disassembly required to recover value from retired goods. For instance, this sort of scheduling has the practical effect of forcing demanufacturing operation practitioners to develop tools for forecasting schedules and making decisions based on data analysis. The proposed model's quick turnaround on recovering EoL component earnings and boosting competition and innovation in the direction of Industry 4.0 could be useful to business owners.

The study presented a clear predictive mechanism required for rail demanufacturing operations planning and execution. In the recovery of EoL railcar components, the use of ANN in scheduling prediction eliminates bias in decision-making. For successful planning and execution of EoL rail component recovery, the combination of PERT and ANN creates the uniqueness of decision-making in a high-uncertainty environment. The data points represented on the ANN graphs strongly suggest that operation planning in the recovery of EoL can be carried out immediately.

In conclusion, this study laid the groundwork for new concepts and theories required in the recovery of EoL railcar components in the age of Industry 4.0. The proposed predictive tool can be used in the future to focus cost estimation relevant to demanufacturing operation activities in the exploded view of the EoL railcar components in the circular economy.

Acknowledgements Funding: The authors disclosed receipt of the following financial support for the research: Technology Innovation Agency (TIA) South Africa, Gibela Rail Transport Consortium (GRTC), National Research Foundation (NRF grant 123575), and the Tshwane University of Technology (TUT).

References

1. Merkisz-Guranowska A, Merkisz J, Jacyna M, Pyza D, Stawecka H (2014) Rail vehicles recycling. *WIT Trans Built Environ* 135:425–436
2. Lamerew YA, Brissaud D (2019) Circular economy assessment tool for end of life product recovery strategies. *J Remanuf* 9:169–185
3. Paterson DA, Ijomah WL, Windmill JF (2017) End-of-life decision tool with emphasis on remanufacturing. *J Clean Prod* 148:653–664
4. Jiang Z, Wang H, Zhang H, Mendis G, Sutherland JW (2019) Value recovery options portfolio optimization for remanufacturing end of life product. *J Clean Prod* 210:419–431
5. Johnson MR, McCarthy IP (2014) Product recovery decisions within the context of extended producer responsibility. *J Eng Tech Manage* 34:9–28
6. Mistry P, Johnson M, Galappaththi U (2021) Selection and ranking of rail vehicle components for optimal lightweighting using composite materials. *Proc Inst Mech Eng Part F: J Rail Rapid Transit* 235:390–402
7. Muvunzi R, Mpofu K, Daniyan I, Fameso F (2022) Analysis of potential materials for local production of a rail car component using additive manufacturing. *Helijon* 8(e09405):1–8
8. Williams JAS (2007) A review of research towards computer integrated demanufacturing for materials recovery. *Int J Comput Integr Manuf* 20:773–780

9. Karakayal I, Emir-Farinas H, Akçal E (2010) Pricing and recovery planning for demanufacturing operations with multiple used products and multiple reusable components. *Comput Ind Eng* 59:55–63
10. Munakata T (2008) Fundamentals of the new artificial intelligence: neural, evolutionary, fuzzy and more, Springer Science and Business Media
11. Mohamed Na, M AS, Abd Wahab D, Abdullah S, Tihth RM (2011) Development of artificial neural network for optimisation of reusability in automotive components. *J Appl Sci* 11:996–1003
12. Lv C, Xing Y, Zhang J, Na X, Li Y, Liu T, Cao D, Wang F-Y (2017) Levenberg–Marquardt backpropagation training of multilayer neural networks for state estimation of a safety-critical cyber-physical system. *IEEE Trans Industr Inf* 14:3436–3446
13. Wang Y, Kim SP, Principe JC (2005) Comparison of TDNN training algorithms in brain machine interfaces. In: Proceedings. 2005 IEEE international joint conference on neural networks. IEEE, 2459–2462
14. Seiffert U (2006) Training of large-scale feed-forward neural networks. The 2006 IEEE international joint conference on neural network proceedings, IEEE, 5324–5329
15. Bilski J, Kowalczyk B, Marchlewska A, Zurada JM (2020) Local Levenberg–Marquardt algorithm for learning feedforwad neural networks. *J Artific Intell Soft Comput Res* 10(4):290–316
16. Wahab DA, Amelia L, Hooi NK, Cheharsu C, Azhari C (2008) The application of artificial intelligence in optimisation of automotive components for reuse. *Mater Manuf Eng* 31(2):595–601
17. Shokohyar S, Mansour S, Karimi B (2014) A model for integrating services and product EoL management in sustainable product service system (S-PSS). *Intell Manuf* 25:427–440
18. Walczak S (2007) Neural networks in organizational research: applying pattern recognition to the analysis of organizational behavior. *Organ Res Methods* 10:710
19. Daniyan IA, Mpofu K, Tlhabadira I, Ramatsetse BI (2021) Process design for milling operation of titanium alloy (Ti6Al4V) using artificial neural network. *Int J Mech Eng Robot Res* 10(11):601–611
20. May RJ, Maier HR, Dandy GC (2010) Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Netw* 23:283–294
21. Shaikhina T, Khovanova NA (2017) Handling limited datasets with neural networks in medical applications: a small-data approach. *Artif Intell Med* 75:51–63
22. Bagshaw KB (2021) PERT and CPM in project management with practical examples. *Am J Oper Res* 11:215–226
23. Boushaala AA (2014) An approach for project scheduling using PERT/CPM and petri nets (PNs) tools. *Indus Eng Oper Manage* 5:939–947
24. Daniyan IA, Bello EI, Ogedengbe TI, Mpofu K (2020) Use of central composite design and artificial neural network for predicting the yield of biodiesel. *Procedia CIRP* 89:59–67
25. Daniyan IA, Tlhabadira I, Mpofu K, Adeodu AO (2020) Development of numerical models for the prediction of temperature and surface roughness during the machining operation of titanium alloy (Ti6Al4V). *Acta Polytech J* 60(5):369–390

Diversified Recommendation Generation Using Graph Convolution Neural Network



Naina Yadav

Abstract Many methods have been proposed for recommendation generation using graph neural networks (GNNs). The advantage of using GNN in a recommendation system is learning the structural information of the user and item and their interaction more efficiently than traditional learning techniques. Most of the proposed models for recommendation generation are concerned only about their accuracy enhancement. Besides accuracy, novelty, diversity, and serendipity in the recommendation are often desirable for a better user experience in a real-world application. Earlier diversity in the recommendation system is achieved using the re-ranking algorithms. These approaches often compromise with accuracy to include diversity in the recommendation. Here, we proposed a methodology for diversity inclusion in the recommendation system using the GNN. We proposed a method based on Cluster-GCN for diversification of the recommendation. In our proposed method, we cluster users' nodes based on their dissimilarity, and further, their subgraph is used for their neighborhood-based representation learning using graph convolution neural network (GCN). The novelty of the work is the clustering for the user's pre-trained diversity enhancement in the recommendation generation. The proposed diversified cluster graph convolution neural network (Div-ClusGCN) model is trained for diversified recommendation generation. We achieved around 7% more diverse recommendations from the other state-of-the-art models.

Keywords Diversity · Clustering · Recommendation system · Graph neural network

N. Yadav (✉)

School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India

e-mail: nainayadav585@gmail.com

1 Introduction

Recommendation systems are proposed for the information overload problem. In today's era, lots of information are now available online. From the vast collection of information, selecting relevant and appropriate information is a tedious task. The recommendation system has been proposed to overcome these information overload problems. There are lots of algorithms presented for the efficient recommendation generation model. The recommendation generation is based on these three algorithms: collaborative filtering (CF), content-based (CBF), and context-based recommendation. The CF-based methods are the most widely used among these algorithms for recommendation generation. These algorithms are only concerned with the model's accuracy where the recommendation generated by these models is monotonous [1, 2]. To overcome these issues, diversity is introduced in the recommendation system. The overall goal of these approaches is to include those items in the recommendation list which are new for the target user as well as relevant. There are lots of techniques that have been proposed for diversified recommendation generation. The proposed diversification methods are based on two approaches: The first is the re-ranking methodology, and the next one is to include the regularization factor to include the more diverse items in the recommendation list. In the re-ranking approach, first step is recommendation generation, and then, we re-arrange the recommendation list based on the dissimilarity of the items. The re-ranking methods are more like a post-processing method for diversified recommendation generation, which only shuffles the list based on their dissimilarity, which is not user-specific. So to overcome this issue, we introduced a GNN-based method for training our model for diverse recommendation generation [3]. We proposed a strategy based on the cluster-based GCN approach, which first clusters users based on their dissimilarities and then applies GCN for structural representation of the user-item interaction [4]. The Light-GCN method is proposed, including the GCN model's essential components. The user and item representation in the LightGCN model are learned with the help of the user-item interaction matrix using the weighted sum of the embedding learned to form all layers of the GCN model and aggregated into final embeddings [5]. The primary concern for the proposed methods is the accuracy of the recommendation model, but in today's era, only accuracy is insufficient for efficient recommendation generation. To overcome this issue in recommendation system, diversity is being introduced for relevant and effective recommendation generation [6, 7]. In this proposed methodology, we try to include diversified recommendations generation for the target user using the GCN model. The proposed model uses the Cluster-GCN model for diversity enhancement in recommendation generation [8]. Earlier approaches follow the post-processing methods for diversified recommendation generation. The drawback of these methods is that they only re-rank the items based on their dissimilarity value and do not provide item domain exposure for the target user. We tried to train our model for diverse recommendation generation to overcome this issue. In our proposed model Div-ClusGCN, we first apply a clustering algorithm to cluster diverse users based on their dissimilarity values. Then, we use the cluster GCN method to generate the final recommendation.

2 Related Work

This section will briefly discuss the current work on the diversification of the recommendation system and GCN-based methods for recommendation generation, which are most relevant to this proposed methodology.

2.1 Diversity in Recommendation System

In the recommendation system, diversity is introduced to overcome the over-fitting problem of recommendation generation, which occurs when we recommend items to target users similar to their past interaction history. The diversification in recommendations will also benefit user satisfaction. In the recommendation system, diversity is introduced by K Bradley, and B Smyth, as a new method for diversity inclusion using a two-step procedure or a re-ranking method [9]. Next, diversity is introduced in the music recommendation system by Stanley et al. in terms of their musical diversity calculated by the music genre they listen to [10]. Zhang et al. proposed a method for diversification using the trust region problem. This proposed method enhances the recommendation diversity with the retrieved recommendation list while maximizing its similarity to the target query as a binary optimization problem [11]. The formula for diversity calculation is as follows:

$$\text{Diversity}(L_1, L_2, N) = \frac{\frac{L_2}{L_1}}{N} \quad (1)$$

where L_1, L_2 is for ranked list generated by CF algorithm and N is the total number of item in the set. Cui et al. proposed a method for diversity in the recommendation system using the new probabilistic genetic operator. They introduced two different objective functions for inclusion of the topic diversity in recommendation generation. It is mainly used to measure the ability of the recommendation algorithm to recommend different topic types of items [12]. Hu, Rong, et al. propose an approach based on a user study that was conveyed to analyze an organization interface, which clubs recommendations into classes, with a standard list interface to perceived categorical diversity. They calculate diversity by a survey conducting between 20 participants [13]. Vargas, Saúl, et al. proposed a methodology based on binomial framework for genre diversity in recommender systems. They also propose an efficient greedy optimization technique to optimize binomial diversity [14]. Hu, Liang, et al. stated that recommendation generation has diversified by using session contexts information with personalized user profiles. The author uses session-based wide-in-wide-out networks that are intended to efficiently learn session profiles across a large number of users and items [15]. Karakaya et al. proposed diversification using re-ranking algorithms that are utilized to aggregate diversity using the ranked list of recommendations [16]. Möller et al. use topic diversity in news recommendations

using different diversity metrics in social science and democracy news [17]. Most of the researchers also focus on the other aspect of diversity in a recommendation, which includes serendipity and accuracy and their effects on diversity. Kotkov et al. proposed a serendipity-oriented, the greedy re-ranking algorithm which improves serendipity of recommendations using feature diversification [18, 19]. Apart from a tradeoff between diversity and other metrics of recommendation, Matt, Christian et al. described different types of diversity in the recommendation. The author proposed algorithmic recommendation diversity, perceived recommendation diversity, and sales diversity and identified different recommendation algorithms and user perception effects on sales recommendation [20].

2.2 *GCN-Based Recommendation System*

Unlike traditional collaborative filtering methods, GCN has become a new state-of-the-art recommendation method and has been widely used because it can capture the high-order similarity of nodes in the interaction graph. Zhang et al. proposed a method GraphRfi to perform user representation and then jointly optimize the model for robust rating prediction and fraudster detection. In the proposed model GraphRfi, GCN is used to precisely capture user preference using node structural information, and the neural random forests achieve satisfying accuracy in fraudster detection [21]. Feng et al. proposed a model AGCN using attention on the GCN model to model users' past interactions and implicit information. The proposed AGCN model uses the Chebyshev polynomial graph filters for complexity reduction as it uses the attention layer to encourage structural graph representation, which improves recommendation accuracy [22]. Zheng et al. proposed a model for diversification in recommendation systems using GCN. They perform a rebalanced neighbor discovering, category-based negative sampling, and adversarial learning for recommendation generation. This method also alleviates the accuracy-diversity tradeoff in recommendation generation [23]. Another model based on GCN is proposed by Ying et al. [24]. They proposed a Web-based pin recommendation model which combines efficient random walks and GCN to generate embeddings of nodes, and then, they incorporate that information for recommendation generation.

3 Preliminaries

Suppose there are a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ which consist of nodes and edges where $|\mathcal{V}| = n$ is the total number of nodes present in the graph \mathcal{G} and $|\mathcal{E}| = m$ is the total relation or edge present in the graph \mathcal{G} . The edge in the graph \mathcal{G} represents the rating relation of the user and item. The \mathcal{A} represents the adjacency matrix of the user and item. The adjacency matrix \mathcal{A} is a $n \times n$ matrix where the entry contains information of the edge between two users i and j . If edge presents between i and j , then the

corresponding entry is 1, else the value will be 0. The goal of the GCN model is to learn the feature representation of each node in f -dimensional space. The feature representation for each node is represented as $X \in R^{n \times f}$. Further, we use l -layer GCN for effective representation of the nodes by smoothing features over the graph. To achieve this, we perform l -layer graph convolution iteratively so that we aggregate the features of the neighbors nodes for the target node. The GCN will aggregate target node's embedding for all neighbor node in the form from the $l - 1$ layers.

$$(Z)^l = A' \cdot (X)^l \cdot (W)^l, \quad (X)^l = \sigma(Z^{l+1}) \quad (2)$$

where $X^l \in R^{n \times F^l}$ is the embedding at the l -th layer for all the N nodes and $X^0 = X$ and A' is the normalized adjacency matrix, $W^l \in R^{F^l \times F^{l+1}}$ is the feature transformation matrix.

4 Proposed Methodology

The former section demonstrates Div-ClusGCN for diversified recommendation generation. In this, we present the effective model based on Cluster-GCN to effectively generate the recommendation for the target user, which are diverse and relevant. The detailed methodology of the proposed Div-ClusGCN is described in the following subsections.

- Bipartite graph
- Clustering algorithm
- Graph convolution
- Layer aggregation for diverse recommendation generation.

In the first step, we will preprocess our dataset for the user-item bipartite graph where users and items are the vertices information, and the their rating relation will be the edge information. The edge relation will be based on the binary matrix; if the user has rated the item, then it will be one, or the edge exists between the user and the item; if the user has not rated that item, then the edge will not be there between a particular user and item. The proposed model's goal is to give diverse and relevant recommendations to the target user. We first use the interaction data, which can be represented as a user-item bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$. If the interaction between user and item presents, then $R_{ui} = 1$, else $R_{ui} = 0$. The user-item interaction matrix consists total n users and m items. The adjacency matrix for these user-item graph is represented as follows:

$$\mathcal{A} = \begin{pmatrix} 0 & R \\ R^T & 0 \end{pmatrix} \quad (3)$$

Next, we partition this \mathcal{G} into c groups where we use clustering based on the dissimilarity of the users. The c groups for the graph \mathcal{G} into $\mathcal{G} = (G_1, G_2, \dots, G_c)$.

Every group G_t contains the user and items into subgroup information and their interaction details. The details of the user and item for each subgroup are clustered into their user-item interaction information.

$$\mathcal{G} = (G_1, G_2, \dots, G_c) = \{(V_1, E_1), (V_2, E_2), \dots, (V_c, E_c)\}. \quad (4)$$

4.1 Clustering Algorithm

In our proposed method, we use a clustering algorithm to create the user-item bipartite graph subgraph. The overall goal of this clustering of users is based on their dissimilar users. In general Cluster-GCN, the nodes are clustered randomly, but in our proposed method, we follow the concept of dissimilarities instead of following the basic collaborative filtering concept. We transform the user-item interaction matrix into various subgraphs, and then for each subgraph, we have different adjacency matrix information. The final adjacency matrix for each c subgraph is described as follows:

$$A = \bar{A} + \Delta = \begin{pmatrix} A_{11} & \dots & A_{1c} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ A_{c1} & \dots & A_{cc} \end{pmatrix} \quad (5)$$

4.2 Graph Convolution

Further, to deepen the concept of subgraph structure of the graph with efficient neighborhood as well as users, which are adequate for the diverse recommendation generation, we followed the LightGCN method [5]. The LightGCN method is more effective for recommendation generation, which follows the concept of collaborative filtering. The technique is more effective than the previous neural-based collaborative filtering method and overcomes the feature transformation effect of the non-linear activation method. So we further use the LightGCN methodology to train our subgraphs for the diversified recommendation generation [25]. The overall methodology for applying GCN in the user-item subgraph is executed by first aggregating the intra-layer neighborhood information followed by the inter-layer combination. In inter-layer aggregation, we compute each user in the subgraph; the embedding of the users is updated by the weighted sum of the embedding generated from the neighboring items node embedding. The aggregation is for neighboring users node which is defined as follows:

$$(e_u)^{k+1} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|} \sqrt{|N_i|}} e_i^k \quad (6)$$

Here in Eq. (6), e_u is the embedding of the user u and the e_i item embedding information at the k th layer. $|N_u|$ is total neighbors of node u , and $|N_i|$ is the total neighbors of item node i . Unlike users updated item embedding is computed using a weighted sum of its neighboring users defined as follows:

$$(e_i)^{k+1} = \sum_{i \in N_i} \frac{1}{\sqrt{|N_i|} \sqrt{|N_u|}} e_u^k \quad (7)$$

After, the K th layer operations over all the users and items node for the final layer (K th). In LightGCN method, we aggregated the embeddings in the final layer GCN which computes a weighted sum of the embedding at the different layer. The final layer embedding for users and items node is as follows:

$$e_u = \sum_{k=0}^K \alpha_k \cdot e_u^k \quad (8)$$

$$e_i = \sum_{k=0}^K \alpha_k \cdot e_i^k \quad (9)$$

where $\alpha \geq 0$. The final layer embedding is used for the final rating prediction.

$$\hat{y}_{ui} = e_u^T \cdot e_i \quad (10)$$

The Bayesian loss function defines the objective function for the model's training which encourages observed user-item predictions to have increasingly higher values than unobserved ones, along with L_2 regularization.

$$L_{\text{BPR}} = - \sum_{u=1}^M \sum_{i \in N_u} \sum_{j \notin N_u} \ln \sigma(\hat{y}_{ui} - \hat{y}_{uj}) + \lambda \|E^0\|^2 \quad (11)$$

where E^0 is the 0th layer embedding.

5 Experiments

We first describe the experimental setting for model training and then discuss the comparative results for our proposed methodology.

5.1 Experimental Settings

The experimental evaluation of our proposed approach Div-ClusGCN is done on a dataset MovieLens 1M¹ and ML-100k.² We use various state-of-the-art algorithms for the comparative analysis of our proposed approach Div-ClusGCN. The comparative results for the state-of-the-art algorithms are obtained using the Microsoft Library Recommenders.³ The comparative algorithms are as follows:

- LightGCN—It uses the collaborative filtering concept using the graph convolution operation and it more effective in terms of complexity and efficiency of the recommendation generation.
- Smart Adaptive Recommendation—SAR is an efficient and adaptive algorithm for personalized recommendations based on user transaction history and item description. It produces easily explainable and relevant recommendations.
- Neural Collaborative Filtering—NCF is an efficient algorithm that uses the concept of collaborative filtering and overcomes the drawback of deep neural networks for learning the interaction from the user-item interaction data.

The hyperparameter setting for the proposed Div-ClusGCN model for embedding dimensions is 64 with batch size information 1024 and learning rate = 0.005. We train our model for total 5 epochs, and we used 3 convolution layers for embedding generation. The training and testing ratio will be 75 and 35% for dissimilarity of users we only set threshold to 0.50, and for clustering, we uses k -means algorithm.

5.2 Results and Discussion

We use precision, recall, ndcg, and MAP for our model's performance evaluation. These metrics are used for the accurate prediction of our model's performance. While for diversity metrics evaluation, we choose the intra-list diversity measure discussed in [6]. The detailed comparative results for our proposed model Div-ClusGCN are shown in Table 1. The complete evaluation of our proposed model is completed on two datasets ML-100k and ML-1M. In our proposed model, we have achieved more than 30% increment in the diversity from the state-of-the-art approaches. Due to accuracy-diversity tradeoff, we faces slight decrement in the accuracy of the model but we also get good comparative diversity results.

¹ <https://files.grouplens.org/datasets/movielens/ml-1m.zip>.

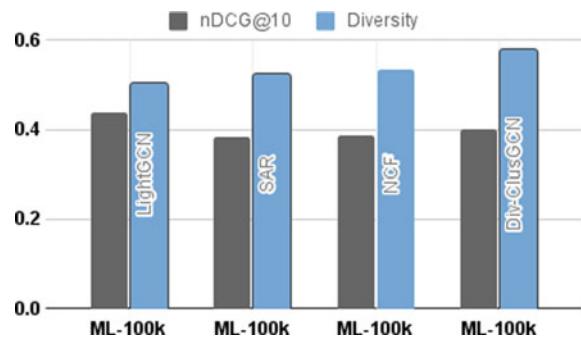
² <https://files.grouplens.org/datasets/movielens/ml-100k.zip>.

³ <https://github.com/microsoft/recommenders>.

Table 1 Comparative results for our proposed model Div-ClusGCN

Dataset	Model	MAP@10	nDCG@10	Precision@10	Recall@10	Diversity
ML-100k	LightGCN	0.12923	0.43629	0.38186	0.20581	0.40214
ML-100k	SAR	0.11059	0.38246	0.33075	0.17638	0.42468
ML-100k	NCF	0.10572	0.38763	0.34219	0.17458	0.43257
ML-100k	Div-ClusGCN	0.09873	0.37887	0.36652	0.15562	0.57814
ML-1M	LightGCN	0.07501	0.37751	0.34567	0.12809	0.49657
ML-1M	SAR	0.06057	0.29924	0.27011	0.10435	0.41243
ML-1M	NCF	0.06282	0.34877	0.32061	0.10812	0.42145
ML-1M	Div-ClusGCN	0.08768	0.32243	0.38876	0.11566	0.52888

The values represented in bold shows the best results

Fig. 1 Accuracy-diversity tradeoff for ML-100k dataset

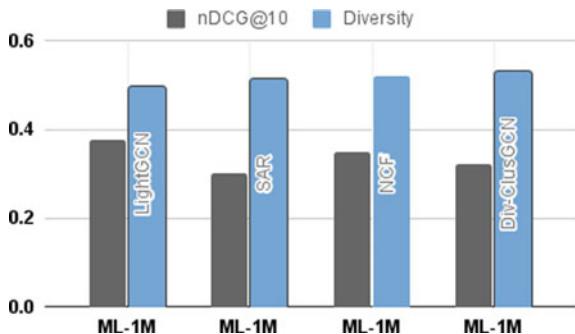
5.3 Accuracy-Diversity Tradeoff

In the recommendation system, if the recommendation list's diversity increases, then the model's accuracy automatically decreases. This is called the accuracy-diversity tradeoff of the model, and it is also an open problem in the recommendation system where increasing the accuracy and diversity at the same time for the recommendation model is challenging. In our proposed model, we also face this challenge, and the result illustration for ML dataset is shown in Figs. 1 and 2.

6 Conclusion

In this paper, we proposed a methodology for diversified recommendation generation by clustering the distinct group of users. Then, we followed the divide their user-item interaction into different subgroups where we made sure that no overlapping of users-

Fig. 2 Accuracy-diversity tradeoff for ML-1M dataset



item interaction would be present. After that, we use the concept of LightGCN for each node's embedding generation. Ultimately, we aggregate this node embedding information for the target user's rating prediction of the items. We significantly improved our model's performance in diversified recommendations while losing a diminutive fall in accuracy. In future, we will try to maintain the accuracy-diversity tradeoff by including some criteria for diverse user selection.

References

1. Nagarnaik P, Thomas A (2015) Survey on recommendation system methods. In: 2015 2nd international conference on electronics and communication systems (ICECS). IEEE, 2015, pp 1603–1608
2. Iwendi C, Ibeke E, Eggoni H, Velagala S, Srivastava G (2022) Pointer-based item-to-item collaborative filtering recommendation system using a machine learning model. *Int J Inf Technol Decis Making* 21(01):463–484
3. Das D, Sahoo L, Datta S (2017) A survey on recommendation system. *Int J Comput Appl* 160(7)
4. Mei D, Huang N, Li X (2021) Light graph convolutional collaborative filtering with multi-aspect information. *IEEE Access* 9:34433–34441
5. He X, Deng K, Wang X, Li Y, Zhang Y, Wang M (2020) LightGCN: simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 639–648
6. Yadav N, Pal S, Singh AK, Singh K (2022) Clus-DR: cluster-based pre-trained model for diverse recommendation generation. *J King Saud Univ Comput Inf Sci*
7. Yadav N, Mundotiya RK, Singh AK, Pal S (2019) Diversity in recommendation system: a cluster based approach. In: International conference on hybrid intelligent systems. Springer, pp 113–122
8. Chiang W-L, Liu X, Si S, Li Y, Bengio S, Hsieh C-J (2019) Cluster-GCN: an efficient algorithm for training deep and large graph convolutional networks. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, pp 257–266
9. Bradley K, Smyth B (2001) Improving recommendation diversity. In: Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science, vol 85, Maynooth, Ireland, pp 141–152
10. Slaney M, White W (2006) Measuring playlist diversity for recommendation systems. In: Proceedings of the 1st ACM workshop on audio and music computing multimedia, pp 77–82

11. Zhang M, Hurley N (2008) Avoiding monotony: improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM conference on recommender systems, pp 123–130
12. Cui L, Ou P, Fu X, Wen Z, Lu N (2017) A novel multi-objective evolutionary algorithm for recommendation systems. *J Parallel Distrib Comput* 103:53–63
13. Hu R, Pu P (2011) Helping users perceive recommendation diversity. In: DiveRS@ RecSys, 2011, pp 43–50
14. Vargas S, Baltrunas L, Karatzoglou A, Castells P (2014) Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: Proceedings of the 8th ACM conference on recommender systems, pp 209–216
15. Hu L, Cao L, Wang S, Xu G, Cao J, Gu Z (2017) Diversifying personalized recommendation with user-session context. *IJCAI* 1858–1864
16. Karakaya MÖ, Aytekin T (2018) Effective methods for increasing aggregate diversity in recommender systems. *Knowl Inf Syst* 56(2):355–372
17. Möller J, Trilling D, Helberger N, van Es B (2018) Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Inf Commun Soc* 21(7):959–977
18. Kotkov D, Veijalainen J, Wang S (2020) How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm. *Computing* 102(2):393–411
19. Garg H, Sharma B, Shekhar S, Agarwal R (2022) Spoofing detection system for e-health digital twin using efficient net convolution neural network. In: Multimedia tools and applications, pp 1–16
20. Matt C, Hess T, Weiß C (2019) A factual and perceptual framework for assessing diversity effects of online recommender systems. In: Internet research
21. Zhang S, Yin H, Chen T, Hung QVN, Huang Z, Cui L (2020) GCN-based user representation learning for unifying robust recommendation and fraudster detection. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 689–698
22. Feng C, Liu Z, Lin S, Quek TQ (2019) Attention-based graph convolutional network for recommendation system. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)
23. Zheng Y, Gao C, Chen L, Jin D, Li Y (2021) DGCN: diversified recommendation with graph convolutional networks. *Proc Web Conf* 2021:401–412
24. Ying R, He R, Chen K, Eksombatchai P, Hamilton WL, Leskovec J (2018) Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, pp 974–983
25. Wang X, He X, Wang M, Feng F, Chua T-S (2019) Neural graph collaborative filtering. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 165–174

Brain Tumor Detection with GLCM Feature Extraction and Hybrid Classification Approach



Shardeep Kaur Sooch and Nitika Kapoor

Abstract Early brain tumor detection has become important to provide timely diagnosis and treatment. Several methodologies are focusing to minimize the manual efforts required for diagnosing by increasing not only the accuracy but also the speed of detection. This proposed methodology includes Otsu's threshold-based segmentation technique after which feature extraction is done by Grey Level Co-Occurrence Matrix (GLCM) to extract 13 intensities-based and textual-based features. The classification is done through the hybrid model of K-nearest neighbor and Random Forest. The final outcome is generated by majority voting which casts its vote to either one of the above hybrid models. The results are compared to existing algorithms on the basis of performance parameters which includes accuracy, recall, specificity, and execution time.

Keywords MRI · Feature extraction · GLCM · Otsu's segmentation · Tumor

1 Introduction

A human brain is the most significant part of the body and consists of 50–100 trillion neurons. The cells in the brain segregate and form new cells. As the old cells age and die, the new cells are responsible for taking the place. When the old cells are not destroyed or the new ones are created when the body doesn't require them, it could lead to extra number of cells in the brain. The lump of tissue caused by the extra cells in brain is called a tumor, and it may affect the performance. As the brain is covered with skull, the complexity increases to analyze and diagnose a disease. The tumor can be benign or malignant. Benign tumors are the non-cancerous tumors, while the malignant are cancerous ones [1]. The brain tumor can also be categorized as primary and metastatic. Primary tumors arise in the brain and affect its performance while the metastatic ones arise somewhere else in the body and gets into the brain through blood stream.

S. K. Sooch (✉) · N. Kapoor
Chandigarh University, Gharuan, Mohali, India

Detection of the tumor can be done using Magnetic Resonance Image (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET). As radiation from the CT scan can be harmful for human brain, MRI scans are widely used from tumor detection. The images in the MRI scans are created using magnetic field and radio waves which provides an accurate visualization of anatomical formation of the brain [2]. Through techniques such as image segmentation, the suspicious regions are extracted from the brain and manually processed by the physicians.

Delay in processing the images or incorrect segmentation can lead to issue in diagnosis. The brightness and contrast of the image can change due to delay in time and the same physician can end up with different results for the image. This increases the need of diagnosis and detection of tumors automatically. As the standard methods are not available to diagnose and detect the tumor, the researchers are working on methodologies to minimize the human efforts and increase the accuracy, precision, and speed of the detection [3].

Any analysis and diagnosis of brain images follow a series of steps for better performance. First the relevant images are collected and processed through several pre-processing techniques. The pre-processing techniques help in de-noising the image and increasing the contrast level through filtering and contrast enhancement. The RGB images are also converted into gray scale, and minor image details are improved. The processed image is segmented using thresholding, region growing, or k-means clustering. Later the important and relevant features are extracted through Discrete Wavelet Transform or Grey Level Co-occurrence Matrix. The final classification can be done using various machine learning algorithms such as support vector machines, decision trees or neural networks [4].

In this paper, a hybrid classification is discussed using K-nearest neighbor and Random Forest. BraTS 2021 brain tumor image dataset is used which consists of 660 cases and 2640 MRI scans. The results are analyzed and compared with different methodologies through various performance parameters.

2 Literature Review

Islam et al. [5] proposed a template-based K-means algorithm where feature extraction is done through superpixels and Principal Component Analysis (PCA) to create an enhanced brain tumor detection technique. Template-based K-means clustering is performed to segment the image and detect the tumor. This proposed method was compared to different other techniques such as thresholding, second order ANN, and region growing. The execution time was decreased to 35–60 s, and 95% accuracy was achieved through it.

Ayadi et al. [6] suggested a model to obtain better quality of Magnetic Resonance Images (MRI) and created distinct feature set using normalization and histogram of gradient technique. Feature extraction was done through Dense Feature points and histogram of gradient. For classification of images, support vector machine was used.

The results were compared with other classifiers such as K-nearest neighbor (KNN) and Random Tree, and overall accuracy of 90.27% was achieved.

Garg et al. [7] used Majority Voting Method to create a hybrid ensemble model which consisted of Random Forest, K-nearest neighbor, and Decision Tree. Otsu's Method was followed for segmenting the image and later the features were extracted through Stationary Wavelet Transform and Principal Component Analysis (PCA). A total 2556 image dataset was used to create a low cost and low computational time methodology giving 97% accuracy.

El Kader et al. [8] proposed a hybrid deep convolution neural network and a deep watershed auto-encoder to train and validate dataset taken from five databases. The pre-processing step included RGB to 1.0 scale conversion and resizing the images to $240 \times 240 \times 3$. The hybrid model was performed in 100 epochs using TensorFlow and Keras framework. The best validation accuracy achieved is 97% which as loss validation of 0.1.

Rinesh et al. [9] suggested a tumor detection model based on performing different operations on hyperspectral images. A combination of k-based algorithm which are K-nearest neighbor and k-means clustering is used, and k value is determined using firefly optimization algorithm. Multilayer feedforward neural network is used to label the parts of brain. This approach resulted in higher peak signal-to-noise ratio and reduction in mean absolute error value. The results were compared with other techniques and overall accuracy of 96.47% was achieved with 96.32% sensitivity and 98.24 specificity.

Dipu et al. [10] used You Only Look Once (YOLO) and FastAI, which is a deep learning library, for tumor detection and classification. A BraTS 2018 dataset containing 1992 MRI images was used. The YOLO model had an accuracy of 85.95%, whereas the FastAI achieved 95.78%. YOLO and FastAI can also be applied for early diagnosis of tumor in real-time tumor detection applications.

Raut et al. [11] suggested a brain tumor model based on Convolutional Neural Network. To have a sufficient set of data, the input Magnetic Resonance Images (MRI) are augmented, and later noise is removed for pre-processing the images. In order to reduce the error and elevate the accuracy, back propagation algorithm is used while training the model. The unnecessary features are removed using autoencoders. K-means algorithm is used to segment the tumor area to achieve a tumor detection accuracy of 97.3%.

Methil et al. [12] proposed a methodology where image pre-processing is done through histogram equalization, and classification is done by Convolutional Neural Network (CNN). A pre-trained learning model, ResNet101v2, is used as a transfer learning over which further training is applied. Several other pre-processing techniques such as Global Thresholding, Adaptive Thresholding, Sobel Filter, High Pass Filter, Median Blur, and Dilation are discussed and tested to choose the required technique. An accuracy of 97.94% was recorded.

Kibriya et al. [13] explored the multiclass classification based on deep learning and machine learning. A total of 15,320 MRI images are used to classify them using Convolutional Neural Network (CNN) models such as RetNet-18 and GoogLeNet.

Support vector machines are used to extract features from the images. This hybrid CNN-SVM based tumor detection technique recorded an accuracy of 98%.

Thachayani et al. [14] proposed detection technique using Convolutional Neural Network (CNN) and sparse stacked autoencoder. A total of 120 MRI images are used, and the methodology is achieved using MATLAB. The combination of Sparse Stacked autoencoder and CNN helped in significantly improve accuracy and the classification effectiveness.

Sadoon et al. [15] classified three types of brain tumor namely pituitary gland, glioma, and meningioma. A comparison was presented between machine learning, deep learning, and Convolutional Neural Network-based models. It also included a pre-processing and post-processing techniques outcomes, and how it affects the performance. The model achieved an overall accuracy of 96.1%. The dataset is divided into 77% for training, 20% for validation, and rest for testing the model. In training, epoch is 100, and initial learning rate is 0.0001 with batch size of 32.

3 Proposed Methodology

The brain tumor detection and classification consist of a series of steps starting from data collection and pre-processing to segmentation and final classification. The methodology used for this paper is shown in Fig. 1.

3.1 *MRI Input*

The MRI dataset of brain tumor images is taken and fed to the model. BraTS 2021 brain tumor image dataset has been used which consists of 660 cases and 2640 MRI scans. Sample MRI image is shown in Fig. 2.

3.2 *Pre-processing*

Pre-processing of the input image is done to de-noise the images using different filtering techniques. It increasing the chance of detection of suspicious region by improving the image details and removing noise. The distortions which are not required are removed, and attributes of the images are improved. The images are resized and converted from RGB to gray scale. By removing the noise, the images become clearer for segmentation, and quality of image is improved. A pre-processed image is shown in Fig. 3.

Fig. 1 Steps for tumor detection

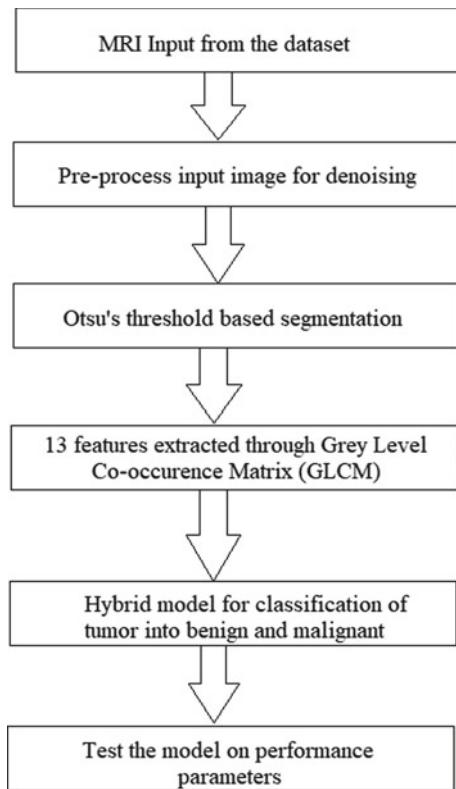


Fig. 2 Input MRI image

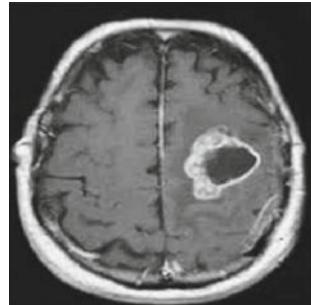


Fig. 3 Pre-processed image

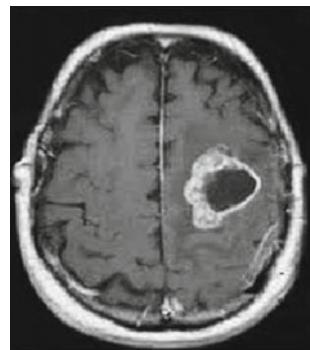


Fig. 4 Image segmentation output



3.3 *Image Segmentation*

Segmentation helps to provide physicians with strong interpretation of the images. Various methods can be used to segment the images such as thresholding, region-based methods, deep learning-based methods, cluster-based methods, and boundary-based methods. Otsu's segmentation is used which performs automatic image thresholding. This process returns a threshold of single intensity which separated the pixels into foreground and background. It achieves the thresholding by maximizing the inter-class variance and minimizing the intra-class intensity variance. A segmented image is shown in Fig. 4.

3.4 *Feature Extraction*

This process extracts the features from the segmented image and verifies if the tumor is low grade or high grade based on shape, size, etc. Grey Level Co-Occurrence Matrix (GLCM) is used to extract 13 intensity and textual-based features from the image. Contrast is a textual-based feature which determines variation in intensity

of threshold and pixel nearest to it, whereas correlation measures the relationship between threshold and the pixel nearest to it. Energy helps to calculate homogeneity and can also be called uniformity. Mean is an intensity-based feature which is an average level of intensity of image, whereas standard deviation is the mean of pixels and their probability density. Entropy is the random variable uncertainty and depends on the probability density. Root Mean Square of RMS is the square root of the mean square. Variance helps to calculate the intensity variation and Kurtosis measures the flatness of histogram. Skewness determines the image symmetry. The local homogeneity is measured through IDM or Inverse Difference Moment.

3.5 Classification

The image after the feature extraction is fed to the classifier to determine whether the image has tumor or not, and if tumor is detected, then is it malignant or benign. A hybrid classifier is created using K-nearest neighbor and Random Forest.

K-nearest neighbor algorithm is a supervised learning algorithm which makes the classifications based on how close the data point is to the group. It works on the assumption that points which are similar to one another can be found near each other. Hence, it is more used for classification solutions rather than regression problems. The class label is assigned to a point by identifying the neighbor point which is nearest.

The Random Forest is generated through integration of multiple decision trees, and each tree contains a set of rules. It is an ensemble model which is used for classification and regression. In order to solve the classification issues, the class which is selected by the majority of the trees is considered the algorithm output. For regression solutions, the output is calculated by the mean of the prediction result by each tree.

The outputs from the K-nearest neighbor and Random Forest are fed to voting algorithm which casts its vote of one of the algorithms in order to achieve final required outcome.

4 Result

The outcome of the feature extraction and classification is compared with existing methodology. The results of the 13 feature extraction parameters are given in Table 1.

The performance parameters include accuracy, precision, recall, specificity, and execution time. The values are calculated from confusion matrix which consists of True Positive, True Negative, False Positive, and False Negative.

True Positive (TP): The model detects tumor and the actual observation indicates the same.

Table 1 Feature extraction parameters

Mean	74.2037
Standard deviation	71.5716
Entropy	6.38931
RMS	94.4357
Variance	3799.8
Smoothness	1
Kurtosis	3.13073
Skewness	0.842085
IDM	255
Contrast	0.36283
Correlation	0.956487
Energy	0.169999
Homogeneity	0.899183

True Negative (TN): The actual observation shows the tumor but model detects it as non-tumor.

False Positive (FP): The actual observation the model both detects the image as non-tumor.

False Negative (FN): The actual observation shows that there is no tumor but the model detects the tumor.

Specificity (β) is determined by the correct classification of the images that has no tumor.

$$\beta = \frac{\text{TN}}{\text{TN} + \text{TP}} \times 100$$

Recall indicates the proportion of actual positives that are identified correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

Accuracy (η) defines the ability to test the classification done correctly.

$$\eta = \frac{\text{Number of correctly classified records}}{\text{Total Record in the test set}} \times 100$$

$$\eta = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$$

The comparison of different performance parameters is given in Table 2.

Table 2 Performance parameter comparison

Algorithm	Recall	Specificity	Accuracy	Execution time
Thresholding	76.9	64.25	72.5	3 min
Region growing	92.8	83.3	90.0	10 min
Second order + ANN	95.5	88.8	92.5	7–15 min
PCA + K means	97.36	100	95	35–60 s
Proposed methodology	97.5	98.6	95.7	2–3 s

5 Conclusion and Future Scope

The proposed methodology is compared with existing ones and an accuracy of 96% is achieved with 2–3 s of execution time. The performance is also compared by recall and specificity parameters. A combination of K-nearest neighbor and Random Forest proved in providing good accuracy with very less execution time. Otsu's segmentation and GLCM feature extraction are able to improve the image quality for better classification.

In the future, work can be done to enhance the accuracy of the detection. Other algorithms can be included in the hybrid model to verify the performance, and model can be made more compatible with deep learning systems. The methodology can be trained and tested on vast MRI dataset which includes real-time clinical data.

References

1. Deepa, Singh A (2016) Review of brain tumor detection from MRI images. In: 2016 3rd International conference on computing for sustainable global development (INDIACoM), pp 3997–4000
2. Sooch SK, Anand D (2021) Emotion classification and facial key point detection using AI. In: 2021 2nd international conference on advances in computing, communication, embedded and secure systems (ACCESS), pp 1–5. <https://doi.org/10.1109/ACCESS51619.2021.9563289>
3. Sooch SK, Anand D (2021) Smart health monitoring during pandemic using internet of things. In: 2021 10th IEEE international conference on communication systems and network technologies (CSNT), pp 489–493. <https://doi.org/10.1109/CSNT51715.2021.9509591>
4. El-Feshawy, Somaya, Shokair M, Saad W, Dessouky MI (2021) Brain tumour classification based on deep convolutional neural networks
5. Islam MK, Ali MS, Miah MS, Rahman MM, Alam MS, Hossain MA (2021) Brain tumor detection in MR image using superpixels, principal component analysis and template-based K-means clustering algorithm. *Mach Learn Appl* 5:100044. ISSN 2666-8270, <https://doi.org/10.1016/j.mlwa.2021.100044>
6. Ayadi W, Charfi I, Elhamzi W, Atri M (2022) Brain Tumor classification based on hybrid approach. *Vis Comput* 38(1):107–117. <https://doi.org/10.1007/s00371-020-02005-1>
7. Garg G, Garg R (2021) Brain tumor detection and classification based on hybrid ensemble classifier. *arXiv preprint arXiv:2101.00216*
8. El Kader IA, Xu G, Shuai Z, Saminu S (2021) Brain tumor detection and classification by hybrid CNN-DWA model Using MR images. *Curr Med Imaging* 17(10):1248–1255. <https://doi.org/10.2174/1573405617666210224113315>. PMID: 33655844

9. Rinesh S, Maheswari K, Arthi B, Sherubha P, Vijay A, Sridhar S, Rajendran T, Waji YA (2022) Investigations on brain tumor classification using hybrid machine learning algorithms. *J Healthcare Eng* 2022;9. Article ID 2761847. <https://doi.org/10.1155/2022/2761847>
10. Dipu NM, Shohan SA, Salam KMA (2021) Deep learning based brain tumor detection and classification. In: 2021 International conference on intelligent technologies (CONIT), pp 1–6. <https://doi.org/10.1109/CONIT51480.2021.9498384>
11. Raut G, Raut A, Bhagade J, Bhagade J, Gayhane S (2020) Deep learning approach for brain tumor detection and segmentation. In: 2020 International conference on convergence to digital World—Quo Vadis (ICCDW), pp 1–5. <https://doi.org/10.1109/ICCDW45521.2020.9318681>
12. Methil AS (2021) Brain tumor detection using deep learning and image processing. In: 2021 International conference on artificial intelligence and smart systems (ICAIS), pp 100–108. <https://doi.org/10.1109/ICAIS50930.2021.9395823>
13. Kibriya H, Masood M, Nawaz M, Rafique R, Rehman S (2021) Multiclass brain tumor classification using convolutional neural network and support vector machine. In: 2021 Mohammad Ali Jinnah University international conference on computing (MAJICC), pp 1–4. <https://doi.org/10.1109/MAJICC53071.2021.9526262>
14. Thachayani M, Kurian S (2021) AI based classification framework for cancer detection using brain MRI images. In: 2021 International conference on system, computation, automation and networking (ICSCAN), pp 1–4. <https://doi.org/10.1109/ICSCAN53069.2021.9526456>
15. Sadoon A, Toqa, Al-Hayani, Mohammed (2021) Deep learning model for glioma, meningioma and pituitary classification. *Int J Adv Appl Sci* 10:88. <https://doi.org/10.11591/ijaas.v10.i1.pp88-98>

Optimization of an Inventory Model with Selling Price and Stock Sensitive Demand Along with Trade Credit Policy



Mamta Kumari , Pankaj Narang , and Pijus Kanti De 

Abstract The demand for a product is influenced by a number of factors, including selling price and displayed stock, among others. Considering this, a novel mathematical model is proposed that takes into account the aforementioned situation where both the selling price and the amount of inventory on hand influence consumer's demand. Besides that, the supplier grants a full trade credit period to the retailer. This policy is very advantageous for both the counterpart—the seller and the buyer. By offering a delay time, the supplier can entice additional clients, while the retailer has the advantage of receiving items without immediate payment. The suggested inventory model aims at determining the optimal selling price and optimal replenishment cycle length so as to maximize the total profit of the retailer per unit time. The suggested inventory model is also demonstrated numerically, as well as an extensive sensitivity analysis is executed to emphasize the outcomes and offer insightful managerial information. Sensitivity analysis can be helpful in figuring out how different cost factors will affect the overall profit earned.

Keywords Inventory · Selling price · Stock-dependent · Trade credit

1 Introduction

Majority of the inventory models assume demand rate to remain the same throughout the inventory cycle. Demand is seen to be influenced by a variety of factors in real life, including selling price, displayed stock level, among others. According to Levin et al. [17], customers exhibit a significant desire to purchase more products when supermarkets display them in the showroom in huge numbers; this frequently raises product demand. In the past few years, different scholars have formulated various

M. Kumari  · P. Narang · P. K. De

Department of Mathematics, National Institute of Technology Silchar, Silchar Assam-788010, India

e-mail: nk656769@gmail.com

P. K. De

e-mail: pkde@math.nits.ac.in

inventory models assuming demand to be dependent on the level of inventory in hand. Datta and Pal [6] assumed demand to be dependent on inventory level. Wu along with his co-researchers [23] evaluated an inventory model with partial backlogging for non-instantaneous deteriorating products where demand was observed to be dependent on stock level.

Pricing is another key aspect that influences the demand for a product. A vital question that arises in the inventory model is what should be an item selling price so that the seller gains maximum benefits while satisfying customer needs. It is also evident that the higher the price, the lesser the demand for the item. In this paper, demand is observed to be exponentially dependent on price following Robinson and Lakhani [21], Teng and Chang [22], and Chang et al. [3]. Chang with his collaborators [4] portrayed an integrated model of inventory with the policy of trade credit tied with order size and demand to be price-dependent. One another common assumption prevalent in the inventory systems is that the buyer should pay for the goods upon receipt of merchandise. There may arise a situation where the buyer may not have enough money to pay instantly for the ordered goods. There arises the need for a trade credit policy which can be rephrased in short as “buy now and pay later”. It allows the buyer to buy goods without the requirement of instant payment and is granted a credit period to settle accounts. The trade credit policy is often utilized to stimulate demand. Haley and Higgins [12] were the first to frame an economic order quantity (EOQ) inventory model considering an allowable delay in payments. Following that, Goyal [10] investigated an EOQ model of inventory with trade credit where the buyer is exempted from clearing the payment and earns interest throughout the credit period. Various trade credit policies have been developed as a result of diversification of trade and changes in the business environment. In the United States, trade credit financed about 60% of small enterprises, and 20% of all investments outside the United States are financed externally through trade credit according to the report of Cuñat and García-Appendini [5]. After that, Khanra and his co-researchers [15] constructed an inventory model allowing delay in payments for a single product with consumer demand to be quadratically dependent on time considering shortages. Cárdenas-Barrón and his collaborators [2] developed an inventory situation with demand to be nonlinearly dependent on stock, allowing partial backlogging and nonlinear holding cost. The retailer was allowed a full trade credit period to clear the debt by the supplier. Ghosh along with his co-researchers [9] formulated an inventory model with multiple advanced and delayed payment policies along with complete back ordering for perishable items. Similarly, other authors have made valuable contributions to the existing literature [1, 8, 11, 14, 16, 18–20].

Some researchers have formulated different inventory models where demand is influenced by both selling price and displayed stock level. By taking into account the exhibited stock level and selling price-dependent demand, Hsieh & Dye [13] constructed a model of inventory for deteriorating items mathematically. To find out the optimal solution, particle swarm optimization is applied. Feng [7] constructed an inventory model with non-zero ending level of inventory for perishable items assuming customer demand to be influenced by selling price, product freshness, and inventory level.

The present paper systematically builds up an inventory situation considering demand to be a function of selling price and displayed stock level. Majority of the existing mathematically formulated supply chain models are predicated on the unrealistic premise that demand is always consistent. The retailer is offered a full trade credit policy by the supplier. This study aims to find out the optimal selling price at which goods are to be sold, optimal replenishment cycle length, optimal order quantity so as to maximize the total generated profit of the retailer. A numerical example is also demonstrated as well as an extensive sensitivity analysis is executed to demonstrate the key outcomes and provide important managerial insights.

The paper is further systematically arranged as follows: The assumptions and notations required to establish the inventory model mathematically are well described in Sect. 2. Section 3 constructs the mathematical model considering trade credit policy and nonlinear demand. Section 4 highlights the outcomes of the presented mathematical model using a numerical example. In Sect. 5, sensitivity analysis is done by changing one parameter at a time and keeping rest of the parameters constant. Finally, Sect. 6 concludes the findings as well as suggests some important future research directions.

2 Assumptions and Notations

2.1 Assumptions

- The inventory system's planning horizon is infinite.
- Rate of replenishment is instantaneous with negligible lead time.
- Demand is assumed to be a function of selling price and stock level given by

$$D(t) = \alpha e^{-\lambda p} [q(t)]^\beta \quad \text{where } \lambda > 0.$$

- Retailer is granted a full trade credit period by the supplier.

2.2 Notations

See Table 1.

3 Mathematical Formulation

In this section, an inventory model is constructed with nonlinear demand and trade credit. The inventory situation is shown in Fig. 1. Demand is a function of selling

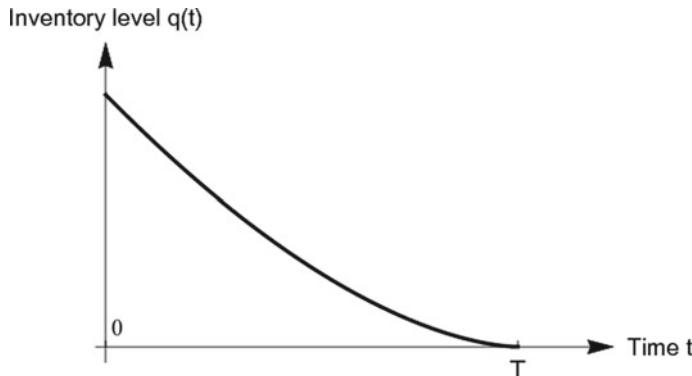
Table 1 Notations used to establish the inventory model

Parameter	Unit	Description	Parameter	Unit	Description
c	\$/unit	Purchasing price per unit	I_p	%/unit time	Interest paid by the retailer
o	\$/order	Ordering cost per order	α		Demand rate scale parameter
h	\$/unit/ unit time	Holding cost per unit item per unit time	Decision variables		
β		Demand elasticity rate; $0 \leq \beta < 1$	p		Selling price per unit item
I_e	%/unit time	Interest earned by the retailer	T	unit time	Replenishment cycle length
M	unit time	trade credit period offered by the supplier to the retailer	Dependent variables:		
$TP(p, T)$	\$/unit time	total profit per unit time	Q	units	Order quantity per cycle

price and displayed stock level. Q units of product initially exist in the inventory. The level of inventory declines owing to the demand patterns throughout the time period $[0, T]$. At time $t = T$, it finally reaches to zero. Then, a replenishment order of Q units is placed which marks the commencement of the subsequent cycle.

The supplier grants a full trade credit period M to the retailer. The situation of the inventory is best governed by the following differential equation:

$$\frac{dq}{dt} = -\alpha e^{-\lambda p} [q(t)]^\beta \quad 0 \leq t \leq T \quad (1)$$

**Fig. 1** Inventory situation at any given time t

The level of inventory at any time t is obtained by solving the differential Eq. (1) along with the boundary conditions $q(T) = 0$ as follows:

$$q(t) = [\alpha e^{-\lambda p}(1 - \beta)(T - t)]^{\frac{1}{1-\beta}} \quad 0 \leq t \leq T \quad (2)$$

Employing the boundary condition $q(0) = Q$ in Eq. (2), the order quantity is computed as follows:

$$Q = [\alpha e^{-\lambda p}(1 - \beta)T]^{\frac{1}{1-\beta}} \quad (3)$$

The following list of expenses is related to the suggested inventory model:

1.

$$\text{Ordering cost} = o \quad (4)$$

2.

$$\text{Purchasing cost(PC)} = c[\alpha e^{-\lambda p}(1 - \beta)T]^{\frac{1}{1-\beta}} \quad (5)$$

3.

$$\text{Sales revenue collected during the time period(SR)} = p[\alpha e^{-\lambda p}(1 - \beta)T]^{\frac{1}{1-\beta}} \quad (6)$$

4.

$$\text{Holding cost(HC)} = \frac{h}{\alpha e^{-\lambda p}(2 - \beta)} [\alpha e^{-\lambda p}(1 - \beta)T]^{\frac{2-\beta}{1-\beta}} \quad (7)$$

Since the retailer is granted a trade credit policy by the supplier, the following two cases arise:

Case 1: $M \leq T$.

Case 2: $M \geq T$.

Case 1: Trade credit period is less than or equal to the cycle length ($M \leq T$).

The trade credit period offered to the retailer by the supplier is less than or equal to the cycle length in this scenario. After the end of credit period, the retailer has to bear the interest charges and needs to pay interest during the interval $[M, T]$. Consequently, the amount of interest paid is computed as follows:

$$\text{IP} = \frac{c I_p \left[[\alpha e^{-\lambda p}(1 - \beta)(T - M)]^{\frac{2-\beta}{1-\beta}} \right]}{\alpha e^{-\lambda p}(2 - \beta)} \quad (8)$$

During the credit period, until time $t = M$, the retailer earns interest.

$$\begin{aligned} \text{IE} = pI_e & \left[M \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} \right. \\ & + \frac{1}{\alpha e^{-\lambda p} (2 - \beta)} \left\{ \left\{ \alpha e^{-\lambda p} (1 - \beta) (T - M) \right\}^{\frac{2-\beta}{1-\beta}} \right. \\ & \left. \left. - \left\{ \alpha e^{-\lambda p} (1 - \beta) T \right\}^{\frac{2-\beta}{1-\beta}} \right\} \right] \end{aligned} \quad (9)$$

The overall profit per unit time is computed as shown below:

$$\text{TP}_1(p, T) = \frac{\text{SR} + \text{IE} - o - \text{PC} - \text{HC} - \text{IP}}{T}$$

Therefore,

$$\begin{aligned} \text{TP}_1(p, T) = & \left[\frac{1}{T} \right] \left[p \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} + pI_e \right. \\ & \left[M \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} + \frac{1}{\alpha e^{-\lambda p} (2 - \beta)} \left\{ \left\{ \alpha e^{-\lambda p} (1 - \beta) (T - M) \right\}^{\frac{2-\beta}{1-\beta}} \right. \right. \\ & \left. \left. - \left\{ \alpha e^{-\lambda p} (1 - \beta) T \right\}^{\frac{2-\beta}{1-\beta}} \right\} \right] - o - c \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} \\ & - \frac{h}{\alpha e^{-\lambda p} (2 - \beta)} \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{2-\beta}{1-\beta}} \\ & \left. - \frac{cI_p \left[\left[\alpha e^{-\lambda p} (1 - \beta) (T - M) \right]^{\frac{2-\beta}{1-\beta}} \right]}{\alpha e^{-\lambda p} (2 - \beta)} \right] \end{aligned} \quad (10)$$

Problem 1

Maximize $\text{TP}_1(p, T) = \frac{S_1}{T}$.

Where $S_1 = \text{SR} + \text{IE} - o - \text{PC} - \text{HC} - \text{IP}$.

Subject to $M \leq T$.

Case 2: Trade credit period is greater than or equal to the cycle length ($M \geq T$).

In this case, the trade credit period granted to the retailer by the supplier is greater than or equal to the cycle length. The retailer is exempted from paying interest in this scenario. Therefore,

$$\text{IP} = 0 \quad (11)$$

The retailer earns interest as long as the credit period is active, up until time $t = M$.

$$IE = pI_e \left[M \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} - \frac{\left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{2-\beta}{1-\beta}}}{\alpha e^{-\lambda p} (2 - \beta)} \right] \quad (12)$$

The overall profit per unit time is computed as shown below:

$$TP_2(p, T) = \frac{SR + IE - o - PC - HC - IP}{T}$$

Therefore,

$$\begin{aligned} TP_2(p, T) = & \left[\frac{1}{T} \right] \left[p \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} + pI_e \left[M \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} \right. \right. \\ & \left. \left. - \frac{\left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{2-\beta}{1-\beta}}}{\alpha e^{-\lambda p} (2 - \beta)} \right] - o - c \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{1}{1-\beta}} \right. \\ & \left. - \frac{h}{\alpha e^{-\lambda p} (2 - \beta)} \left[\alpha e^{-\lambda p} (1 - \beta) T \right]^{\frac{2-\beta}{1-\beta}} \right] \end{aligned} \quad (13)$$

Problem 2

Maximize $TP_2(p, T) = \frac{S_2}{T}$.

Where $S_2 = SR + IE - o - PC - HC - IP$.

Subject to $M \geq T$.

4 Numerical Example

The aforementioned described inventory model highlights the outcomes of the presented mathematical model using a numerical example. The aim is to find out the optimal selling price (p^*), optimal replenishment length of the cycle (T^*), order quantity (Q^*) so as to maximize the overall profit earned by the retailer per unit time. It is solved by using the graphical method in MATHEMATICA software. The profit function plot is shown in Fig. 2. The various input parameters are listed below:

$\lambda = 0.01$; $I_p = 15\% /year$; $I_e = 10\% /year$; $M = 60/365 year$; $\alpha = 50$; $o = \$200/order$; $\beta = 0.4$; $c = \$60/unit$; $h = \$15/unit/year$.

Hence, the optimal solution is:

$p^* = 159.643$; $T^* = 4.55528$; $Q^* = 253.431$; $TP^*(p, T) = 3303.02$.

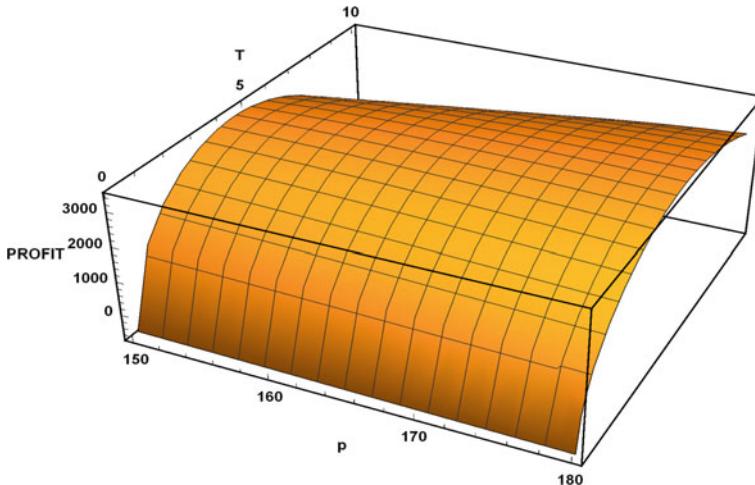


Fig. 2 Change in profit function in response to the decision variables p and T

5 Sensitivity Analysis

The impact of over or underestimation of input parameters on the optimal values of selling price (p^*), replenishment cycle length (T^*), order quantity (Q^*), and the total profit per unit time ($TP^*(p, T)$) of the inventory system is studied using the above numerical example. The extensive sensitivity analysis is executed by adjusting the values of input parameters from -20% to 20%. It is accomplished by changing the input parameters one at a time and keeping rest of the parameters constant. Table 2 presents the outcomes listed below:

1. It can be seen that with the increase in purchasing price c , the optimal selling price (p^*) increases. The replenishment cycle length (T^*), order quantity (Q^*), and the overall profit per unit time decrease. This is obvious.
2. It is observed that as λ increases, the optimal selling price (p^*), replenishment cycle length (T^*), order quantity (Q^*), and the overall profit per unit time decreases. As β increases, total profit per unit time increases.
3. With the increase in the value of h , optimal selling price (p^*) increases. The replenishment cycle length (T^*), order quantity (Q^*), and the total profit per unit time decreases. This is an apparent result since the retailer will try to minimize its cost by keeping fewer items. As α increases, order quantity (Q^*) and the total profit per unit time increases. With the increase in ordering cost o , a decline in overall profit per unit time is noticed.
4. As I_p increases, overall profit per unit time declines. It is visible that with the increment in the value of I_e , total profit per unit time increases. As M increases,

Table 2 Sensitivity analysis in response to the changing input parameters

Parameter	% change in parameter	Change in optimal values			
		p^*	T^*	TP^*	Q^*
c	-20	150	5.08219	4241.32	357.177
	-10	153.571	4.73092	3742.81	298.672
	10	165	4.37965	2916.71	217.082
	20	170.893	4.20401	2576.85	183.799
λ	-20	184.821	5.65949	5865.08	442.868
	-10	170.893	5.08219	4358.3	335.237
	10	150	4.06091	2541.28	191.395
	20	130	3.32583	1932.63	159.418
β	-20	161.107	3.6025	2011.34	109.289
	-10	159.821	4.06091	2527.35	165.291
	10	158.929	5.20965	4528.53	427.847
	20	158.929	5.89457	6592.65	750.577
h	-20	158.571	5.08219	3619.69	309.631
	-10	159.6	4.90656	3452.46	287.039
	10	159.821	4.20401	3167	221.047
	20	160.179	4.02838	3044.66	204.647
α	-20	159.107	4.55528	2263.48	176.288
	-10	159.643	4.55528	2764	212.616
	10	159.107	4.55528	3879.15	299.728
	20	159.107	4.55528	4491.39	346.504
o	-20	159.643	4.55528	3311.8	253.431
	-10	159.643	4.55528	3307.41	253.431
	10	159.643	4.55528	3298.63	253.431
	20	159.643	4.55528	3294.23	253.431
I_p	-20	159.643	4.90656	3467.92	286.833
	-10	159.643	4.73092	3382.77	269.925
	10	159.286	4.37965	3228.13	238.772
	20	158.929	4.20401	3157.56	224.358
I_e	-20	159.643	4.55528	3302.14	253.431
	-10	159.643	4.55528	3302.58	253.431
	10	159.643	4.55528	3303.45	253.431
	20	159.643	4.55528	3303.89	253.431
M	-20	160.179	4.55528	3285.82	251.177
	-10	159.643	4.55528	3294.43	253.431
	10	159.107	4.55528	3311.65	255.705
	20	159.107	4.55528	3320.39	255.705

the retailer has the chance to sell more goods and collect sales revenue. The retailer has to pay interest charges for a lesser number of goods; hence, the overall profit per unit time increases.

6 Conclusion

The primary goal of this investigation is to analyze an inventory situation with trade credit agreement between the merchant and the buyer. In the course of daily living, demand for a product is influenced by different factors including selling price, displayed stock, among others. In order to compete in the business era, it is very crucial to decide the selling price of the item since it directly impacts customer choice. The retailer must appropriately decide the selling price in order to generate profit rather than suffer loss. Another aspect is to decide efficiently the replenishment cycle length so as to avoid shortages and run the business smoothly. In the present paper, an effort is made to build an inventory model where the demand for an item is influenced by selling price as well as displayed stock. The supplier grants a full trade credit period to the retailer which implies that the latter has the opportunity to buy goods instantly without making any immediate payment. The present study aims to determine the optimal selling price, order quantity, as well as replenishment cycle length so as to maximize the total profit generated by the retailer. The extension of this research work is manifold such as analyzing partial trade credit policy, inflation, shortages, order size-dependent trade credit, as well as fuzzy-valued inventory costs, among others.

References

1. Alfares HK, Ghaithan AM (2016) Inventory and pricing model with price-dependent demand, time-varying holding cost, and quantity discounts. *Comput Ind Eng* 94:170–177
2. Cárdenas-Barrón LE, Shaikh AA, Tiwari S, Treviño-Garza G (2020) An EOQ inventory model with nonlinear stock dependent holding cost, nonlinear stock dependent demand and trade credit. *Comput Ind Eng* 139:105557
3. Chang CT, Goyal SK, Teng JT (2006) On “An EOQ model for perishable items under stock-dependent selling rate and time-dependent partial backlogging” by Dye and Ouyang. *Eur J Oper Res* 174(2):923–929
4. Chang H-C, Ho C-H, Ouyang L-Y, Su C-H (2009) The optimal pricing and ordering policy for an integrated inventory model linked to order quantity. *Appl Math Model* 33:2978–2991
5. Cuñat V, García-Appendini E (2012) Trade credit and its role in entrepreneurial finance. Oxford handbook of entrepreneurial finance, 526–557
6. Datta TK, Pal AK (1990) A note on an inventory model with inventory-level dependent demand rate. *J Oper Res Soc* 41(10):971–975
7. Feng L, Chan YL, Cárdenas-Barrón LE (2017) Pricing and lot-sizing policies for perishable goods when the demand depends on selling price, displayed stocks, and expiration date. *Int J Prod Econ* 185:11–20

8. Garg P, Chauhan Gonder SS, Singh D (2022) Hybrid crossover operator in genetic algorithm for solving N-queens problem. In: *Soft computing: theories and applications*, 91–99, Springer, Singapore
9. Ghosh PK, Manna AK, Dey JK, Kar S (2021) An EOQ model with backordering for perishable items under multiple advanced and delayed payments policies. *J Manage Anal* <https://doi.org/10.1080/23270012.2021.1882348>
10. Goyal SK (1985) Economic order quantity under conditions of permissible delay in payments. *J Oper Res Soc* 36(4):335–338
11. Gupta H, Kumar S, Yadav D, Verma OP, Sharma TK, Ahn CW, Lee JH (2021) Data analytics and mathematical modeling for simulating the dynamics of COVID-19 epidemic—a case study of India. *Electronics* 10(2):127
12. Haley CW, Higgins HC (1973) Inventory policy and trade credit financing. *Manage Sci* 20(4):464–471
13. Hsieh TP, Dye CY (2017) Optimal dynamic pricing for deteriorating items with reference price effects when inventories stimulate demand. *Eur J Oper Res* 262(1):136–150
14. Khanna A, Kishore A, Sarkar B, Jaggi CK (2020) Inventory and pricing decisions for imperfect quality items with inspection errors, sales returns, and partial backorders under inflation. *RAIRO-Oper Res* 54(1):287–306
15. Khanra S, Mandal B, Sarkar B (2013) An inventory model with time dependent demand and shortages under trade credit policy. *Econ Model* 35:349–355
16. Kumar A, Sharma TK, Verma OP, Poonia AS, Bisht M (2022) COVID-19 Cases in India: prediction and analysis using machine learning. In: *Soft computing: theories and applications*. Springer, Singapore, 551–563
17. Levin RI, McLaughlin CP, Lamone RP, Kattas JF (1972) *Productions/operations management: contemporary policy for managing operating systems*. McGraw Hill, New York
18. Mishra U, Cárdenas-Barrón LE, Tiwari S, Shaikh AA, Treviño-Garza G (2017) An inventory model under price and stock dependent demand for controllable deterioration rate with shortages and preservation technology investment. *Ann Oper Res* 254(1–2):165–190
19. Musa A, Sani B (2012) Inventory ordering policies of delayed deteriorating items under permissible delay in payments. *Int J Prod Econ* 136:75–83
20. Ouyang L-Y, Ho C-H, Su C-H (2009) An optimization approach for joint pricing and ordering problem in an integrated inventory system with order-size dependent trade credit. *Comput Ind Eng* 57:920–930
21. Robinson B, Lakhani C (1975) Dynamic price models for new-product planning. *Manage Sci* 21(6):1113–1122
22. Teng JT, Chang CT (2005) Economic production quantity models for deteriorating items with price- and stock-dependent demand. *Comput Oper Res* 32(2):297–308
23. Wu KS, Ouyang LY, Yang CT (2006) An optimal replenishment policy for non-instantaneous deteriorating items with stock-dependent demand and partial backlogging. *Int J Prod Econ* 101(2):369–384

A New Family of Generalized Euler-Genocchi Polynomials Associated with Hermite Polynomials



Azhar Iqbal and Waseem A. Khan

Abstract In this paper, we introduce a new class of generalized Hermite-based Euler-Genocchi polynomials and present some properties and identities of these polynomials. We derive some explicit and implicit summation formulas of these polynomials. Also, we derive some symmetry identities of these polynomials by applying generating functions.

Keywords Hermite polynomials · Generalized Euler-Genocchi polynomials · Summation formulae · Symmetric identities

Mathematics Subject Classification 05A10 · 05A19 · 11B68 · 33C45

1 Introduction

The Hermite polynomials play an essential position within extension of the classical, unique features. Starting from the Hermite polynomials, it has already been possible to attain a few extensions of a few classically unique sets of capabilities, which include Bessel capabilities, Dickson polynomials, Laguerre polynomials, Chebyshev polynomials (see [1–3, 8, 14, 15]).

The Hermite (or 2 variable Kampé de Fériet) polynomials is given by (see [1, 2])

$$\mathbb{H}_v(\xi, \eta) = v! \sum_{r=0}^{\lfloor \frac{v}{2} \rfloor} \frac{\eta^r \xi^{j-2r}}{r!(j-2r)!}. \quad (1)$$

A. Iqbal (✉) · W. A. Khan

Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, P.O. Box 1664, Al Khobar 31952, Saudi Arabia
e-mail: aiqbal@pmu.edu.sa

The generating function of the Hermite polynomials is defined by

$$e^{\xi\omega+\eta\omega^2} = \sum_{v=0}^{\infty} \mathbb{H}_v(\xi, \eta) \frac{\omega^v}{v!}, \quad (2)$$

letting $\eta = -1$ and $\xi \rightarrow 2\xi$, we get

$$e^{2\xi\omega-\omega^2} = \sum_{v=0}^{\infty} \mathbb{H}_v(\xi) \frac{\omega^v}{v!}.$$

The falling factorial sequence is defined by

$$(\xi)_0 = 1, \quad (\xi)_v = \xi(\xi - 1) \cdots (\xi - v + 1), \quad (v \geq 1). \quad (3)$$

The first kind of Stirling numbers are defined by

$$(\xi)_v = \sum_{\eta=0}^v S_1(v, \eta) \xi^{\eta} \quad (v \geq 0) \quad (\text{see [1-10]}). \quad (4)$$

The Stirling numbers of the second kind are given by (see [4-11])

$$\xi^v = \sum_{\eta=0}^v S_2(v, \eta) (\xi)_{\eta}. \quad (5)$$

By Eqs. (1.4) and (1.5), the first and second kind of Stirling numbers are defined by generating function, respectively (see [11-17])

$$\frac{1}{\eta!} (\log(1 + \omega))^{\eta} = \sum_{v=\eta}^{\infty} S_1(v, \eta) \frac{\omega^v}{v!}, \quad (6)$$

and

$$\frac{1}{\eta!} (e^{\omega} - 1)^{\eta} = \sum_{v=\eta}^{\infty} S_2(v, \eta) \frac{\omega^v}{v!} \quad (\eta \geq 0). \quad (7)$$

Let r and m be integers with $r \geq 0$ and $m \geq 0$, Frontczak et al. [5] introduced and studied a new class of generalized Euler-Genocchi polynomials as

$$\sum_{v=0}^{\infty} A_v^{(r,m)}(\xi) \frac{\omega^v}{v!} = \left(\frac{2\omega^r}{e^{\omega} + 1} \right)^m e^{\xi\omega}, \quad |\omega| < \pi, \quad (8)$$

with $A_v^{(r,0)}(\xi) = \xi^v$ for all $r \geq 0$ and $A_v^{(r,m)}(\xi) = 0$ for all $v < rm$. We call the numbers $A_v^{(r,m)}(0) = A_v^{(r,m)}$ the generalized Euler-Genocchi numbers of order m .

It is clear that $A_v^{(r,1)}(\xi) = A_v^{(r)}(\xi)$ and ${}_H A_0^{(r,m)}(\xi, \eta) = 0$ for $r, m \geq 1$. Furthermore, we note that $A_v^{(0,m)}(\xi) = E_v^{(m)}(\xi)$ and $A_v^{(1,m)}(\xi) = G_v^{(m)}(\xi)$ are the generalized Euler and Genocchi polynomials as follows (see [1, 7, 8]):

$$\sum_{v=0}^{\infty} E_v^{(m)}(\xi) \frac{\omega^v}{v!} = \left(\frac{2}{e^{\omega} + 1} \right)^m e^{\xi\omega} \quad | \omega | < \pi, \quad (9)$$

and

$$\sum_{v=0}^{\infty} G_v^{(m)}(\xi) \frac{\omega^v}{v!} = \left(\frac{2\omega}{e^{\omega} + 1} \right)^m e^{\xi\omega} \quad | \omega | < \pi. \quad (10)$$

For each $\rho \geq 0$, $S_\rho(v)$ [17] defined by

$$S_\rho(v) = \sum_{\eta=0}^v \eta^\rho, \quad (11)$$

$$\sum_{\rho=0}^{\infty} S_\rho(v) \frac{\omega^\rho}{\rho!} = 1 + e^\omega + e^{2\omega} + \cdots + e^{v\omega} = \frac{e^{(v+1)\omega} - 1}{e^\omega - 1}. \quad (12)$$

In this paper, we have presented the Hermite-based generalized Euler-Genocchi polynomials and discussed, in particular, some interesting series representations. We have deduced some relevant properties by using the structure and the relations satisfied by the recently generalized Hermite polynomials. Section 2 incorporates the definition of Hermite-based generalized Euler-Genocchi polynomials and a preliminary study of these polynomials. Some theorems on implicit summation formulae for Hermite-based generalized Euler-Genocchi polynomials ${}_H A_v^{(r,m)}(\xi, \eta)$ and their special cases are given in Sect. 3. Finally, symmetry identities for Hermite-based generalized Euler-Genocchi polynomials are given in Sect. 4.

2 Hermite-Based Generalized Euler-Genocchi Polynomials

This section incorporates the definition of Hermite-based generalized Euler-Genocchi polynomials and a preliminary study of these polynomials.

Let r and m be integers with $r \geq 0$ and $m \geq 0$. We define Hermite-based generalized Euler-Genocchi polynomials as

$$\sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} = \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\xi\omega + \eta\omega^2} \quad | \omega | < \pi, \quad (13)$$

with ${}_H A_v^{(r,0)}(\xi, \eta) = H_v(\xi, \eta)$ for all $r \geq 0$ and ${}_H A_v^{(r,m)}(\xi, \eta) = 0$ for all $v < rm$. We call the numbers ${}_H A_v^{(r,m)}(0, 0) = A_v^{(r,m)}$ the generalized Euler-Genocchi numbers of order m .

It is obvious that ${}_H A_v^{(r,1)}(\xi, \eta) = {}_H A_v^{(r)}(\xi, \eta)$ and ${}_H A_0^{(r,m)}(\xi, \eta) = 0$ for $r, m \geq 1$. Furthermore, we note that ${}_H A_v^{(0,m)}(\xi, \eta) = {}_H E_v^{(m)}(\xi, \eta)$ and ${}_H A_v^{(1,m)}(\xi, \eta) = {}_H G_v^{(m)}(\xi, \eta)$ are the Hermite-based generalized Euler and Genocchi polynomials as follows (see [1, 8]):

$$\sum_{v=0}^{\infty} {}_H E_v^{(m)}(\xi, \eta) \frac{\omega^v}{v!} = \left(\frac{2}{e^{\omega} + 1} \right)^m e^{\xi\omega + \eta\omega^2} \quad | \omega | < \pi,$$

and

$$\sum_{v=0}^{\infty} {}_H G_v^{(m)}(\xi, \eta) \frac{\omega^v}{v!} = \left(\frac{2\omega}{e^{\omega} + 1} \right)^m e^{\xi\omega + \eta\omega^2} \quad | \omega | < \pi.$$

Proposition 2.1 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi, \eta) = \sum_{\mu=0}^v \binom{v}{\mu} A_{\mu}^{(r,m)} H_{v-\mu}(\xi, \eta). \quad (14)$$

Proposition 2.2 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi, \eta) = \sum_{\mu=0}^{\lfloor \frac{v}{2} \rfloor} \binom{v}{2\mu} A_{v-2\mu}^{(r,m)}(\xi) \mu^l. \quad (15)$$

Theorem 2.3 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi, \eta) = \sum_{\rho=0}^v \sum_{v=0}^{\rho} \binom{v}{\rho} (\xi)_v S_2(\rho, v) {}_H A_{v-\rho}^{(r,m)}(0, \eta). \quad (16)$$

Proof By (1.7), (1.8) and (2.1), we have

$$\begin{aligned}
\sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\xi\omega + \eta\omega^2} \\
&= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\eta\omega^2} (e^\omega - 1 + 1)^\xi \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} \sum_{v=0}^{\infty} (\xi)_v \frac{1}{v!} (e - 1)^v \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} \sum_{v=0}^{\infty} (\xi)_v \sum_{\rho=v}^{\infty} S_2(\rho, v) \frac{\omega^\rho}{\rho!} \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} \sum_{\rho=0}^{\infty} \sum_{v=0}^{\rho} (\xi)_v S_2(\rho, v) \frac{\omega^\rho}{\rho!} \\
&= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \sum_{v=0}^{\rho} \binom{v}{\rho} (\xi)_v S_2(\rho, v) {}_H A_{v-\rho}^{(r,m)}(0, \eta) \right) \frac{\omega^v}{v!}, \tag{17}
\end{aligned}$$

yields the proof of Theorem (2.3). \square

Theorem 2.4 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi + \alpha, \eta) = \sum_{\rho=0}^v \sum_{v=0}^{\rho} \binom{v}{\rho} (\xi)_v S_2(\rho + \alpha, v + \alpha) {}_H A_{v-\rho}^{(r,m)}(0, \eta). \tag{18}$$

Proof Suppose $\xi \rightarrow \xi + \alpha$ in (2.1), we see

$$\begin{aligned}
\sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi + \alpha, \eta) \frac{\omega^v}{v!} &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\xi\omega + \eta\omega^2} e^{\alpha\omega} \\
&= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\eta\omega^2} e^{\alpha\omega} (e^\omega - 1 + 1)^\xi \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} e^{\alpha\omega} \sum_{v=0}^{\infty} (\xi)_v \frac{1}{v!} (e^\omega - 1)^v \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} e^{\alpha\omega} \sum_{v=0}^{\infty} (\xi)_v \sum_{\rho=v}^{\infty} S_2(\rho, v) \frac{\omega^\rho}{\rho!} \\
&= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(0, \eta) \frac{\omega^v}{v!} \sum_{\rho=0}^{\infty} \sum_{v=0}^{\rho} (\xi)_v S_2(\rho + \alpha, v + \alpha) \frac{\omega^\rho}{\rho!} \\
&= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \sum_{v=0}^{\rho} \binom{v}{\rho} (\xi)_v S_2(\rho + \alpha, v + \alpha) {}_H A_{v-\rho}^{(r,m)}(0, \eta) \right) \frac{\omega^v}{v!}, \tag{19}
\end{aligned}$$

yields the proof. \square

Theorem 2.5 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi, \eta) = \sum_{\rho=0}^v \binom{v}{\rho} {}_H A_{v-\rho}^{(r,m-k)}(\xi, \eta) {}_H A_{\rho}^{(r,k)}(0, 0). \quad (20)$$

Proof We observe that

$$\begin{aligned} \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{\xi\omega + \eta\omega^2} \\ &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^{m-k} \left(\frac{2\omega^r}{e^\omega + 1} \right)^k e^{\xi\omega + \eta\omega^2} \\ &= \sum_{v=0}^{\infty} {}_H A_v^{(r,m-k)}(\xi, \eta) \frac{\omega^v}{v!} \sum_{\rho=0}^{\infty} {}_H A_{\rho}^{(r,k)}(0, 0) \frac{\omega^\rho}{\rho!} \\ &= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \binom{v}{\rho} {}_H A_{v-\rho}^{(r,m-k)}(\xi, \eta) {}_H A_{\rho}^{(r,k)}(0, 0) \right) \frac{\omega^v}{v!}. \end{aligned} \quad (21)$$

In view of (2.9), we get the required result. \square

Theorem 2.6 Let $r \geq 0$ and $m \geq 0$. Then

$$\begin{aligned} \sum_{\sigma=0}^v \sum_{\rho=0}^{\lfloor \frac{v-\sigma}{2} \rfloor} \left(\frac{\xi}{\eta^2} - \frac{\eta}{\xi^2} \right)^\rho &\frac{{}_H A_{v-2\rho-\sigma}^{(r,k)}(\xi, \eta) {}_H A_{\sigma}^{(r,k)}}{m! \sigma! (v-\sigma-2\rho)! \eta^\sigma \xi^{v-\sigma-2\rho}} \\ &= \sum_{\rho=0}^v \frac{{}_H A_{v-\rho}^{(r,k)}(\eta, \xi)}{(v-\rho)! \rho! \xi^\rho \eta^{v-\rho}}. \end{aligned} \quad (22)$$

Proof Let $\omega \rightarrow \frac{\omega}{\xi}$ and $m \rightarrow k$, Eq. (2.1) as

$$\sum_{v=0}^{\infty} {}_H A_v^{(r,k)}(\xi, \eta) \frac{\omega^v}{\xi^v v!} = \left(\frac{2(\frac{\omega}{\xi})^r}{e^{\frac{\omega}{\xi}} + 1} \right)^k e^{\omega + \eta \frac{\omega^2}{\xi^2}}. \quad (23)$$

Now, interchanging ξ by η , we have

$$\sum_{j=0}^{\infty} {}_H A_j^{(r,k)}(\eta, \xi) \frac{z^j}{\eta^j j!} = \left(\frac{2(\frac{\omega}{\eta})^r}{e^{\frac{\omega}{\eta}} + 1} \right)^k e^{\omega + \xi \frac{\omega^2}{\eta^2}}. \quad (24)$$

Comparison of (2.11) and (2.12) yields

$$\begin{aligned}
& e^{\xi \frac{\omega^2}{\eta^2} - \eta \frac{\omega^2}{\xi^2}} \left(\frac{2(\frac{\omega}{\xi})^r}{e^{\frac{\omega}{\xi}} + 1} \right)^k \sum_{v=0}^{\infty} {}_H A_v^{(k)}(\xi, \eta) \frac{\omega^v}{\xi^v v!} \\
& = \left(\frac{2(\frac{\omega}{\xi})^r}{e^{\frac{\omega}{\xi}} + 1} \right)^k \sum_{v=0}^{\infty} {}_H A_v^{(k)}(\eta, \xi) \frac{\omega^v}{\eta^v v!} \\
& = \sum_{\rho=0}^{\infty} \frac{(\frac{\xi}{\eta^2} - \frac{\eta}{\xi^2})^{\rho}}{\rho!} \omega^{2\rho} \sum_{\sigma=0}^{\infty} A_{\sigma}^{(r,k)}(\lambda) \frac{\omega^{\sigma}}{\eta^{\sigma} \sigma!} \sum_{v=0}^{\infty} {}_H A_v^{(r,k)}(\xi, \eta) \frac{\omega^v}{\xi^v v!} \\
& = \sum_{\rho=0}^{\infty} A_{\rho}^{(r,k)} \frac{\omega^{\rho}}{\xi^{\rho} \rho!} \sum_{v=0}^{\infty} {}_H A_v^{(r,k)}(\eta, \xi) \frac{\omega^v}{\eta^v v!} \\
& = \sum_{v=0}^{\infty} \left(\sum_{q=0}^j \sum_{l=0}^{\lfloor \frac{j-q}{2} \rfloor} \left(\frac{\xi}{\eta^2} - \frac{\eta}{\xi^2} \right)^l \frac{{}_H A_{j-2l-q}^{(r,k)}(\xi, \eta) A_q^{(r,k)}}{l! q! (j-q-2l)! \eta^q \xi^{j-q-2l}} \right) z^j \\
& = \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \frac{A_{\rho}^{(r,k)} {}_H A_{v-\rho}^{(r,k)}(\eta, \xi)}{(v-\rho)! \rho! \xi^{\rho} \eta^{v-\rho}} \right) \omega^v. \tag{25}
\end{aligned}$$

By (2.1) and (2.16), we obtain the result (2.13). \square

3 Implicit Formulae Involving Hermite-Based Generalized Euler-Genocchi Polynomials

We begin by considering the theorems on implicit summation formulae for Hermite-based generalized Euler-Genocchi polynomials ${}_H A_j^{(r,m)}(\xi, \eta)$ and their special cases.

Theorem 3.1 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_{\sigma+\rho}^{(r,m)}(\zeta, \eta) = \sum_{v,v=0}^{\sigma, \rho} \binom{\sigma}{v} \binom{\rho}{v} (\zeta - \xi)^{v+\nu} {}_H A_{\sigma+\rho-v-v}^{(r,m)}(\xi, \eta). \tag{26}$$

Proof Let $\omega \rightarrow \omega + u$ in (2.1), we see

$$\left(\frac{2(\omega+u)^r}{e^{\omega+u}+1} \right)^m e^{\eta(\omega+u)^2} = e^{-\xi(\omega+u)} \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\xi, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!}, \quad (\text{see [17]}). \quad (27)$$

Again let $\xi \rightarrow \zeta$ in (3.2), we have

$$e^{(\zeta-\xi)(\omega+u)} \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\xi, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!} = \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\zeta, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!}. \quad (28)$$

$$\sum_{N=0}^{\infty} \frac{[(\zeta-\xi)(\omega+u)]^N}{N!} \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\xi, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!} = \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\zeta, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!}, \quad (29)$$

$$\sum_{N=0}^{\infty} f(N) \frac{(\xi+\eta)^N}{N!} = \sum_{v, m=0}^{\infty} f(\omega+m) \frac{\xi^v}{v!} \frac{\eta^m}{m!} \quad (30)$$

in the left hand side becomes

$$\sum_{v, \nu=0}^{\infty} \frac{(\zeta-\xi)^{v+\nu} \omega^\nu u^\nu}{v! \nu!} \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\xi, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!} = \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\zeta, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!} \quad (31)$$

$$\begin{aligned} & \sum_{\sigma, \rho=0}^{\infty} \sum_{v, \nu=0}^{\sigma, \rho} \frac{(\zeta-\xi)^{v+\nu}}{v! \nu!} {}_H A_{\sigma+\rho-v-v}^{(r,m)}(\xi, \eta) \frac{\omega^\sigma}{(\sigma-v)!} \frac{u^\rho}{(\rho-v)!} \\ &= \sum_{\sigma, \rho=0}^{\infty} {}_H A_{\sigma+\rho}^{(r,m)}(\zeta, \eta) \frac{\omega^\sigma}{\sigma!} \frac{u^\rho}{\rho!}. \end{aligned} \quad (32)$$

Thus, by (3.7), we get (3.1). \square

Remark 3.1 Letting $\rho = 0$ in Theorem 3.1, we get

Corollary 3.1 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_{\sigma}^{(r,m)}(\zeta, \eta) = \sum_{v=0}^{\sigma} \binom{\sigma}{v} (\zeta-\xi)^v {}_H A_{\sigma-v}^{(r,m)}(\xi, \eta). \quad (33)$$

Remark 3.2 Let $\zeta \rightarrow \zeta + \xi$ and taking $\eta = 0$, Eq. (3.1) as

$${}_H A_{\sigma+\rho}^{(r,m)}(\zeta + \xi) = \sum_{v,v=0}^{\sigma,\rho} \binom{\sigma}{v} \binom{\rho}{v} \zeta^{v+v} {}_H A_{\sigma+\rho-v-v}^{(r,m)}(\xi). \quad (34)$$

Theorem 3.2 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_{\omega}^{(r,m)}(\xi + \zeta, \eta + u) = \sum_{\mu=0}^v \binom{v}{\mu} {}_H A_{v-\mu}^{(r,m)}(\xi, \eta) \mathbb{H}_{\mu}(\zeta, u). \quad (35)$$

Proof By Eq. (2.1) as

$$\begin{aligned} \left(\frac{2\omega^r}{e^{\omega} + 1} \right)^m e^{(\xi+\zeta)\omega + (\eta+u)\omega^2} &= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} \sum_{\mu=0}^{\infty} \mathbb{H}_{\mu}(\zeta, u) \frac{\omega^{\mu}}{\mu!} \\ \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi + \zeta, \eta + u) \frac{\omega^v}{v!} &= \sum_{v=0}^{\infty} \left(\sum_{\mu=0}^v \binom{v}{\mu} {}_H A_{v-\mu}^{(r,m)}(\xi, \eta) \mathbb{H}_{\mu}(\zeta, u) \right) \frac{\omega^v}{v!}, \end{aligned}$$

which the complete of the proof. \square

Theorem 3.3 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_{\omega}^{(r,m)}(\eta, \xi) = \sum_{\mu=0}^{\lfloor \frac{v}{2} \rfloor} A_{v-2\mu}^{(r,m)}(\eta) \frac{v! \xi^{\mu}}{(v-2\mu)! \mu!}. \quad (36)$$

Proof Suppose $\xi \rightarrow \eta$ and $\eta \rightarrow \xi$ in (2.1), we have

$$\begin{aligned} \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\eta, \xi) \frac{\omega^v}{v!} &= \sum_{v=0}^{\infty} A_v^{(r,m)}(\eta) \frac{\omega^v}{v!} \sum_{s=0}^{\infty} \frac{\xi^s z^{2s}}{s!} \\ &= \sum_{v=0}^{\infty} \left(\sum_{\mu=0}^{\lfloor \frac{v}{2} \rfloor} A_{v-2\mu}^{(r,m)}(\eta) \frac{\xi^{\mu}}{(v-2\mu)! \mu!} \right) \omega^v. \end{aligned}$$

The complete of the proof. \square

Theorem 3.4 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_{\omega}^{(r,m)}(\eta, \xi) = \sum_{v=0}^v \binom{v}{v} A_{v-v}^{(r,m)}(\xi - \zeta) \mathbb{H}_v(\zeta, \eta). \quad (37)$$

Proof Equation (2.1), we note that

$$\begin{aligned} \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e^{(\xi - \zeta)\omega} e^{\zeta\omega + \eta\omega^2} &= \sum_{v=0}^{\infty} A_v^{(r,m)}(\xi - \zeta | \lambda) \frac{\omega^v}{v!} \sum_{v=0}^{\infty} \mathbb{H}_v(\zeta, \eta) \frac{\omega^v}{v!} \\ \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} &= \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} A_{v-v}^{(r,m)}(\xi - \zeta) \mathbb{H}_v(\zeta, \eta) \right) \frac{\omega^v}{v!}, \end{aligned} \quad (38)$$

yields the proof.

□

Theorem 3.5 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi + 1, \eta) = \sum_{v=0}^v \binom{v}{v} \xi^v {}_H A_{v-v}^{(r,m)}(\xi, \eta). \quad (39)$$

Proof In (2.1), we have

$$\begin{aligned} \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi + 1, \eta) \frac{\omega^v}{v!} &- \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} \\ &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m (e^\omega - 1) e^{\xi\omega + \eta\omega^2} \\ &= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} \left(\sum_{v=0}^{\infty} \xi^v \frac{\omega^v}{v!} - 1 \right) \\ &= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} \sum_{v=0}^{\infty} \xi^v \frac{\omega^v}{v!} - \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} \\ &= \sum_{v=0}^{\infty} \sum_{v=0}^v \binom{v}{v} \xi^v {}_H A_{v-v}^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!} - \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta) \frac{\omega^v}{v!}, \end{aligned}$$

which the complete of the proof.

□

4 Symmetry Identities

In this section, we give general symmetry identities for Hermite-based generalized Euler-Genocchi polynomials ${}_H A_j^{(r,m)}(\xi, \eta)$ by applying the generating functions (1.2) and (2.1).

Theorem 4.1 Let $r \geq 0$ and $m \geq 0$. Then

$$\begin{aligned} & \sum_{v=0}^v \binom{v}{v} a^{v-v} b^v {}_H A_{v-v}^{(r,m)} (b\xi, b^2\eta) {}_H A_v^{(r,m)} (a\xi, a^2\eta) \\ &= \sum_{v=0}^v \binom{v}{v} b^{v-v} a^v {}_H A_{v-v}^{(r)} (a\xi, a^2\eta) {}_H A_v^{(r,m)} (b\xi, b^2\eta). \end{aligned} \quad (40)$$

Proof Let

$$F(\omega) = \left(\frac{4(ab\omega)^r}{(e^{a\omega} + 1)(e^{b\omega} + 1)} \right)^r e^{ab\xi\omega + a^2 b^2 \eta \omega^2}.$$

$$F(\omega) = \sum_{j=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} a^{v-v} b^v {}_H A_{v-v}^{(r,m)} (b\xi, b^2\eta) {}_H A_v^{(r,m)} (a\xi, a^2\eta) \right) \frac{\omega^v}{v!}.$$

On the similar lines, we can show that

$$F(\omega) = \sum_{j=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} b^{v-v} a^v {}_H A_{v-v}^{(r)} (a\xi, a^2\eta) {}_H A_v^{(r,m)} (b\xi, b^2\eta) \right) \frac{\omega^v}{v!},$$

which implies the desired result. \square

Remark 4.1 Letting $b = 1$, Eq. (4.1), we have

Corollary 4.1 Let $r \geq 0$ and $m \geq 0$. Then

$$\sum_{v=0}^v \binom{v}{v} a^{v-v} {}_H A_{v-v}^{(r,m)} (\xi, \eta) {}_H A_v^{(r,m)} (a\xi, a^2\eta) \quad (41)$$

$$= \sum_{v=0}^v \binom{v}{v} a^v {}_H A_{v-v}^{(r)} (a\xi, a^2\eta) {}_H A_v^{(r,m)} (\xi, \eta). \quad (42)$$

Theorem 4.2 Let $r \geq 0$ and $m \geq 0$. Then

$$\sum_{v=0}^v \sum_{\theta=0}^{a-1} \sum_{\rho=0}^{b-1} \binom{v}{v} a^{v-v} b^v {}_H A_{v-v}^{(r,m)} \left(b\eta + \frac{b}{a}\theta + \rho, b^2\zeta \right) {}_H A_v^{(r,m)} (a\eta)$$

$$= \sum_{v=0}^v \sum_{\rho=0}^{a-1} \sum_{\theta=0}^{b-1} \binom{v}{\rho} b^{v-\rho} a^v {}_H A_{v-\rho}^{(r,m)} \left(a\eta + \frac{a}{b}\theta + \rho, a^2 \zeta \right) A_v^{(r,m)}(b\eta). \quad (43)$$

Proof Let

$$H(\omega) = \left(\frac{4(ab\omega)^r}{(e^{a\omega} + 1)(e^{b\omega} + 1)} \right)^m \frac{(e^{ab\omega} - 1)^2}{(e^{a\omega} - 1)(e^{b\omega} - 1)} e^{ab(\xi+\eta)\omega + a^2 b^2 \zeta \omega^2} \quad (44)$$

$$= \left(\frac{2(a\omega)^r}{e^{a\omega} + 1} \right)^m e^{ab\xi\omega + a^2 b^2 \zeta \omega^2} \sum_{\theta=0}^{a-1} e^{b\omega} \left(\frac{2(b\omega)^r}{e^{b\omega} + 1} \right)^m e^{ab\eta\omega} \sum_{\rho=0}^{b-1} e^{a\omega\rho} \quad (45)$$

$$= \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \sum_{\theta=0}^{a-1} \sum_{\rho=0}^{b-1} \binom{v}{\rho} a^{v-\rho} b^v {}_H A_{v-\rho}^{(r,m)} \left(b\eta + \frac{b}{a}\theta + \rho, b^2 \zeta \right) A_v^{(r,m)}(a\eta) \right) \frac{\omega^v}{v!}.$$

On the other hand, we have

$$H(\omega) = \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \sum_{\rho=0}^{a-1} \sum_{\theta=0}^{b-1} \binom{v}{\rho} b^{v-\rho} a^v {}_H A_{v-\rho}^{(r,m)} \left(a\eta + \frac{a}{b}\theta + \rho, a^2 \zeta \right) A_v^{(r,m)}(b\eta) \right) \frac{\omega^v}{v!},$$

which provide the desired result. \square

Theorem 4.3 Let $r \geq 0$ and $m \geq 0$. Then

$$\begin{aligned} & \sum_{\rho=0}^v \binom{v}{\rho} a^{v-\rho} b^{\rho} {}_H A_{v-\rho}^{(r,m)}(b\xi, b^2 \zeta) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(b-1) A_{\rho-\theta}^{(r,m)}(a\eta) \\ &= \sum_{\rho=0}^v \binom{v}{\rho} b^{v-\rho} a^{\rho} {}_H A_{v-\rho}^{(r,m)}(a\xi, a^2 \zeta) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(a-1) A_{\rho-\theta}^{(r,m)}(b\eta). \end{aligned} \quad (46)$$

Proof From (1.12) and (2.1), we see

$$\begin{aligned} Q(\omega) &= \left(\frac{2(ab\omega)^r}{(e^{a\omega} + 1)(e^{b\omega} + 1)} \right)^m \frac{(e^{ab\omega} - 1)^2}{(e^{a\omega} - 1)(e^{b\omega} - 1)} e^{ab(\xi+\eta)\omega + a^2 b^2 \zeta \omega^2} \\ &= \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(b\xi, b^2 \zeta) \frac{(a\omega)^v}{v!} \sum_{\theta=0}^{\infty} S_{\theta}(b-1) \sum_{\rho=0}^{\infty} A_{\rho}^{(r,m)}(a\eta) \frac{(b\omega)^{\rho}}{\rho!} \end{aligned}$$

$$= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \binom{v}{\rho} a^{v-\rho} b^{\rho} {}_H A_{v-\rho}^{(r,m)}(b\xi, b^2\zeta) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(b-1) A_{\rho-\theta}^{(r,m)}(a\eta) \right) \frac{\omega^v}{v!}.$$

On the other hand,

$$Q(\omega) = \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \binom{v}{\rho} b^{v-\rho} a^{\rho} {}_H A_{v-\rho}^{(r,m)}(a\xi, a^2\zeta) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(a-1) A_{\rho-\theta}^{(r,m)}(b\eta) \right) \frac{\omega^v}{v!}.$$

The complete proof of this theorem. \square

5 Concluding Remarks

In 2003, Dattoli et al. [4] introduced the truncated exponential polynomials by means of the generating function as

$$\sum_{v=0}^{\infty} H_v(\xi, \eta|s) \frac{\omega^v}{v!} = e_s(\xi\omega + \eta\omega^2). \quad (47)$$

By using (1.8) and (5.1), we can define truncated Hermite-based generalized Euler-Genocchi polynomials by means of the following generating function as

$$\left(\frac{2\omega^r}{e^{\omega} + 1} \right)^m e_s(\xi\omega + \eta\omega^2) = \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta : s) \frac{\omega^v}{v!}, \quad (48)$$

where $e_s(\xi)$ is the truncated exponential function defined by [4]

$$e_s(\xi) = \sum_{v=0}^s \frac{\xi^v}{v!}, \quad \frac{e^{\xi\omega}}{1 - \omega} = \sum_{v=0}^{\infty} \omega^v e_v(\xi) \quad |\omega| < 1. \quad (49)$$

Theorem 5.1 Let $r \geq 0$ and $m \geq 0$. Then

$${}_H A_v^{(r,m)}(\xi, \eta : s) = \sum_{v=0}^v \binom{v}{v} A_{v-v}^{(r,m)} \mathbb{H}_v(\xi, \eta : s). \quad (50)$$

Proof By using definition (1.8) (5.1) and (5.2), we get

$$\begin{aligned}
 \sum_{v=0}^{\infty} {}_H A_v^{(r,m)}(\xi, \eta : s) \frac{\omega^v}{v!} &= \left(\frac{2\omega^r}{e^\omega + 1} \right)^m e_s(\xi\omega + \eta\omega^2) \\
 &= \sum_{v=0}^{\infty} A_v^{(r,m)} \frac{\omega^v}{v!} \sum_{v=0}^{\infty} \mathbb{H}_v(\xi, \eta : s) \frac{\omega^v}{v!} \\
 &= \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} A_{v-v}^{(r,m)} \mathbb{H}_v(\xi, \eta : s) \right) \frac{\omega^v}{v!}.
 \end{aligned}$$

The complete proof of the theorem. \square

References

1. Andrews LC (1985) Special functions for engineers and mathematicians. Macmillan. Co., New York
2. Bell ET (1934) Exponential polynomials. Ann Math 35:258–277
3. Dattoli G, Chiccoli C, Lorenzutta S, Maino G, Torre A (1994) Theory of generalized Hermite polynomials. Comput Math Appl 28(4):71–83
4. Dattoli G, Cesarano C, Sacchetti D (2003) A note on truncated polynomials. Appl Math Comput 134(2–3):595–605
5. Frontczak R, Tomovski Ž (2019) Generalized Euler-Genocchi polynomials and Lucas numbers. Integers 19:1–17
6. Khan WA, Muhiuddin G, Muhyi A, Al-Kadi D (2021) Analytical properties of type 2 degenerate poly-Bernoulli polynomials associated with their applications. Adv Diff Equ 2021(420):1–18
7. Khan WA, Pathan MA (2019) On generalized Lagrange-Hermite-Bernoulli and related polynomials, Acta et. Commentationes Universitatis Tartuensis de Mathematica 23(2):211–224
8. Khan WA, Haroon H (2016) Some symmetric identities for the generalized Bernoulli, Euler and Genocchi polynomials associated with Hermite polynomials. Springer Plus 5:1–21
9. Khan WA, Muhyi A, Ali R, Alzobydi KAH, Singh M, Agarwal P (2021) A new family of degenerate poly-Bernoulli polynomials of the second kind with its certain related properties. AIMS Math 6(11):12680–12697
10. Khan WA, Acikgoz M, Duran U (2020) Note on the type 2 degenerate multi-poly-Euler polynomials. Symmetry 12:1–10
11. Khan WA, Ali R, Alzobydi KAH, Ahmed A (2021) A new family of degenerate poly-Genocchi polynomials with its certain properties. J Funct Spaces 2021, Article ID 6660517, 8p
12. Khan WA (2022) A note on q -analogues of degenerate Catalan-Daehee numbers and polynomials. J Math 2022, Article ID 9486880, 9p
13. Khan WA (2022) A note on q -analogue of degenerate Catalan numbers associated with p -adic integral on \mathbb{Z}_p . Symmetry 14(1119):1–10
14. Khan WA (2022) On generalized Lagrange-based Apostol type and related polynomials. Kragujevac J Math 46–22:865–882

15. Khan WA (2022) A new class of higher-order hypergeometric Bernoulli polynomials associated with Hermite polynomials. *Boletim da Sociedade Paranaense de Matematica* 40:1–14. <https://doi.org/10.5269/bspm.51845>
16. Khan W (2022) A, A study on q -analogue of degenerate $\frac{1}{2}$ -Changhee numbers and polynomials. *SE Asian J Math Math Sci* 18(2):1–12
17. Pathan MA, Khan WA (2021) A new class of generalized polynomials associated with Hermite and poly-Bernoulli polynomials. *Miskolc Math J* 22(1):317–330

6G-Enabled Internet of Medical Things



Sumit Singh Dhanda, Tarun Kumar Sharma, Brahmjit Singh, Poonam Jindal, and Deepak Panwar

Abstract Health care is the foremost concern for any country. UN has also fixed “good health and well-being” as its third Sustainable Development Goal. COVID-19 era has shown the vulnerability of health infrastructure worldwide. As per WHO, aging population of the world would put enormous pressure on health infrastructure in coming decades. With lack of healthcare professionals and limited budget to achieve UNSDG 3, IoMT with its huge umbrella of medical services and applications can address these issues. The 5G has been deployed in many countries, and vision for 6G has been finalized by different research groups. The 6G will help Internet of Medical Things to realize its full potential. In this work, an open “6G-enabled IoMT” architecture has been presented. It can integrate a wide number of services. The limitation of 5G in catering to such a network is also highlighted. The challenges and open issues are presented in light of services and applications that can be provided by the purposed architecture. A brief overview of 6G is also provided.

Keywords Internet of medical things (IoT) · eHealth · m-Health · 6G · Deep neural network (DNN) · Artificial intelligence (AI) · Information security

1 Introduction

WHO says that the population of people above 60 years will be 22% of the total by year 2050 and 77% of them will be suffering from at least two chronic-disease [1]! Though, the healthcare infrastructure and services have improved to exemplary levels in last two decades due to enormous technological advancements. Increasing

S. S. Dhanda (✉) · B. Singh · P. Jindal

Department of Electronics and Communication, National Institute of Technology, Kurukshetra, Haryana, India

e-mail: dhandasumit@gmail.com

T. K. Sharma

Shobhit University, Meerut, India

D. Panwar

Manipal University Jaipur, Jaipur, India

world's population has posed a critical challenge to the medical system. Current health infrastructure faces three critical issues.

1. The conventional patient–doctor appointment is losing its effectiveness due to huge number of unplanned visits to the doctors. It is the outcome of intensification of aging process.
2. Physical limitation of hospital system. When medical inpatient bed occupancy rate exceeds 100%, it leads to unsatisfied service to the patients. Especially, during COVID-19 times, bed occupancy continuously exceeded 100% mark.
3. Huge investment is required in preventative and long-term cares (LTCs) to prepare for the demographic challenge.

With the increase in demand for LTC services, it has become hard to find qualified healthcare employees for the same. Hence, the need is to change on-site medical interaction to online medical interaction. This challenge also brings out the opportunity for the technology to address this issue. National Health Expenditure of USA was \$ 4.1 trillion in 2020, with per capita expenditure at \$12,530. It is 19.7% of their GDP and is expected to grow at a rate of 5.4% per year to reach \$ 6.2 trillion by 2028 [2]. Use of technology can reduce this expenditure and save a lot of money apart from relieving pressure from health infrastructure. As per Precedence Research [3], the global valuation of smart healthcare market was \$145.32 billion in 2020, and at a rate of 19.7% per year, it is expected to reach \$ 482.25 billion by 2027. Its highest portion will be occupied by m-Health at 47%, while North America has a share of 34.81%.

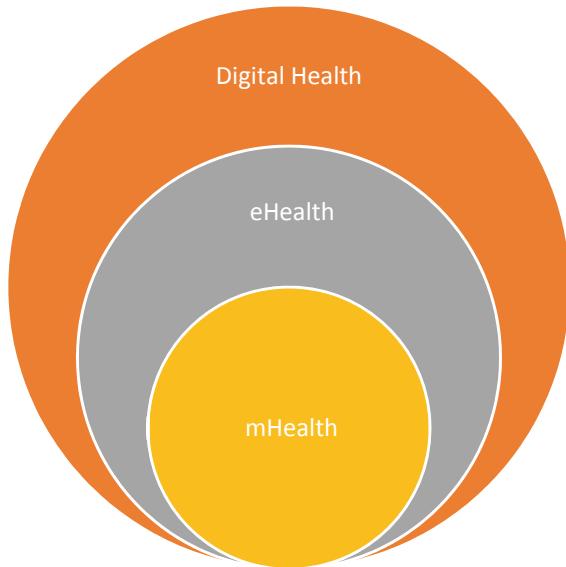
Digital Health is a very broad term which is eHealth and uses big data, artificial intelligence and other advanced computing technologies to provide better health care. It is shown in Fig. 1 with its subsets. Improvement in the availability and effectiveness of healthcare services with the use of information and communication technology (ICT) is referred as “eHealth”. The m-Health is a subset of eHealth. When telecommunication technology and multimedia services are used to disseminate health information and deliver health services, it is called m-Health.

With eHealth, there has been a paradigm shift to a “predictive and preventive” healthcare model from the older curative healthcare model. Data is collected and transmitted from the body and stored and shared for predictive analysis through examination. It results in “predictive and preventive” model. In Internet of Medical Things (IoMT), medical equipment are embedded sensors to connect with hospitals and patients via internet to achieve “predictive and preventive” medical model. But IoMT tries to encompass whole of the Digital Health umbrella by catering vast number of medical services.

In this work, the issues of 5G have been explored that will create hindrance in realizing the full potential of IoMT. While highlighting these problems, solution is also searched out of the expected capabilities and vision of 6G that has been finalized by Hexa-X, an industry-academia consortium in Europe.

Rest of the paper is organized as follows: State of the art of IoMT field has been discussed in Sect. 2. An open 6G-enabled IoMT architecture has been presented, and limitations of 5G, that will limit this architecture from achieving its full potential,

Fig. 1 Digital health and its subset as per WHO [1]



have been discussed in Sects. 3. Section 4 provides the details of different services and application while differentiating between these two. Section 5 presents the open issues and challenges that IoMT faces currently. Finally, conclusions are drawn and future work directions are mentioned in Sect. 6.

2 Related Work

In [4], authors have presented a deep learning algorithm that can be used in 6G-enabled IoMT for the detection of lung nodules with low-dose computed tomography (CT). A combination of residual and convolutional neural network (CNN) has been used for the purpose of learning of lung nodules. This information is transferred in a memory, and features are optimized for the early screening of lung cancer. Authors expect this model to be more reliable and accurate than existing methods. The technique can further be used for the exchange and analysis of medical data and diagnosis results which will enable an effective and convenient online diagnosis in the future.

The paradigm of health care has shifted from traditional nuances. With technological advancements, the health care has become smart. In [5], authors have discussed about evolution of this emerging concept.

To realize Internet of Medical Things, information of all the participating persons and things must be collected and transmitted to cloud and processed to get the required analysis out of it. In [6], authors have presented an intelligent medical interactive system that can identify and recognize the equipment, doctors, samples, patients

and medical records. It has three layers in this framework to improve the patient's experience. It is a hospital-specific model which does not have medical record sharing for improved diagnosis that is missing.

Optimum utilization of medical infrastructure and efficiency of medical diagnosis are two key areas for the IoMT. A smart medical service system (SMSS) architecture is presented in [7] to address these two issues.

Artificial intelligence (AI) will play a key role in the IoMT. It will help in effective diagnosis and service discovery as well. In [8], authors have explored the use of AI in analysis of MRI. In [9], authors have used deep learning techniques to analyze the discomfort of a person. While deep learning is applied on CT images to predict lung cancer in [10], IoT is used for fitness monitoring in [11], and details of various sensors are presented. Artificial intelligence is used for EEG classification in [12]. Authors also provision security of the data in their work. Use of IoT in dementia care has been demonstrated in [13].

Storage of medical data is an important issue because IoMT will generate massive big data whose storage and processing will be key challenge. In [14], authors have presented a multi-cloud platform named tri-storage failure recovery system (Tri-SFRS). It also improves the reliability in IoMT by ensuring storage, access and scalability.

3 IoMT and Enabling Technologies

United Nations has defined 17 Sustainable Development Goals which must be achieved by 20xx. Among the seventeen United Nation Sustainable Development Goals (UNSDGs), Goal 3, which is “good Health and well-being”, can be directly achieved with the help of Internet of Medical Things (IoMT), to tackle the concept of sustainability from the various angles (environmental, societal and economic) and come within the scope of the UN SDGs. Internet of Medical Things (IoMT) is an environment where IoT use-cases are created for the delivery of medical services. Now, the spectrum of such services is becoming vast and complex. In such complex use-cases or business scenarios, IoT is used to improve the efficiency of medical diagnosis and improve the utilization of medical resources.

Lack of convenient and practical information about patient's condition can lead to dissatisfactory/erroneous treatment. Timely communication and supervision can help improve the efficacy of treatment. Auxiliary diagnostic for early screening of different diseases will help in reducing the risk of life. It will also help in improving the effectiveness of medical diagnosis along with raising the patient's requirements for the convenience.

Figure 2 shows the architecture for IoMT. It has five layers. Bottom layer is “perception layer” which is constituted by different sensing devices and machines that are embedded with sensors. These all enabled to connect with internet via a gateway. These equipment, machines and devices use different link layer technologies like ZigBee, Wi-Fi, near-field communication (NFC) and 4G/5G mobile networks. The

main task of this layer is to collect the medical data and transmit it to next level. In IoMT, the requirement of accuracy and reliability for devices, media and protocols is very stringent. Because slight variation in information may prove fatal. These effective and fast communications result in the improvement of work efficiency. These advancements have resulted in higher expectations from patients and doctors. They expect that medical records should be shared between different hospitals so that medical diagnosis can be online and its results be more effective and accurate. It will also help in contacting the experts in relevant hospitals. Information related to differential diagnoses, such personal data and family's medical history, electronic health records (EHR) and disease condition, may be shared for better treatment. At present, the IoMT has been used in many medical fields; hence, it will generate massive data. Its transmission and processing are two main constraints on IoMT development.

At adaptation layer, gateway plays an important role of translator for these resource-constrained devices. It will also perform the role of broker between service provider servers and equipment seeking information, e.g., in-house monitoring and automatic drug delivery systems. Network layer also demands for adaptive protocols with high security. Here, different types of networks will be interacting with each other as shown in Fig. 2. Service aggregation layer shown with cloud storage will be used to provide the base for service-oriented architectures and interoperability of

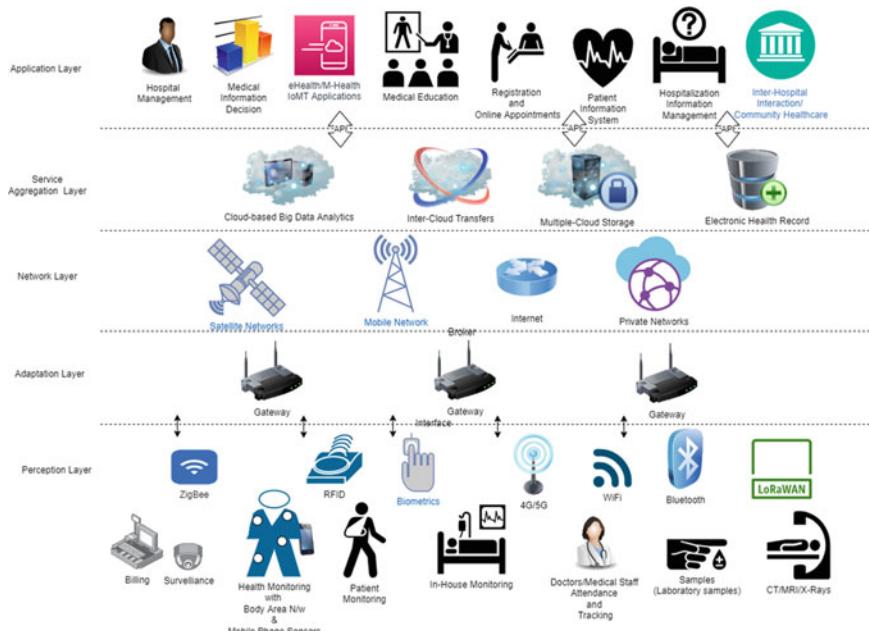


Fig. 2 IoMT architecture

different systems. AI algorithms will provide the capabilities to serve multiple applications with same information by extracting context-aware information and enabling semantic interoperability. Finally, applications can be multi-faceted. It can be monitoring of doctors and staff by hospital management or in-house patient monitoring. Online registrations, medical education, hospital information systems, etc. are other mentions.

AI technology has matured with the development of many intelligent algorithms. It has helped in achieving the structural redundancy in medical big data. Arrival of 5G technology has helped in transmission of medical images and data in real time. The combination of AI and 5G guarantees the improvement of the IoMT system. There are two important advantages of 5G over 4G in context of IoMT: (i) high bandwidth which allows 5G to achieve transmission rate up to 10 Gbps, but it is only for stationary user in uncrowded environment. In this way, 5G is hundred times faster than 4G (100 Mbps); (ii) low delay. Ideally, 5G can achieve 1 ms of delay, while in real life, it is around 10 ms in comparison to 100 ms of 4G. Although, 5G can fulfill the needs of big data transmission and other security requirements. But 5G still has some room for improvement in terms of transmission immediacy and the coverage to dead corner of the network. A case study can be considered to understand the problem in “5G-enabled IoMT”. Assuming that medical images and big data, take 0’s, a delay of 10 ms will be introduced by 5G, and the best treatment opportunity might be missed.

Many research organizations are decided on the vision for 6G. Some of the most important use-cases involve massive twinning of cities and factories; very high link rates up to 100 Gbps; high scalability; accurate situational awareness with <1 cm of error and delays of 100 μ s, all are to be achieved in 6G. Now, if the “6G-enabled IoMT” is considered under the above-mentioned conditions, delay of 100 μ s gives a safe window in emergency medical treatment. (If the computing and backhaul delay is considered, it can be in seconds for 5G). Higher bandwidths in 6G will also enable the communication and analysis of videos in 4 K, 8 K and even 16 K resolutions.

Artificial intelligence will be at the helm of the affairs in 6G where a large number of use-cases will be provided to users with the help of different AI techniques. Cooperative robots or Cobots are one such example that can be used for the support of elderly and incapacitated persons.

AI-based medical service program; AI-based medical service data management system; AI-based medical service data expert processing system; AI-based medical service information system; AI-based medical consultation system; AI-based medical monitoring system; AI-based medical terminal management; AI-based medical service telemedicine; smart buildings in smart lighting, AC control, etc. But increased latency unsuitable for the applications such as real-time vehicular control or smart grid applications.

4 Services and Applications

Though it is not a standard definition, there is a basic difference between services and applications. Services are general set of utilities that are used as the base to develop specific solutions or applications. Figure 3 shows various IoMT services that can be provided.

Ambient-assisted living (AAL) is a separate IoMT service which uses AI to provide convenient and independent life to elderly people and incapacitated individuals. Cobots, in 6G as mentioned in previous section, can help them with greater autonomy by providing servant like assistance. Adverse drug reaction (ADR) is a situation that may arise out of wrong combination of drug or by intake of a drug. An intelligent pharmaceutical information system can be designed which corroborates the drug with allergy profile and EHR of the patient.

Community healthcare is another cooperative network structure created by IoMT by incorporating a municipal hospital with a rural community or nearby residential area. It is like a network of networks which is again another capability that 6G must acquire. Children health information (CHI) is an AI-based IoMT service that will help in identifying the mental, emotional and behavioral problems of children and improve their cognitive skills [15, 16].

Wearable device access (WDA) integration of non-intrusive sensors for data collection which can form the body area network is a challenge for researchers. As heterogeneous nature of such products poses numerous challenges. Semantic medical access (SMA) is another important service which must include AI technology to ensure the interoperability of semantics and ontologies in sharing of medical information and its proper use in different use-cases. Embedded gateway configuration (EGC) will provide the connection between medical equipment, network nodes and internet. Embedded context prediction (ECP) enables the context-aware applications [17].

IoMT Services	Ambient Assisted Living
	Adverse Drug Reactions
	Community Healthcare
	Indirect Emergency Healthcare
	Embedded Gateway Configuration
	Embedded Context prediction
	Children Health Care
	Wearable Device Access
	Semantic Medical Access

Fig. 3 Services for Internet of Things for Healthcare

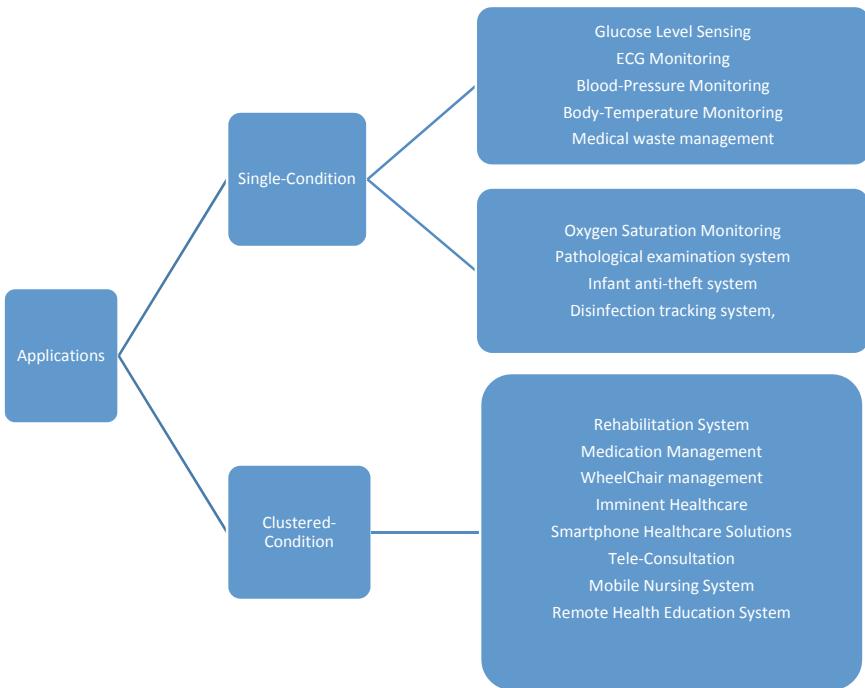


Fig. 4 Application of IoT for Healthcare

IoMT applications are user-centric and can be single-condition or clustered-condition applications. Some applications in which a sample or image needs to be collected for the medical data are single-condition applications. Here, the focus is only on the single function which is called single-condition applications. These are shown in Fig. 4 under the same segment. Clustered-condition applications achieve more than one objective such as rehabilitation system, medication management, wheel chair management, imminent health care, smartphone healthcare solutions, and tele-consultation.

5 Challenges and Open Issues

As the issues regarding 6G and 5G networks are already discussed in the earlier sections, this section will focus on the challenges for IoMT.

- *Standardization* is an important issue to be addressed. As large number of vendors are producing different IoMT products, interoperability can become an issue.

- *Low-power protocols* are the most important requirement. Small devices cannot transmit continuously. Handshakes and other procedure should be such that life time of such small devices should not be affected.
- *Scalability* issues will always be there in such a network. A large number of nodes will be added even for single patient. Hence, to cater a community health care, scalability must be included not only for perception layer but for application and service aggregation layer as well.
- *Quality of Service (QoS)* should be very high and must be ensured through increased network and device reliability. Healthcare services cannot be tolerant to link failures or high noise data. Hence, there should be an alternative plan for the emergency and algorithms to minimize such situations.
- *Data protection* as massive data would be captured and transmitted via IoMT. Hence, security measures must be designed to address cloud security. There must be schemes that can provide security on low costs as most of the devices are resource-constrained. Physical security of devices is also important due to open deployment of sensors. Researchers must also design the means to secure personal data with assured availability at the time of requirement.
- *Mobility* as many of the emergency applications are mobile in nature for IoMT. One must design protocols and security procedures to address these issues. For example, 6LoWPAN must be standardized to include mobility.
- *Inclusion of Medical Experts in design process*. Medical experts who belong to an authorization body must be included in the applications and system design process.

6 Conclusion

In this work, a brief discussion is carried out on the need and impact of IoMT. An open “6G-enabled IoMT” architecture that can be included to integrate a wide number of services is presented. The limitations of 5G in catering to such a network are also highlighted. The challenges and open issues are presented in light of services and applications provided by such “6G-enabled IoMT architecture”. A brief overview of 6G is also provided, and its capabilities that will inherently support IoMT architecture and services are presented. In the future, we will develop the technologies to realize this architecture and try to implement a working model after that.

References

1. Philip NY et al (2021) Internet of things for in-home health monitoring systems: current advances, challenges, and future directions. IEEE J Select Area Commun 39(2):300–309
2. National Health Expenditures 2020 Highlights Retrieved from: <https://www.cms.gov/files/document/highlights.pdf>
3. <https://www.precedenceresearch.com/smart-healthcare-market>

4. Wei W, Liu F, Zhi X, Zhang T, Huang C (2021) An integrated deep learning algorithm for detecting lung nodules with low-dose CT and its application in 6G-enabled internet of medical things. *IEEE Internet of Things J* 8(7):5274–5283
5. Zhu H, Wu CK, Koo CH, Tsang YT, Liu Y, Chi HR, Tsang KF Smart healthcare in the era of internet of things. *IEEE Consumer Electronics Magazine* 26–30. <https://doi.org/10.1109/MCE.2019.2923929>
6. Jia N, Zheng C Design of intelligent medical interactive system based on internet of things and cloud platform. In: 2018 10th International conference on intelligent human–machine systems and cybernetics, 28–31
7. Lu S, Wang A, Jing S, Shan T, Zhang X, Guo Y, Liu Y (2019) A study on service-oriented smart medical systems combined with key algorithms in the IoT environment. *China Commun* 235–249
8. Daoud MK, Otair M (2020) The role of artificial intelligence and the internet of things in the development of medical radiology (an experimental study on magnetic resonance imaging). In: 2020 International conference on intelligent computing and human-computer interaction (ICHCI) | 978-1-6654-2316-8/20/\$31.00 ©2020 IEEE, 17–20. <https://doi.org/10.1109/ICHC51889.2020.00011>
9. Ahmed I, Jeon G, Piccialli F (2021) A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of internet of things. *IEEE Internet of Things J* 8(13):10318–10326
10. Han T, Nunes VX et al Internet of medical things based on deep learning techniques for segmentation of lung and stroke regions in CT scans. Digital Object Identifier <https://doi.org/10.1109/ACCESS.2020.2987932>
11. Qiu Y, Zhu X, Lu J Fitness monitoring system based on internet of things and big data analysis. Digital Object Identifier. <https://doi.org/10.1109/Access.2021.3049522>
12. Singh R, Ahmed T, Singh AK, Chanak P, Singh SK (2021) SeizSClas: an efficient and secure internet-of-things-based EEG classifier. *IEEE Internet of Things J* 8(8):6214–6221
13. Enshaeifar S, Barnaghi P et al (2018) The internet of things for dementia care. *IEEE Internet Comput IEEE Comput Soc* 8–17
14. Cao R, Tang Z, Liu C, Veeravalli B (2020) A scalable multicloud storage architecture for cloud-supported medical internet of things. *IEEE Internet Things J* 7(3):1641–1654
15. Uusitalo MA et al (2021) 6G vision, value, use cases and technologies from European 6G flagship project hexa-X. *IEEE Access* 9:160004–160021
16. Sharma N, Panwar D (2021) Advance security and challenges with intelligent IoT devices. In: Proceedings of second international conference on smart energy and communication. Springer, Singapore, pp 177–189
17. Sharma N, Panwar D (2020) Green IoT: advancements and sustainability with environment by 2050. In: 2020 8th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO). IEEE, 1127–1132

An Inductively Degenerated LNA for ISM Applications: Design and Performance Comparison



P. K. Verma  , Himanshu Katiyar, Prashant Pandey, Vikas, and D. K. Tripathi

Abstract This paper states that an inductive source degeneration topology has been used to offer precise low noise figure (LNA), low power consumption (LPC), and high forward gain (S21). The system design and simulation of the proposed LNA is operating over the band of frequency 2.4–2.5 GHz, which is very much applicable for industrial, scientific, and medical applications. The stated work is standing on fall in noise figure through rising in power gain at low power consumption as paralleled to published preceding analysis at the power supply of 1.2 V. The system simulation was gone through using Advance Design System (ADS) software. The three different schematic designs have been presented for design and performance analysis, which start with a very basic LNA to the proposed one.

Keywords S parameters · Noise figure · Power consumption · ADS · ISM applications

1 Introduction

Through the highly advance progress in the field in communication, several varieties of wireless strategies are required. LNA is the very much critical element of radio frequency (RF) receiver [1–3]. LNA escalates the gain level and upturns the noise

P. K. Verma  · H. Katiyar · P. Pandey · D. K. Tripathi
Electronics Engineering Department, Rajkiya Engineering College, Sonbhadra, UP, India
e-mail: pkverma@recsonbhadra.ac.in

H. Katiyar
e-mail: hkatiyar@recsonbhadra.ac.in

P. Pandey
e-mail: ppandey@recsonbhadra.ac.in

D. K. Tripathi
e-mail: dktripathi@recsonbhadra.ac.in

Vikas
Institute of Electronics Engineering, National Tsing Hua University, Hsinchu, Taiwan
e-mail: vikas.rohilla11@gmail.com

figure (NF) of the input signal. Whereas scheming an LNA, highly recommended high gain, high IIP3, low return loss, noise figure, and consumption of low power are taken into contemplation. In command to accomplish the essential gain, it is required to minimize these parameters such as input reflection coefficient (S_{11}) and output reflection coefficient (S_{22}) [4, 5]. In the investigation, LNA design has been quantified with lessening in power consumption and NF as well as increase in gain (S_{21}) in [6–13].

There are several topologies for LNA design given as a very first resistive matching (RM), second is common gate amplifier (CGA), third is shunt series feedback common source (SSFCS), fourth is inductive source degeneration (ISD), and final one is the cascode inductive source degeneration (C-ISD) [6]. The proposed paper contributes an ISD topology for figuring out, high S_{21} , low NF, and LPC. The recommended design has been designed and imitated over an ADS software.

An ADS software is communication system and microwave software industrialized by Keysight Technologies (formerly Agilent Technologies). It has much more dominant capacity in RF circuit scheme, and simulation of different model has been investigated. In proposed model of simulation, scattering parameters provide evidence almost stability, gain, linearity, and noise figure.

The presentation of paper is as tracks. In Sect. 2, altered forms of topologies of LNA are offered. Section 3 designates the input/output identical of LNA, and Sect. 4 describes the three different designs, and simulated outcomes are in Sect. 5. Section 6 contributes the conclusion of projected LNA.

2 Types of Topologies

The several number of topologies for LNA design are recommended as RM, CGA, SSCFS, ISD, and C-ISD. These retained topologies have their specific qualities/shortcomings which are charted in Table 1.

Table 1 Qualities/shortcomings of LNA topologies

Type of topology	Qualities	Shortcomings
RM [17, 22]	Broadband amplifier	Addition of noise from resistor
CGA [16, 18, 20, 22]	The input impedance (Z_{in}) = $1/g_m$, which is very informal to have 50Ω	The impedance swings by bias current
SSCFS [17–19, 27]	Broadband amplifier	Calculation of noise from resistor
ISD [14, 19, 22, 24, 25]	The source and gate inductors mark Z_{in} 50Ω and not providing further noise from input	The inductor provides very low frequency and low isolation at off chip
C-ISD [15, 21, 23, 25]	Low NF, higher S_{21} and also provide good isolation of input/output	The inductor provides very low frequency at off chip

In the above-projected topologies, ISD arrangement is finest from other topologies. So, it has preferred ISD, with inductive load. It offers enough gain deprived of adding major noise and also affords enhanced matching in evaluation to rest of the topologies.

3 Input/Output Matching

A very basic design of the LNA will be matching of input, which assures that most of the input signal that is accepted hooked on LNA for meting out and in mandate to attain required power, and the signal is acknowledged by the LNA to contribute minimum reflection. Allowing the maximum power theorem (MPT), the Z_{in} of the LNA must be in complex conjugate of the source impedance. For the straightforwardness, it is presumed that the source impedance is real and having the value 50Ω . The proposed methodology is perfectly coordinated to a 50Ω source by the ISD inductor L_s . The MPT, from preceding phase to LNA, is accomplished by reducing the S_{11} [13]. The mathematical equation of input impedance for the low noise amplifier is formulated [26, 28] by means of

$$Z_{in} = \frac{V_g}{I_g} \quad (1)$$

$$Z_{in} = \frac{I_g R_g + V_c + I_s \cdot j\omega L_s}{I_g} \quad (2)$$

where $V_c = \frac{I_g}{sC_{gs}} = \frac{I_g}{j\omega C_{gs}}$ and $I_s = I_g + g_m V_c$

From Eq. (2), we have

$$Z_{in} = \frac{I_g R_g + \frac{I_g}{sC_{gs}} + s L_s \left(I_g + g_m \cdot \frac{I_g}{sC_{gs}} \right)}{I_g} \quad (3)$$

$$Z_{in} = R_g + \frac{L_s g_m}{C_{gs}} + s \left(L_s + \frac{1}{s^2 C_{gs}} \right) \quad (4)$$

Putting $s = j\omega$ in Eq. (4), we have

$$Z_{in} = R_g + \frac{L_s g_m}{C_{gs}} + j\omega \left(L_s + \frac{1}{(j\omega)^2 C_{gs}} \right) \quad (5)$$

$$Z_{in} = R_g + \frac{L_s g_m}{C_{gs}} + j \left(\omega L_s - \frac{1}{\omega C_{gs}} \right) \quad (6)$$

Equation (6) can be rewritten as,

$$Z_{\text{in}} = R_g + R_a + j(X_{L_s} - X_{C_{\text{gs}}}) \quad (7)$$

$$\text{where } R_a = \frac{L_s g_m}{C_{\text{gs}}}$$

In this proposed work, MOSFETs' gate resistance (R_g) is assumed as zero. Therefore, without feedback, the Z_{in} of the MOSFET is given as

$$Z_{\text{in}} = -jX_{C_{\text{gs}}} \quad (8)$$

By the addition of series feedback that adds $R_a + jX_{L_s}$ to the obtained input impedance.

To accomplish input matching, the input impedance must be 50Ω . So,

$$Z_{\text{in}} = \frac{L_s g_m}{C_{\text{gs}}} \quad (9)$$

So to give required Z_{in} , value of L_s is taken randomly, and the values of C_{gs} and g_m are calculated using formulae.

4 Circuit Design

To design the proposed LNA, comprise three different stages. In the very first stage, basic LNA by means of a bias resistor containing a current mirror in addition to current source is shown in Fig. 1. In the simulation on ADS, ideal inductor, i.e., not value of Q has been considered.

The simulations of LNA through further cascode stage have been proposed in second step. In this, schematic ISDFS amplifier with the L_g for matching of input impedance is presented in Fig. 2

The L_s is used to offer the chosen Z_{in} of 50Ω and to accomplish synchronized input and noise matching. The cascode arrangement is of common source by means of load of common gate (CG). Further, cascode circuit has been maintained as a diode. To boost improvement reaction of LNA, an inductor sandwiched between cascode source and power supply has been delivered. The rate may be speckled, and the situation also interrupts RF leaky to the power supply.

In the last phase, i.e., recommended LNA presented in Fig. 3, to improve the S_{21} of LNA, additional C-S stage might be further added to cascode output. The output phase is DC coupled to the output of first cascode phase, so it collects the accurate bias. Enhanced S_{21} is significantly enlightening the NF at receiver end.

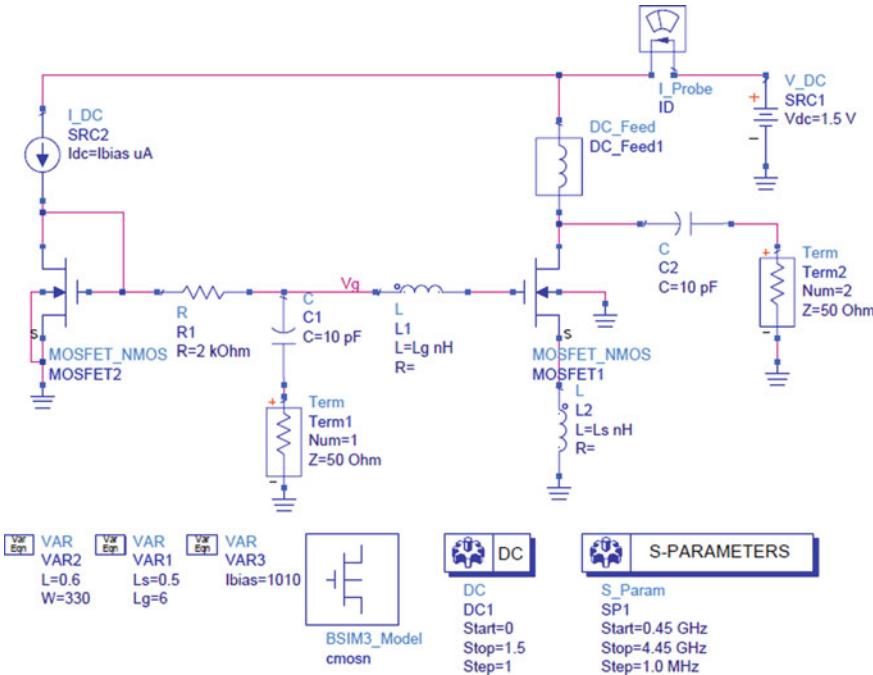


Fig. 1 Schematic circuit of basic LNA design

5 Simulation Results

The proposed LNA is simulated by means of ADS software. The LNA delivers a max S_{21} of 26.486 dB as plotted in Fig. 4. With the help of (9), C_{GS} and g_m are considered to mark input matching, the Z_{in} as close to 50Ω . Proposed design is functioned at 1.2 V steam voltage. The S_{12} (Fig. 5) is better through more than—52.533 dB at the functioning frequency. An NF_{min} of 1.044 dB is attained in proximity of selected over 2.4–2.5 GHz band of frequency for recommended LNAs as simulated in Fig. 8.

Simulation results of Fig. 1 schematic circuit of basic LNA design, Fig. 2 schematic circuit of LNA design with additional cascode stage, and Fig. 3 schematic circuit diagram of the C-S LNA by means of C-S stage further auxiliary to cascode are shown, respectively, for different parameters.

5.1 S Parameters

S_{21} is nothing but the forward transmission coefficient and denotes just exactly how healthy the signal inserts commencing input to output. It is acknowledged as forward gain as presented in Fig. 4.

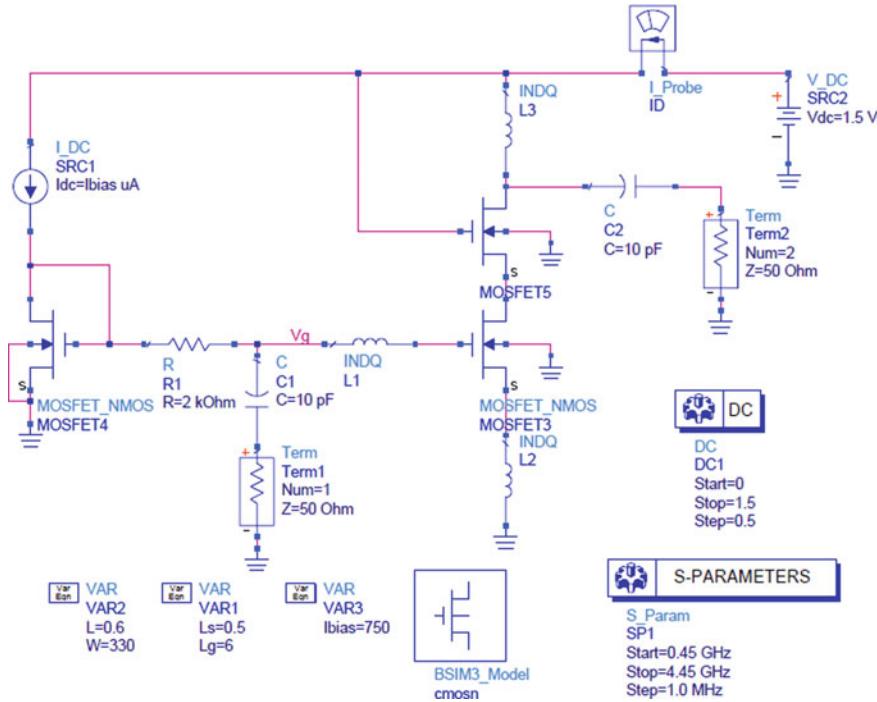


Fig. 2 Schematic circuit of LNA design with additional cascode stage

The reverse transmission coefficient is denoted by S_{12} , and it extends how greatly the input signal is returned back, and it is recognized as reverse isolation as shown in Fig. 5.

S_{11} is input reflection coefficient, which characterizes the amount, to match the input impedance with the reference impedance. It is identified as input return loss as shown in Fig. 6.

The output reflection coefficient is reflected by S_{22} , and it signifies that amount of output impedance is harmonized to load impedance. It is recognized as output return loss as revealed in Fig. 7.

5.2 Noise Figure

The NF designates the noise performances of device at 2.45 GHz, it quantified the minimum NF, and 1.044 dB has been achieved for recommended LNA as presented in Fig. 8.

The simulated ICP1 and IIP3 are shown in Fig. 9 and the stability factor is shown in Fig. 10.

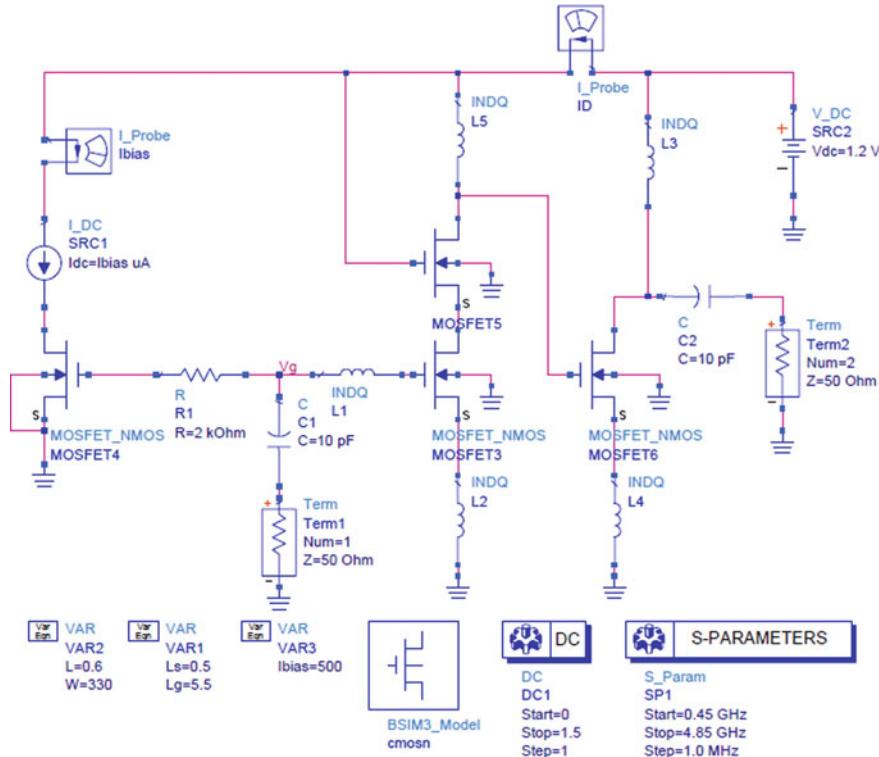


Fig. 3 Schematic circuit diagram of the C-S LNA by way of C-S stage added to cascade

5.3 Power Consumption

The PC of proposed LNA along with first and second circuit strategy is obtained by the succeeding current and voltage standards provided in Table 2.

5.4 Linearity

The linearity is another very crucial parameter that must be taken into the consideration. For any device, linear operation is precisely important, mainly when the input signal is very fragile in very close proximity of inferring signal, i.e., noisy signal. Blocking and cross-modulation of signal occurred very repeatedly in such type of scenario with the possibility of undesired inter-modulation distortion. The odd mandate misrepresentation shaped by an LNA can contribute increase to misrepresentation yields, which can limit by means of the preferred signal.

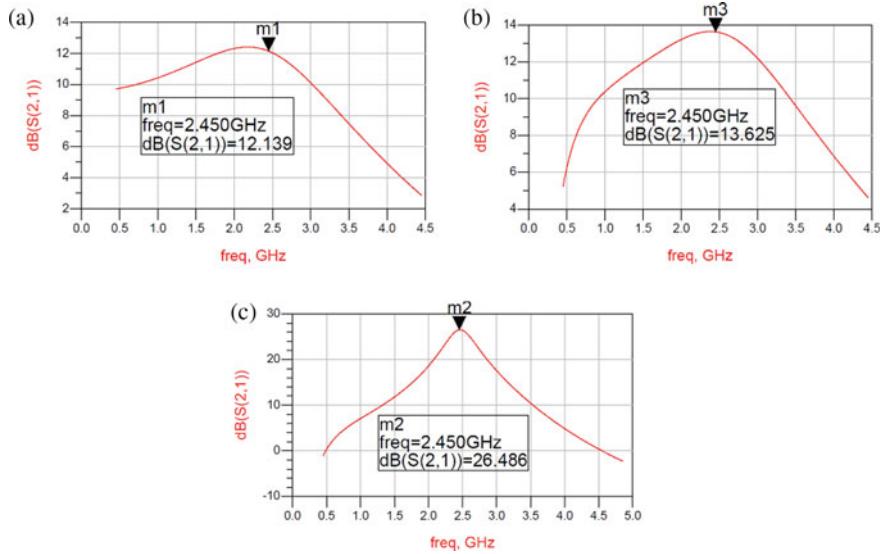


Fig. 4 Forward gain (S_{21}). Forward gain (S_{21}). Forward gain (S_{21})

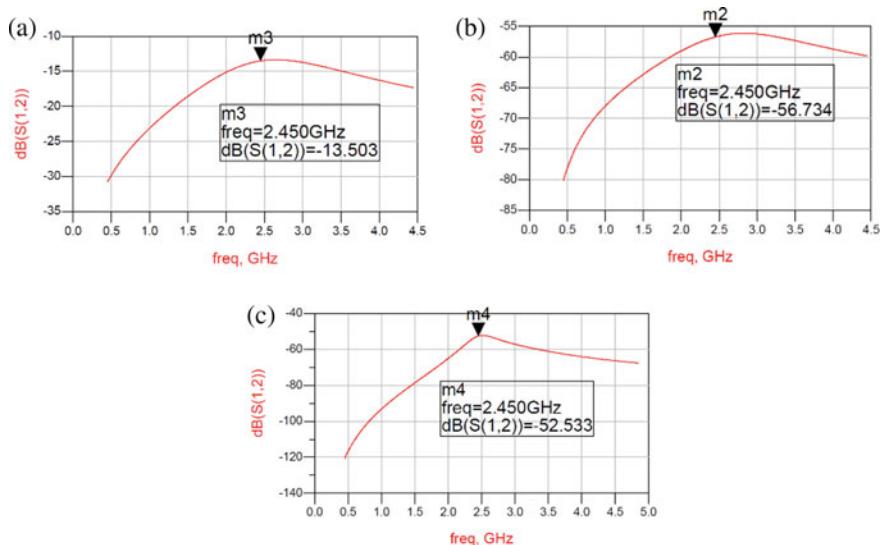


Fig. 5 Scattering parameter (S_{12}). Scattering parameter (S_{12}). Scattering parameter (S_{12})

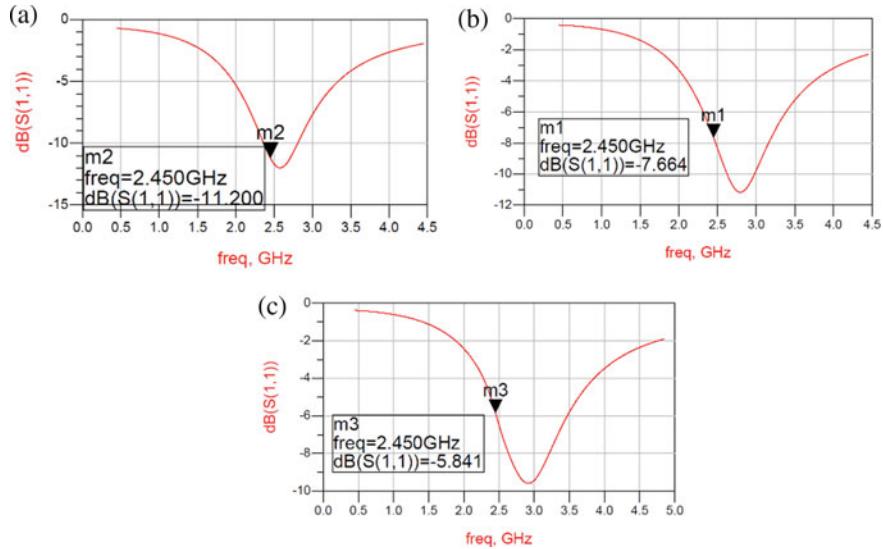


Fig. 6 Scattering parameter (S_{11}). Scattering parameter (S_{11}). Scattering parameter (S_{11})

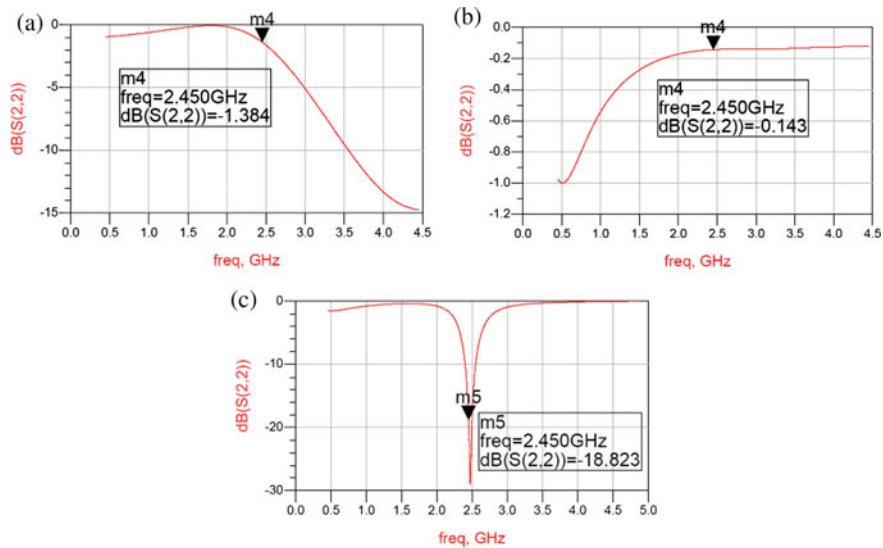


Fig. 7 Scattering parameter (S_{22}). Scattering parameter (S_{22}). Scattering parameter (S_{22})

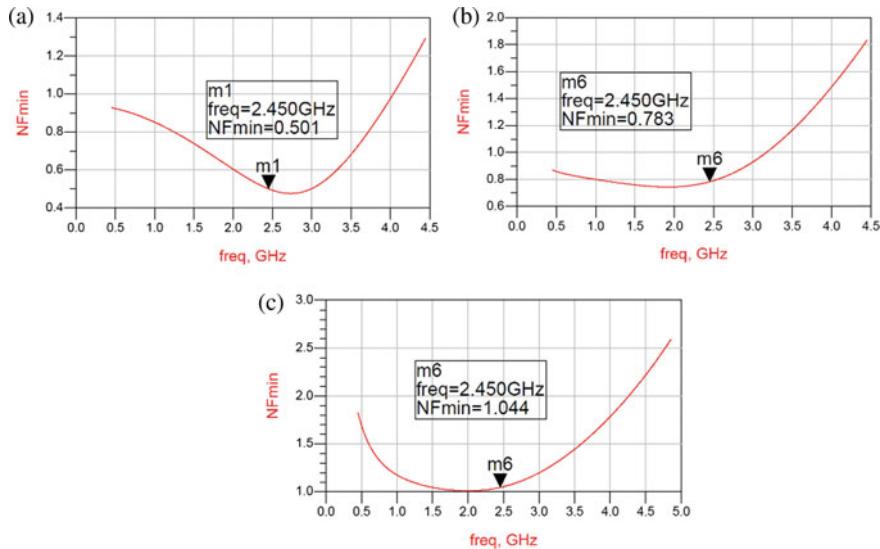


Fig. 8 NF_{min} , NF_{min} , NF_{min}

Fig. 9 Simulated ICP1 and IIP3

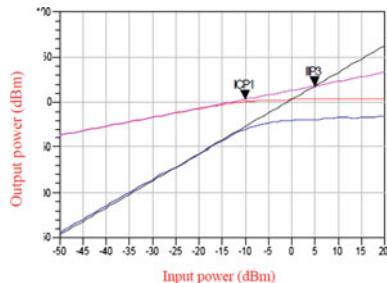
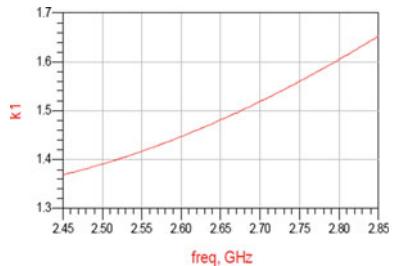


Fig. 10 Stability factor (K1)



The assessment of various constraints of LNA: Constructed on the model outcomes of LNA, Table 3 contributes the conclusion that recommended LNA for ISM, which has enhanced outcomes than the preceding outcomes.

Table 2 Standards of current and voltage

S. No.	I_D, I (mA)	V_g (mV)	Power consumed (mW)
Schematic circuit of basic LNA design (Fig. 1)			
1	11.03	1.055	11.63
Schematic circuit of LNA design with additional cascode stage (Fig. 2)			
2	7.918	1.023	8.10
Schematic circuit diagram of C-S LNA by means of C-S stage added to cascode (Fig. 3)			
3	4.865	398.7	1.93

Table 3 Comparison of various parameters of LNA

[Ref.]	Power (mW)	Gain (dB)	NF (dB)	IIP3 (dBm)	F_0 (GHz)
[6]	30	22	3.5	-9.5	1.5
[7]	12	22	2.5	-10	2.5
[8]	22.4	19.8	3	4.5	2.4
[9]	7.2	15	2.2	1.3	2.4
[10]	4.5	13.4	3	0	2.5
[11]	-	15.9	2.88	-2.6/-11.2	2.45
[12]	13.2	11	2.2	3	2.4
[13]	7.33	20.34	1.98	5	2.4-2.5
[29]	2.00	25.18	2.22	2	2.4
[30]	4.5	24.34	3.65	1.6	2.4
Proposed LNA	1.93	26.48	1.044	5	2.45

6 Conclusion

The paper designates the methodology of scheming an LNA functioning at 2.4 - 2.5 GHz with three different designing i.e. from precisely basic LNA proposal to recommended LNA design with high gain, low noise figure and low power consumption which is much applicable for wireless and ISM applications. At 1.93 mW, power consumption, 26.486 dB of forward gain and 1.044 of noise figure are achieved. Along with these prodigious characteristic, 10 dBm of IIP3, which corroborated the worthy linearity have been perceived. This associates satisfactorily to the previously existing circuit schemes in literature survey.

References

1. Comer I, DJ, Comer DT (2004) Using the weak inversion region to optimize input stage design of CMOS op amps [J]. IEEE Trans Circuits Syst 51(1):8-14

2. Mou SX, Ma JG, Seng YK (2005) A modified architecture used for input matching in CMOS low-noise amplifiers [J]. *IEEE Trans Circuits Syst* 52(11):784–788
3. Liao CH, Chuang HR (2003) A 5.7 GHz 0.18- μ m CMOS gain-controlled differential LNA with current reuse for WLNA receiver [J]. *IEEE Microwave Wireless Components Lett* 13(12):526–528
4. Floyd BA, Mehta J, Gamero C, Kenneth KO (1999) A 900-MHz, 0.8 μ m CMOS low noise amplifier with 1.2-dB noise figure. *IEEE Custom Integrated Circuit Conference*, pp 661–664
5. Konishi Y, Honjo K (1993) *Microwave semiconductor circuits*, Tokyo, Japan, Nikan, 114
6. Shaaffer DK, Lee TH (1997) A 1.5 V, 1.5 GHz CMOS low noise amplifier. *IEEE J Solid-State Circuits* 32(5):745–759
7. Rafila RA, El-Gamal MN (1999) Design of a 1.5 V CMOS integrated 3 GHz LNA. In: *Proceedings of the 1999 IEEE Int'l Symposium on Circuits and Systems* 2:440–443
8. Huang C, Weng RM, Lung HC, Lin KY (2001) A 2 V 2.4 GHz fully integrated CMOS LNA with Qenhancement circuit. *Asia-Pacific Microwave Conf* 3:1028–1031
9. Yang X, Wu T, McMacken J (2001) Design of LNA at 2.4 GHz using 0.25 μ m technology. In: *Topical meeting on silicon monolithic integrated circuits in RF systems*, pp 12–17
10. Fournier TJM, Haidar J (2001) Noise contribution in a fully integrated 1-V, 2.5-GHz LNA in CMOS-SOI technology. In: *I8th IEEE international conference on electronic circuits and systems*, pp 1611–1614
11. Li X, Brogan T, Esposito M, Myers B, KK O (2001) A comparison of CMOS and SiGe LNA's and mixers for wireless LAN application. In: *IEEE conference on custom integrated circuits*, pp 531–534
12. Lagnado I, de la Houssaye PR, Dubbelday WB, Koester SJ, Hammond R, Chu JO, Ott JA, Mooney PM, Perraud L, Jenkins KA (2000) Silicon-onsapphire for RF Si systems. In: *2000 Topical meetings on silicon monolithic integrated circuits in RF systems*, pp 79–82
13. Nadia A, Belgacem H, Aymen F (2013) A low power low noise CMOS amplifier for Bluetooth applications. In: *International conference on applied electronics (AE)*
14. Andreani P, Sjoland H (2001) Noise optimization of an inductively degenerated CMOS low noise amplifier. *IEEE Trans Circuits Syst-II: Analog Digital Signal Process* 48(9)
15. Lerdworatawee J, Namgoong W (2005) Wide-band CMOS cascode low-noise amplifier design based on source degeneration topology. *IEEE Trans Circuits Syst-I: Regular Pap* 52(11)
16. Liscidini A, Martini G, Mastantuono D, Castello R (2008) Analysis and design of configurable LNAs in feedback common. *IEEE Trans Circuits Syst-I: Express Briefs* 55(8)
17. Chi B, Zhang C, Wang Z (2008) Bandwidth extension for ultra-wideband CMOS low-noise amplifiers. In: *IEEE International symposium on circuits and systems (ISCAS)*
18. Balashov EV, Korotkov AS (2008) Ultra wideband low noise amplifier with source degeneration and shunt series feedback. In: *4th European conference on circuits and Systms for communications (ECCSC)*
19. Balashov EV, Korotkov AS (2009) Dual feedback low noise amplifier for ultra wideband application. *IEEE EUROCON*
20. Wang H, Zhang L, Yu Z (2010) A wideband inductorless LNA with local feedback and noise cancelling for low-power low-voltage applications. *IEEE Trans Circuits Syst-I: Regular Pap* 57(8)
21. Lian LL, Noh NM, Mustaffa MT, Manaf ABA, Sidek OB (2011) A dual-band LNA with 0.18- μ m CMOS switches. In: *IEEE Regional symposium on micro and nanoelectronics (RSM)*
22. Chien KH, Chiou HK (2013) A 0.6–6.2 GHz wideband LNA using resistive feedback and gate inductive peaking techniques for multiple standards application. In: *Asia-Pacific microwave conference proceedings*
23. Murad SAZ, Ismail RC, Isa MNM, Ahamd MF, Han WB (2013) High gain 2.4 GHz CMOS low noise amplifier for wireless sensor network applications. In: *IEEE international RF and microwave conference (RFM)*
24. Kurniawan TA, Wibisono G (2013) Noise modeling of source inductive degeneration low noise amplifier in 0.18- μ m CMOS technology. In: *IEEE International conference on communication, networks and satellite (COMNETSAT)*

25. Naveen Motamarri, Munshi Nurul Islam, and Apurbaranjan P. Deepu S P. Prasantakumar S: A high-gain Source Degenerative Cascode LNA for Wi-Max and W-CDMA Applications at 3.5 GHz. *IEEE International Conferences on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, (2014).
26. Verma PK, Jain P (2015) A low noise high gain low noise amplifier for wireless applications. In: IEEE International conference on communication control and intelligent systems (CCIS), GLA University, Mathura, India, pp 381–385
27. Vimalan L, Devi S (2018) Performance analysis of various topologies of common source low noise amplifier (CS-LNA) at 90 nm technology. In: 3rd IEEE international conference on recent trends in electronics, information and communication technology (RTIECT), Banglore, India, pp 1687–1692
28. Tong X, Zhang L, Zheng P, Zhang S, Xu J, Wang R (2020) An 18–56-GHz wideband GaN low-noise amplifier with 2.2–4.4-dB noise figure. *IEEE Microwave Wirel Compon Lett* 30(12):1153–1156
29. Gupta A, Kushwaha A, Mehra G (2021) A 2.4 GHz low power and high gain LNA in 0.18 μ m CMOS for radio applications. In: IEEE international conference on advanced computing & communication system (ICACCS), Coimbatore, India, pp 1390–1393
30. Ananda M, Kalpana AB (2020) Design of high gain low noise amplifier for FR receiver systems. In: 3rd international conference on intelligent sustainable system (ICISS), Thoothukudi, India, pp 1500–1504

Prostate Cancer Risk Analysis Using Artificial Neural Network



Anjali Patel, Subhankar Jana, and Juthika Mahanta

Abstract This research developed an expert system based on the neural network to analyze prostate cancer risk. This model does not diagnose prostate cancer but helps a medical practitioner avoid unnecessary biopsies. An artificial neural network is created using the data from 119 patients with four attributes of prostate cancer (PSA, % free PSA, prostate volume, and age) as input parameters, and biopsy results are used as outputs. Outputs are divided into two classes positive and negative. The 70% data is used for training the network, and 30% is used for validation and testing. The results are demonstrated by confusion matrix and ROC curve. The suggested approach yielded an accuracy of 72.2%, which is higher than other existing methods.

Keywords Neural network · Prostate cancer · Prostate-specific antigen · Prostate volume

1 Introduction

In 1956, for the first time, the term artificial intelligence (AI) was coined by John McCarthy. John McCarthy defined AI as the science and engineering of making intelligence of machines. AI is a method or a mechanism by which we build a machine that simulates human behavior. AI is a vast field; under AI, we have machine learning or deep learning, object detection, image processing, expert Systems, and so on. AI plays a vital role in health care. The expansion of AI in the medical field and the rapid development of AI techniques are substantiating disease diagnosis. AI in

A. Patel (✉) · S. Jana · J. Mahanta
NIT Silchar, Silchar, Assam 788010, India
e-mail: anjali_rs@math.nits.ac.in

S. Jana
e-mail: subhankar_rs@math.nits.ac.in

J. Mahanta
e-mail: juthika@math.nits.ac.in

medical diagnostics has allowed computer simulations to incorporate expert human thinking. One of the prominent machine learning algorithms is the neural network, which is usually employed in medical diagnostics these days.

An artificial neural network (ANN) is a functional unit of deep learning. Deep learning solves complex data-driven challenges using an ANN that mimics human brain functioning. In the human body, a neural network is made up of a huge network of associated neurons, each counting in the billions. These linked neurons in the human body facilitate parallel processing. Similarly, an ANN is mainly composed of a large number of basic processing units that are layered and linked. ANN provides a number of benefits over traditional statistical classification methods. When standard classification algorithms fail due to unclear or insufficient data, ANNs are a great choice. In multidimensional classification situations with such a coefficient of determination degree, neural networks are also effective. Disease diagnosis is an outstanding demonstration of a challenging categorization problem. This dependency can be enhanced by using ANN correctly in this domain to get the connection of symptoms and accurate diagnosis.

Several studies have been proposed based on ANN models to diagnose various diseases like Alzheimer's disease [1], schizophrenia, periodontal gum disease [2], Parkinson's disease [3], breast cancer [4, 5], skin disease [6], and heart disease [7, 8].

Most industrialized countries have a high mortality rate from prostate cancer, which is the most prevalent form of cancer among males. As reported by the American Cancer Society, there will be around 191,930 new cases and 33,330 fatalities in the United States due to prostate cancer [9]. Prostate cancer is more likely to strike older men and males of African descent. An early prostate cancer diagnosis is essential for successful treatment. It is cancer that develops in the male reproductive system's prostate gland. Prostate cancer develops when cells in the prostate change and grow uncontrollably. Those cells can move from the bladder to certain other regions of the body, particularly the bones and lymphatic system. Men above the age of fifty are more likely to develop prostate cancer. To help the doctors and medical practitioners, many researchers [10–16] developed different systems for early diagnosis of prostate cancer.

The remaining paper has been structured as follows: Sect. 2 provides a literature review; Sect. 3 details the materials and techniques used; Sect. 4 presents the research outcome, and Sect. 5 offers a summary and future work.

2 Related Work

In this section, some of the recent studies related to our topic of prostate cancer diagnosis have been discussed. Saritas et al. [12] devised an ANN model to predict prostate cancer. This system does not diagnose prostate cancer but helps doctors and other medical practitioners decide whether a biopsy is needed. It provides information about the patient's chance of having prostate cancer. Data of 121 patients with input

parameters tPSA, fPSA, and age have been taken to prepare this model. The results of the biopsy and the ANN model were compared using a confusion matrix and ROC curve. A success rate of 94.11 % was achieved, which shows that this system can be used for better prediction of prostate cancer.

Tsao et al. [14] developed an ANN model to predict the pathological stage of prostate cancer. The pathological and clinical data of 299 patients have been taken to train and test the ANN model. They evaluated ANN's prediction performance with that of LR, and the areas under the receiver operating characteristic curve (AUCs) were used to determine the validity of the 2007 Partin tables [17]. The AUCs of the ANN model were higher, and the pathologic stage of prostate cancer could be predicted with more accuracy. Prostate cancer with extracapsular extension (ECE) affected 109 of the 299 individuals, while organ-confined illness affected 190. Mesrabadi et al. [18] designed a model using different machine learning algorithms. They used data from 50 patients. Three methods have been used to diagnose prostate cancer: ANN, support vector machine(SVM), and deep learning method. Scaled conjugate gradient (SCG), Broyden–Fletcher–Goldfarb–Shanno (BFGS), and Levenberg–Marquardt (LM) are three techniques used in ANN to categorize related data. Using the same data, these three algorithms had an average accuracy of 79.1%.

Mahanta and Panda [16] designed a fuzzy expert system (FES) to predict prostate cancer risk. They used data from 119 patients with four attributes, prostate volume (PV), age, prostate-specific antigen (PSA), and percent free PSA. These four attributes were used as input of the FES, and the output was prostate cancer risk. The results were compared with the biopsy results and found 69% true prediction. Erdem et al. [19] studied machine learning algorithms for prostate cancer diagnosis and found that the MLP classifier had the maximum accuracy rate of 97% when compared to other approaches. This suggests that the MLP model is the most effective approach for diagnosing prostate cancer.

3 Materials and Methods

The ANN model was constructed using data from 119 people, including age, PSA, PV, % FPSA value, and biopsy findings. We have the biopsy results for the given data, which will be used as a target dataset. We divided biopsy results into two classes, class 1 and class 2. In class 1, we have given positive to 1 and negative to 0 values, and in class 2, we have given negative to 1 and positive to 0. To classify between positive with prostate cancer, we are using MATLAB neural network toolbox software to devise the ANN [20].

3.1 Datasets

Based on previous studies [15, 16, 21], we found that % free PSA (% FPSA), along with age, prostate-specific antigen (PSA), and PV, are significant factors for better prediction of prostate cancer. Therefore, we created a neural network model by taking care of all these parameters. We have data from 119 patients, and this data are divided into three parts data for training, validation, and testing. Validation is the process where we can examine how well our model performs against real data.

PSA Men's PSA levels are critical for PC. A protein known as PSA is produced by both healthy and cancerous prostate gland cells. Because some of this protein is secreted into the bloodstream, the usual PSA level increases. When compared to a malignant prostate, a healthy one produces less PSA in the blood. Consequently, an increase in PSA levels over the usual range may indicate prostate cancer. Prostate cancer can be detected early on with a PSA blood test. PSA levels fluctuate from one guy to the next. At any given point in time, it can change a man's life. For the most part, doctors said that PSA values less or equal to 4.0 nanograms per milliliter (ng/ml) were deemed normal. Doctors frequently advise a prostate biopsy for males with a PSA of 4 ng/ml or higher. Changes in PSA level don't mean that man has prostate cancer. There is a number of other factors that can raise or reduce the PSA level.

Percentage free PSA PSA is an enzyme that binds to proteins in semen. PSA is a protein found in the blood that is produced by the prostate gland. It assists in maintaining the liquid state of the semen. PSA that isn't attached to any proteins is called free PSA. The ratio of unbound PSA to bound (that is not tied to protein) PSA is measured by the free PSA test, whereas the PSA test examines the total PSA level in the blood (both bound and free). Percentage free PSA is measured by $\frac{\text{free PSA}}{\text{total PSA}} \times 100$. The higher the PSA level and the lower the free PSA level can increase the prostate cancer risk (PCR). As you get older, PSA levels can rise in a man without having prostate cancer. Thus, % PSA has a vital role in the early diagnosis of prostate cancer.

PV Doctors consider PV a key metric when dealing with people who have cancer or have lower urinary tract symptoms. By using univariate analysis, Young et al. [22] discovered that PV and PSA are crucial predictors in prostate cancer detection. Men's prostates grow as they become older. For prostate assessment, transrectal ultrasonography (TRUS) is often used. The prostate gland is situated in front of the rectum, beneath the bladder neck. A pathologic zonal architecture is used to characterize the gland. For example, the peripheral zone is made up of the fibromuscular stroma without glandular tissue and is separated into the periurethral and transitional zones. Approximately, 70% of all prostate tumors are found on the peripheral, with 20% arising from the transition region, while 10% occur in the middle. The prolate ellipsoid formula is given by $0.524 \times H \times W \times L \times \frac{\pi}{6}$, where H is the largest anteroposterior (height), W is transverse (width), and L is cephalocaudal (length) prostate diameters. The ellipsoid method is commonly used since it has a high correlation with the actual PV.

Age We already discussed that age is an essential factor for the early detection of PC. Elevation of PSA levels in the blood is associated with both prostate gland health and an individual's age. Chances of having prostate cancer increase after the age of 50 years.

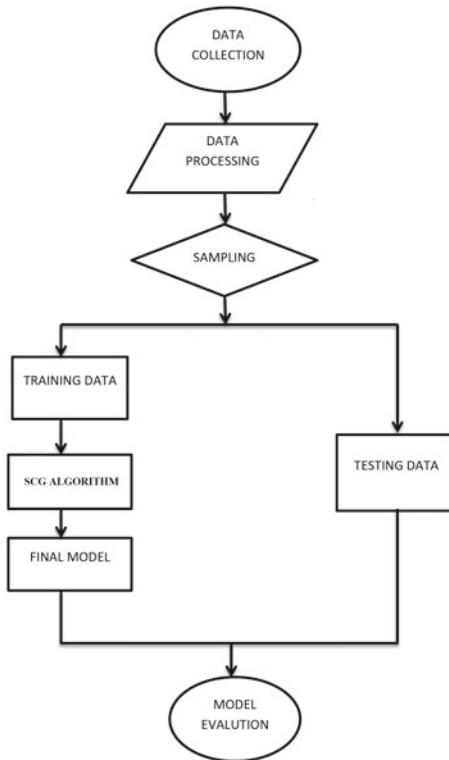
3.2 Artificial Neural Network

ANN learns the skills by gathering knowledge from the environment. It must be trained through the use of a number of examples. An ANN learns from examples provided by the data that are fed. Neurons make up the layers that make up a neural network. The network's processing power is concentrated in these neurons. Each layer accepts input and sends it on to the next, with the output layer predicting our ultimate output. In between, there are one or more hidden layers that perform lots of computations required by the network. Each layer has a fixed number of neurons connected, and each connection is associated with some weight shown by a numerical value on it.

A neural network is trained on a collection of inputs to create a set of target outputs associated with the inputs. Once the neural network has been trained, it may be utilized to create results for inputs; it hasn't previously been trained on because of the link it develops between input and output. For our model, the input data are represented as a matrix of dimension 119×4 and output data as a matrix of size 119×2 . The matrix 119×4 represent 119 samples of 4 attributes that are PSA, % free PSA, age, and PV. The matrix 119×2 represent 119 samples of 2 classes of positive and negative prostate cancer.

The architecture for our model has been presented in Fig. 1. In our study, we have constructed a feed-forward network with two levels of storage. A sigmoid transfer function is used in the hidden layer, whereas a softmax transfer function is used in the output layer. The number of hidden neurons determines the size of the hidden layer. The default layer size is 10. Since there are two classes, the number of output neurons should be adjusted at two as well. Training data, validation data, and testing data are all grouped together. We used 70% of the data for training, 15% for validation, and 15% for testing. The network is trained to predict the desired output based on inputs fed to the network in the training phase. The architecture of the ANN is presented in Fig. 2. Input nodes for this network are the diagnostic variables (PV, PSA, age, and % fPSA), multiplied by the weight values and sum together, then added some bias to it to optimize the prediction. Using a nonlinear activation function (the sigmoid function), we can get this total to fall inside a certain range. The neural network is trained on data with known outputs to update the parameter and bias weights. Information in the form of diagnosis is processed from the input to the output layer via feeding forward. Then, reverse propagation is used to train the network, called backpropagation. Error in the output is calculated by the difference of network output values obtained from feed-forward and the desired output values and sometimes squared that obtained value. When there is an error in the output,

Fig. 1 Flow chart of the proposed model



it is propagated back through the network, and the weights are changed in order to reduce the error as much as possible. Backpropagation is the process of updating the weights of the network in order to reduce the error in prediction. In this system, the network is trained by using the scaled conjugate gradient (SCG) [23] algorithm. The MATLAB function ‘trainscg’ is used to train the network by updating weight and bias values using SCG algorithm. It can train any network that has derivative functions for its weight, net input, and transfer functions. The step size in the SCG method is determined by a quadratic approximation of the error function, making it more resilient and independent of user-defined parameters. After that, the output is compared to the original results in every iteration, and multiple iterations are done to get the maximum accuracy. With every iteration, the weights at every interconnection are adjusted based on the error. After various iterations of backpropagation, weights are assigned the appropriate value. At this point, our network is trained and can be used to make predictions. Then, this trained model is applied to test data.

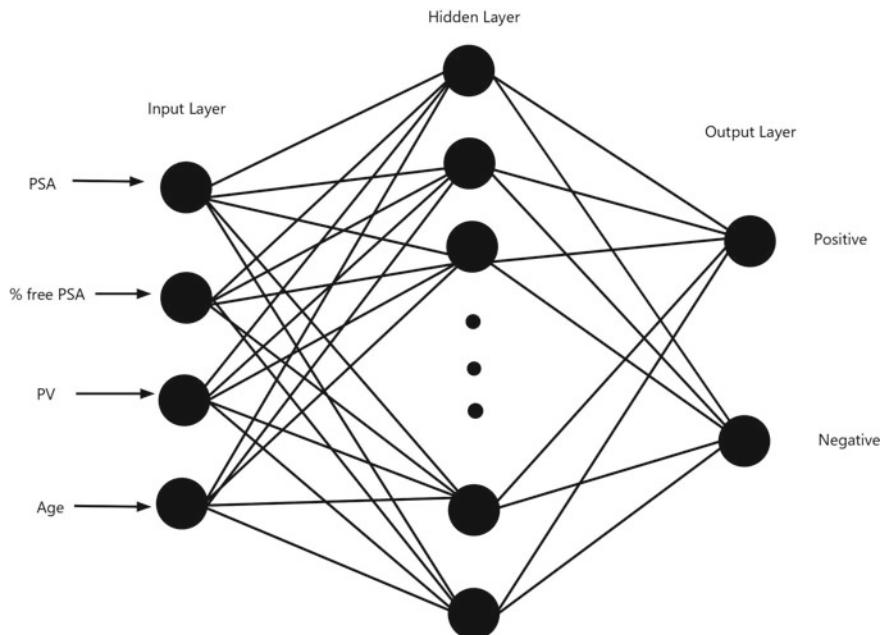
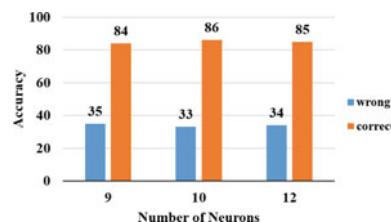


Fig. 2 The architecture of the ANN

Fig. 3 Prediction accuracy for 9, 10, and 12 neurons in the hidden layer



4 Results and Discussion

We have trained the network using a different number of neurons in the hidden layer and then fixed the neurons on which network gives the best result. We set the number of neurons in the hidden layer by trial and error. The following Fig. 3 shows the results on different hidden neurons. At first, we trained the network using nine neurons in the hidden layer, and it was diagnosed correctly 84 out of 119, by 10 neurons network correctly diagnosed 86 samples, and by using 12 neurons, we have 85 out of 119 correctly diagnosed as demonstrated in Fig. 3.

Out of all the patients, 61 have positive biopsy report, and 58 have negative results. Our proposed model classifies that 39 (63.93%) people have higher chances of being positive for prostate cancer, and 47 (81.03%) people have no chance of being positive for prostate cancer. The confusion matrix in Table 1 illustrates the results by

Table 1 The confusion matrix represents the results of classification using 10 neurons

Actual class	Predicted class	
	Positive	Negative
Positive	39	11
Negative	22	47

Table 2 Comparison with other methods using the same dataset

Methods	Sarita et al. [21]	Mahanta and Panda [16]	The proposed method
Correct prediction	77	82	86
Incorrect prediction	42	37	33
Accuracy (%)	64.71	68.91	72.2

the proposed method. The following expression is used to measure the accuracy

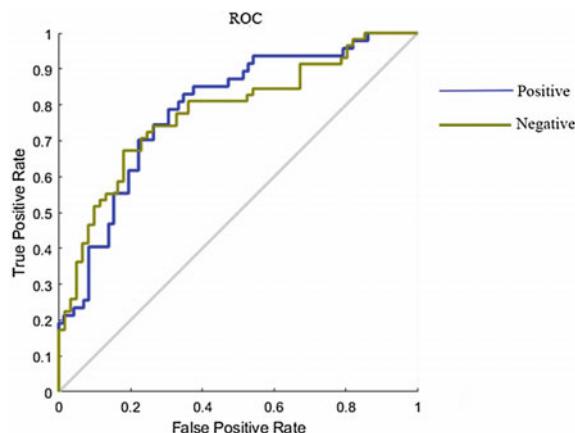
$$\text{Accuracy (A)} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}} \quad (1)$$

where tp is true positive, tn is true negative, fp is false positive, and fn is false negative. By using (1), we have found accuracy for our problem is 72.2%, which is better than the fuzzy expert system prepared by Sarita et al. [21] and Mahanta and Panda [16] by using the same data. They developed a fuzzy system using the same set of data and compared the results with the biopsy results. It is clear from Table 2 that they found average 68.91% accurate prediction for their FES, whereas our proposed ANN model has 72.2% accuracy.

We can see our system performance by the receiver operating curve (ROC) presented in Fig. 4. The receiver operating characteristic plot is another indicator of how effectively the neural network has suited the data. This graph depicts how the true-positive rate (TPR) and false-positive rate (FPR) change when the output thresholding is changed from 0 to 1. The lines should appear in the upper-left corner of the screen for the perfect result. The network does a good job with our problem.

ANN can handle complex data problems with high efficiency, unlike other statistical methods, which follow a set of rules and process information step by step. With high predictive accuracy, a properly trained neural network is superior to other predictive techniques. Neural networks do not need any series of rules and expertise to design the rule like other rule-based systems. We can train the neural network to perform different types of tasks. In FES, the inclusion of new variable results increases the number of rules and computation time, and problems become more complex. But, ANN can handle a large amount of data with high efficiency. The inclusion of a new parameter doesn't affect the calculation time or system efficiency. Thus, we can say that our ANN model helps reduce unnecessary biopsies and save medical costs.

Fig. 4 ROC for positive and negative classes of prostate cancer



5 Conclusion and Future Work

We have developed an artificial neural network-based expert system to predict prostate cancer by using some set of inputs. For this expert system, we used four parameters, PSA, % fPSA, age, and PV. We have explained how these parameters are significant for the early diagnosis of prostate cancer. We have trained our network using 70% of data, and 30% of data is used to test the network. Our proposed model classified the prostate cancer as positive and negative with an accuracy of 72.2%, which is higher than the other models [16, 21]. The goal of the suggested model is to help clinicians decide whether or not to do a biopsy on a patient who may or may not have prostate cancer. Therefore, ANNs can be used to help doctors and other medical practitioners to diagnose prostate cancer. Our study's limitations include relatively small sample sizes, and the accuracy of our model could have been increased by incorporating new biomarkers and taking some other variables like a prostate history of the family and lifestyle. One can develop a new model by taking care of some new variables and biomarker data and increasing the sample size.

Conflict of Interests The authors declares that there is no conflict of interests regarding the publication of this manuscript.

References

1. Aljović A, Badnjević A, Gurbeta L (2016) Artificial neural networks in the discrimination of Alzheimer's disease using biomarkers data. In: 2016 5th Mediterranean conference on embedded computing (MECO). IEEE, pp 286–289
2. Lee J-H, Kim D-H, Jeong S-N, Choi S-H (2018) Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 77:106–111

3. Pereira LAM, Rodrigues D, Ribeiro PB, Papa JP, Weber SAT (2014) Social-spider optimization-based artificial neural networks training and its applications for Parkinson's disease identification. In: 2014 IEEE 27th international symposium on computer-based medical systems. IEEE, pp 14–17
4. Gupta KK, Vijay R, Pahadiya P (2020) A review paper on feature selection techniques and artificial neural networks architectures used in thermography for early stage detection of breast cancer. *Soft Comput Theor Appl* 455–465
5. Hakkoum H, Idri A, Abnane I (2021) Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification. In: Computer methods in biomechanics and biomedical engineering: imaging and visualization, pp 1–13
6. Kalbande DR, Khopkar U, Sharma A, Daftary N, Kokate Y, Dmello R (2020) Early stage detection of psoriasis using artificial intelligence and image processing. In: *Soft computing: theories and applications*. Springer, pp 1199–1208
7. Ajam N (2015) Heart diseases diagnoses using artificial neural network. *IISTE Netw Complex Syst* 5(4)
8. Chung C-C, Chiu W-T, Huang Y-H, Chan L, Hong C-T, Chiu H-W (2021) Identifying prognostic factors and developing accurate outcome predictions for in-hospital cardiac arrest by using artificial neural networks. *J Neurol Sci* 425:117445
9. Prashant M, Krishnan S, Meesha C, Priyanka D, Lakshminarayana SK, Stephen S, Vinodh N, Anish J, Sandeep N, Selvaraj RF et al (2020) Cancer statistics, 2020: report from national cancer registry programme, India. *JCO Glob Oncol* 6:1063–1075
10. Benecchi L (2006) Neuro-fuzzy system for prostate cancer diagnosis. *Urology* 68(2):357–361
11. Barlow H, Mao S, Khushi M (2019) Predicting high-risk prostate cancer using machine learning methods. *Data* 4(3):129
12. Saritas I, Ozkan IA, Sert IU (2010) Prognosis of prostate cancer by artificial neural networks. *Exp Syst Appl* 37(9):6646–6650
13. Srivenkatesh M (2020) Prediction of prostate cancer using machine learning algorithms. *Int J Recent Technol Eng (IJRTE)* 8:5353–5362
14. Tsao C-W, Liu C-Y, Cha T-L, Sheng-Tang W, Sun G-H, Dah-Shyong Yu, Chen H-I, Chang S-Y, Chen S-C, Hsu C-Y (2014) Artificial neural network for predicting pathological stage of clinically localized prostate cancer in a Taiwanese population. *J Chin Med Assoc* 77(10):513–518
15. Seker H, Odetayo MO, Petrovic D, Naguib RNG (2003) A fuzzy logic based-method for prognostic decision making in breast and prostate cancers. *IEEE Trans Inf Technol Biomed* 7(2):114–122
16. Mahanta J, Panda S (2020) Fuzzy expert system for prediction of prostate cancer. *New Math Nat Comput* 16(01):163–176
17. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD (2001) Contemporary update of prostate cancer staging nomograms (Partin tables) for the new millennium. *Urology* 58(6):843–848
18. Mersrabadi HA, Faez K (2018) Improving early prostate cancer diagnosis by using artificial neural networks and deep learning. In: 2018 4th Iranian conference on signal processing and intelligent systems (ICSPIS). IEEE, pp 39–42
19. Erdem E, Bozkurt F (2021) A comparison of various supervised machine learning techniques for prostate cancer prediction. *Avrupa Bilim Teknol Derg* 21:610–620
20. Demuth H, Beale M, Hagan M (1994) Neural network toolbox. Mathworks
21. Saritas I, Allahverdi N, Sert IU (2013) A fuzzy approach for determination of prostate cancer. *Int J Intell Syst Appl Eng* 1(1):1–7
22. Kim YM, Park S, Kim J, Park S, Lee JH, Ryu DS, Choi SH, Cheon SH (2013) Role of prostate volume in the early detection of prostate cancer in a cohort with slowly increasing prostate specific antigen. *Yonsei Med J* 54(5):1202–1206
23. Møller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw* 6(4):525–533

On Two Bivariate Kinds of (p, q) -Euler Polynomials



Atul K. Singh, Idrees A. Khan, Nidal Abu-Libdeh, and Waseem A. Khan

Abstract In this paper, we introduce (p, q) -Sine Euler polynomials and (p, q) -Cosine Euler polynomials and investigate some properties of these polynomials. We derive some recurrence relations and relationships for those polynomials. Furthermore, we derive the (p, q) -derivative operator and (p, q) -representations for the (p, q) -Sine Euler polynomials and (p, q) -Cosine Euler polynomials.

Keywords (p, q) -calculus, Trigonometric functions, (p, q) -Euler polynomials

Mathematics Subject Classification 33D15 · 11B68 · 11B73 · 11B83

1 Introduction

The (p, q) -numbers $0 < |q| < |p| \leq 1$ are defined as

$$[\omega]_{p,q} = p^{\omega-1} + p^{\omega-2}q + p^{\omega-3}q^2 + \cdots + pq^{\omega-2} + q^{\omega-1} = \frac{p^\omega - q^\omega}{p - q},$$

A. K. Singh (✉) · I. A. Khan

Department of Mathematics and Statistics, Faculty of Science, Integral University, Lucknow 226026, India

e-mail: satul545@gmail.com

N. Abu-Libdeh · W. A. Khan

Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, P.O. Box 1664, Al Khobar 31952, Saudi Arabia

e-mail: nabulibdeh@pmu.edu.sa

W. A. Khan

e-mail: wkhan1@pmu.edu.sa

and

$$[\omega]_{p,q} = p^{\omega-1} [\omega]_{q/p},$$

where $[\omega]_{q/p}$ are called the q -calculus defined by

$$[\omega]_{q/p} = \frac{(q/p)^{\omega-1}}{(q/p) - 1}.$$

Replacing q by q/p , we cannot find (p, q) -numbers. The (p, q) -numbers and q -numbers are different, when $p = 1$, (p, q) -numbers reduces to q -numbers (see []).

The (p, q) -number of derivative operator \hbar with respect to ξ is defined by

$$D_{p,q} \hbar(\xi) = \frac{\hbar(p\xi) - \hbar(q\xi)}{(p - q)\xi}, (\xi \neq 0) \quad (1)$$

and $(D_{p,q} \hbar(0)) = \hbar'(0)$, delivered that \hbar is differentiable at 0. The number (p, q) -derivative operator holds the following properties:

$$D_{p,q}(\hbar(\xi)g(\xi)) = g(p\xi)D_{p,q}\hbar(\xi) + \hbar(q\xi)D_{p,q}g(\xi), \quad (2)$$

and

$$D_{p,q} \left(\frac{\hbar(\xi)}{g(\xi)} \right) = \frac{g(q\xi)D_{p,q}\hbar(\xi) - \hbar(q\xi)D_{p,q}g(\xi)}{g(p\xi)g(q\xi)}. \quad (3)$$

Let $\omega \geq 1$, the (p, q) -generalization of addition $(\xi + a)^\omega$ is given by

$$(\xi + a)_{p,q}^\omega = (\xi + a)(p\xi + aq) \cdots (p^{\omega-2}\xi + aq^{\omega-2})(p^{\omega-1}\xi + aq^{\omega-1}),$$

$$= \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p,q} p^{\binom{\omega}{2}} q^{\binom{\omega-\delta}{2}} \xi^\delta a^{\omega-\delta}, \quad (4)$$

where

$$\binom{\omega}{\delta}_{p,q} = \frac{[\omega]_{p,q}!}{[\omega-\delta]_{p,q}! [\delta]_{p,q}!} (\omega \geq \delta) \text{ and } [\omega]_{p,q}! = [\omega]_{p,q} \cdots [2]_{p,q} [1]_{p,q}, (\omega \in \mathbb{N}).$$

The (p, q) -extension of exponential function is defined by

$$e_{p,q}(\xi) = \sum_{\omega=0}^{\infty} \frac{p^{\binom{\omega}{2}} \xi^\omega}{[\omega]_{p,q}!} \quad \text{and} \quad E_{p,q}(\xi) = \sum_{\omega=0}^{\infty} \frac{q^{\binom{\omega}{2}} \xi^\omega}{[\omega]_{p,q}!}, \quad (5)$$

holds the identities

$$e_{p,q}(\xi)E_{p,q}(-\xi) = 1 \quad \text{and} \quad e_{p-q}(-\xi) = E_{p,q}(\xi), \quad (6)$$

and

$$D_{p,q} e_{p,q}(\xi) = e_{p,q}(p\xi) \quad \text{and} \quad D_{p,q} E_{p,q}(\xi) = E_{p,q}(q\xi). \quad (7)$$

The (p, q) -analogue integral is defined by

$$\int_0^a f(\xi) d_{p,q} \xi = (p - q)a \sum_{\delta=0}^{\infty} \frac{p^\delta}{q^{\delta+1}} f\left(a \frac{p^\delta}{q^{\delta+1}}\right),$$

in conjunction with

$$\int_a^b \hbar(\xi) d_{p,q} \xi = \int_0^b \hbar(\xi) d_{p,q} \xi - \int_0^a \hbar(\xi) d_{p,q} \xi, \quad (\text{see [10]}). \quad (8)$$

From (1.5), we can derive

$$e_{p,q}(i\xi) = \sum_{\omega=0}^{\infty} \frac{p^{\binom{\omega}{2}} (i\xi^\omega)}{[\omega]_{p,q}!} = \sum_{\omega=0}^{\infty} \frac{(-1)^\omega p^{\binom{\omega}{2}} \xi^{2\omega}}{[\omega]_{p,q}!} + i \sum_{\omega=0}^{\infty} \frac{(-1)^\omega p^{\binom{2\omega+1}{2}} \xi^{2\omega+1}}{[2\omega+1]_{p,q}!}. \quad (9)$$

By (1.9), the (p, q) -extension Sine and (p, q) -extension Cosine functions are defined (see [3, 12]) as follows:

$$Cos_{p,q}(\xi) = \sum_{\omega=0}^{\infty} \frac{(-1)^\omega p^{\binom{\omega}{2}} \xi^{2\omega}}{[2\omega]_{p,q}!}, \quad (10)$$

$$Sin_{p,q}(\xi) = \sum_{\omega=0}^{\infty} \frac{(-1)^\omega p^{\binom{2\omega+1}{2}} \xi^{2\omega+1}}{[2\omega+1]_{p,q}!}. \quad (11)$$

The (p, q) -extension Bernoulli, the (p, q) -extension Euler and the (p, q) -extension Genocchi polynomials are defined (see [2–4, 6, 7])

$$\left(\frac{v}{e_{p,q}(v) - 1} \right)^\alpha e_{p,q}(\xi v) E_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} \mathbb{B}_\omega^{(\alpha)}(\xi, \eta : p, q) \frac{v^\omega}{[\omega]_{p,q}!} \quad |v| < 2\pi, \quad (12)$$

$$\left(\frac{2}{e_{p,q}(v) + 1} \right)^\alpha e_{p,q}(\xi v) E_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} \mathbb{E}_\omega^{(\alpha)}(\xi, \eta : p, q) \frac{v^\omega}{[\omega]_{p,q}!} \quad |v| < \pi, \quad (13)$$

and

$$\left(\frac{2v}{e_{p,q}(v) + 1} \right)^\alpha e_{p,q}(\xi v) E_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} \mathbb{G}_\omega^{(\alpha)}(\xi, \eta : p, q) \frac{v^\omega}{[\omega]_{p,q}!} \quad |v| < \pi. \quad (14)$$

It is clear that

$$\mathbb{B}_\omega^{(\alpha)}(0, 0 : p, q) = \mathbb{B}_\omega^{(\alpha)}(p, q), \quad \mathbb{E}_\omega^{(\alpha)}(0, 0 : p, q) = \mathbb{E}_\omega^{(\alpha)}(p, q),$$

and

$$\mathbb{G}_\omega^{(\alpha)}(0, 0 : p, q) = \mathbb{G}_\omega^{(\alpha)}(p, q) \quad (\omega \in \mathbb{N}).$$

2 A Bivariate Kind of (p, q) -Euler Polynomials

Let $\xi, \eta \in \mathbb{R}$. It is well-known that the Taylor expansion of the two functions $e^{\xi v} \cos(\eta v)$ and $e^{\xi v} \sin(\eta v)$ are as follows (see [3]):

$$e^{\xi v} \cos(\eta v) = \sum_{\omega=0}^{\infty} C_\omega(\xi, \eta) \frac{v^\omega}{\omega!}, \quad (15)$$

and

$$e^{\xi v} \sin(\eta v) = \sum_{\omega=0}^{\infty} S_\omega(\xi, \eta) \frac{v^\omega}{\omega!}, \quad (16)$$

where

$$C_\omega(\xi, \eta) = \sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^\delta \binom{\omega}{2\delta} \xi^{\omega-2\delta} \eta^{2\delta}, \quad (17)$$

and

$$S_\omega(\xi, \eta) = \sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^\delta \binom{\omega}{2\delta+1} \xi^{\omega-2\delta-1} \eta^{2\delta+1}. \quad (18)$$

Recently, Sadjang and Duran [12] introduced a (p, q) -extension of the two above polynomials $C_{\omega,p,q}(\xi, \eta)$ and $S_{\omega,p,q}(\xi, \eta)$ as

$$e_{p,q}(\xi v) \cos_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} C_{\omega,p,q}(\xi, \eta) \frac{v^\omega}{[\omega]_{p,q}!}, \quad (19)$$

and

$$e_{p,q}(\xi v) \operatorname{Sin}_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} S_{\omega,p,q}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p,q}!}, \quad (20)$$

where

$$C_{\omega,p,q}(\xi, \eta) = p^{\binom{\omega}{2}} \sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^{\delta} \binom{\omega}{2\delta} {}_{p,q} p^{2\delta(\delta-\omega)} \xi^{\omega-2\delta} \eta^{2\delta}, \quad (21)$$

and

$$S_{\omega,p,q}(\xi, \eta) = p^{\binom{\omega-1}{2}} \sum_{\delta=0}^{\lfloor \frac{\omega-1}{2} \rfloor} (-1)^{\delta} \binom{\omega}{2\delta+1} {}_{p,q} p^{4\delta^2-2\delta\omega} \xi^{\omega-2\delta-1} \eta^{2\delta+1}. \quad (22)$$

Let $\xi, \eta \in \mathbb{R}$, we consider two bivariate kinds of (p, q) -Cosine Euler polynomials $\mathbb{E}_{\delta,p,q}^{(c)}(\xi, \eta)$ and (p, q) -Sine Euler polynomials $\mathbb{E}_{\omega,p,q}^{(s)}(\xi, \eta)$ as

$$\frac{2}{e_{p,q}(v) + 1} e_{p,q}(\xi v) \operatorname{Cos}_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega,p,q}^{(c)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p,q}!}, \quad (23)$$

and

$$\frac{2}{e_{p,q}(v) + 1} e_{p,q}(\xi v) \operatorname{Sin}_{p,q}(\eta v) = \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega,p,q}^{(s)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p,q}!}, \quad (24)$$

respectively.

Note that

$$\mathbb{E}_{\omega,p,q}^{(c)}(0, 0) = \mathbb{E}_{\omega,p,q}, \quad \mathbb{E}_{\omega,p,q}^{(s)}(\xi, \eta) = 0.$$

Theorem 2.1 Let $\omega \geq 0$. Then

$$\mathbb{E}_{\omega,p,q}^{(c)}(\xi, \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta} {}_{p,q} \mathbb{E}_{\delta,p,q} C_{\omega-\delta, p,q}(\xi, \eta), \quad (25)$$

$$\mathbb{E}_{\omega,p,q}^{(s)}(\xi, \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta} {}_{p,q} \mathbb{E}_{\delta,p,q} S_{\omega-\delta, p,q}(\xi, \eta). \quad (26)$$

Proof By (1.13), (2.5) and (2.9), we have

$$\begin{aligned}
& \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{\nu^{\omega}}{[\omega]_{p, q}!} = \frac{2}{e_{p, q}(\nu) + 1} e_{p, q}(\xi \nu) \cos_{p, q}(\eta \nu) \\
& = \left(\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)} \frac{\nu^{\omega}}{[\omega]_{p, q}!} \right) \left(\sum_{\omega=0}^{\infty} C_{\omega, p, q}(\xi, \eta) \frac{\nu^{\omega}}{[\omega]_{p, q}!} \right) \\
& = \sum_{\omega=0}^{\infty} \left(\sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} \mathbb{E}_{\delta, p, q}^{(c)} C_{\omega-\delta, p, q}(\xi, \eta) \right) \frac{\nu^{\omega}}{[\omega]_{p, q}!},
\end{aligned}$$

yields the proof (2.11). The proof of (2.12) is similar. \square

Theorem 2.2 For $\omega \geq 0$, we have

$$\mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) = \sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^{\delta} \binom{\omega}{2\delta}_{p, q} p^{\binom{2\delta}{2}} \mathbb{E}_{\omega-2\delta, p, q}(\xi) \eta^{2\delta}, \quad (27)$$

$$\mathbb{E}_{\omega, p, q}^{(s)}(\xi, \eta) = \sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^{\delta} \binom{\omega}{2\delta+1}_{p, q} p^{\binom{2\delta+1}{2}} \mathbb{E}_{\omega-1-2\delta, p, q}(\xi) \eta^{2\delta+1}. \quad (28)$$

Proof In (1.10) and (2.9), we see

$$\begin{aligned}
& \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{\nu^{\omega}}{[\omega]_{p, q}!} = \frac{2}{e_{p, q}(\nu) + 1} e_{p, q}(\xi \nu) \cos_{p, q}(\eta \nu) \\
& = \left(\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{\omega^{\omega}}{[\omega]_{p, q}!} \right) \left(\sum_{\omega=0}^{\infty} \frac{(-1)^{\omega} p^{\binom{2\omega}{2}}}{[2\omega]_{p, q}!} (\eta \nu)^{2\omega} \right) \\
& = \sum_{\omega=0}^{\infty} \left(\sum_{\delta=0}^{\lfloor \frac{\omega}{2} \rfloor} (-1)^{\delta} \binom{\omega}{2\delta}_{p, q} p^{\binom{2\delta}{2}} \mathbb{E}_{\omega-2\delta, p, q}(\xi) \eta^{2\delta} \right) \frac{\nu^{\omega}}{[\omega]_{p, q}!}.
\end{aligned}$$

Comparing the coefficients of $\frac{\nu^{\omega}}{[\omega]_{p, q}!}$ in both sides we get (2.13). The proof of (2.14) is similar via (1.11). \square

Theorem 2.3 For $\omega \geq 0$, we have

$$C_{\omega, p, q}(\xi, \eta) = \frac{1}{2} \sum_{\delta=0}^{\omega} p^{\binom{\delta}{2}} \binom{\omega}{\delta}_{p, q} E_{\omega-\delta, p, q}^{(c)}(\xi, \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta), \quad (29)$$

$$S_{\omega, p, q}(\xi, \eta) = \frac{1}{2} \sum_{\delta=0}^{\omega} p^{\binom{\delta}{2}} \binom{\omega}{\delta}_{p, q} E_{\omega-\delta, p, q}^{(s)}(\xi, \eta) + \mathbb{E}_{\omega, p, q}^{(s)}(\xi, \eta). \quad (30)$$

Proof In (2.9), we have

$$\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} = \frac{2}{e_{p, q}(v) + 1} e_{p, q}(\xi v) \cos_{p, q}(\eta v).$$

Hence

$$\begin{aligned} \sum_{\omega=0}^{\infty} C_{\omega, p, q}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} &= \frac{e_{p, q}(v) + 1}{2} \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} \\ &= \frac{1}{2} \left(\sum_{\omega=0}^{\infty} \frac{p^{(\delta)}}{[\delta]_{p, q}} \frac{v^{\delta}}{[\delta]_{p, q}!} + 1 \right) \left(\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} \right) \\ &= \sum_{\omega=0}^{\infty} \left(\frac{1}{2} \sum_{\delta=0}^{\omega} p^{(\delta)} \binom{\omega}{\delta}_{p, q} E_{\omega-\delta, p, q}^{(c)}(\xi, \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \right) \frac{v^{\omega}}{[\omega]_{p, q}!}. \end{aligned}$$

Complete the proof of (2.15). The proof of (2.16) is similar. \square

Theorem 2.4 For $\omega \geq 0$, we have

$$\mathbb{E}_{\omega, p, q}^{(c)}((1 \oplus_{p, q} \xi), \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) = [2]_{p, q} C_{\omega, p, q}(\xi, \eta), \quad (31)$$

$$\mathbb{E}_{\omega, p, q}^{(s)}((1 \oplus_{p, q} \xi), \eta) + \mathbb{E}_{\omega, p, q}^{(s)}(\xi, \eta) = [2]_{p, q} S_{\omega, p, q}(\xi, \eta). \quad (32)$$

Proof From (2.9), we have

$$\begin{aligned} \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}((1 \oplus_{p, q} \xi), \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} &= \frac{2e_{p, q}[(1 \oplus \xi)v] \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \\ &= \frac{2e_{p, q}(\xi v)[e_{p, q}(v) - 1 + 1] \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \\ &= 2e_{p, q}(\xi v) \cos_{p, q}(\eta v) - \frac{2e_{p, q}(\xi v) \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \\ &= 2 \sum_{\omega=0}^{\infty} C_{\omega, p, q}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} - \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \frac{v^{\omega}}{[\omega]_{p, q}!}, \end{aligned}$$

yields the proof of (2.17). The proof of (2.18) is similar. \square

Corollary 2.1 For $\omega \geq 0$, we have

$$\mathbb{E}_{\omega, p, q}^{(c)}(1, \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(0, \eta) = [2]_{p, q}(-1)^\omega p^{\binom{2\omega}{2}} \eta^{2\omega},$$

and

$$\mathbb{E}_{\omega, p, q}^{(s)}(1, \eta) + \mathbb{E}_{\omega, p, q}^{(s)}(0, \eta) = [2]_{p, q}(-1)^\omega p^{\binom{2\omega+1}{2}} \eta^{2\omega+1}.$$

Proof If we replace ω by 2ω in (2.17) and ξ by 0, we obtain

$$\mathbb{E}_{\omega, p, q}^{(c)}(1, \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(0, \eta) = [2]_{p, q} C_{2\omega, p, q}(0, \eta).$$

The first relation is proved since from (2.7), we have $C_{2\omega, p, q} = (-1)^\omega p^{\binom{2\omega}{2}} y^{2\omega}$. The second relation is proved similarly. \square

Theorem 2.5 For $\omega \geq 0$, we have

$$\mathbb{E}_{\omega, p, q}^{(c)}((\xi \oplus_{p, q} \zeta), \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} \mathbb{E}_{\delta, p, q}^{(c)}(\xi) C_{\omega-\delta, p, q}(\eta, \zeta), \quad (33)$$

$$\mathbb{E}_{\omega, p, q}^{(s)}((\xi \oplus_{p, q} \zeta), \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} \mathbb{E}_{\delta, p, q}^{(s)}(\xi) S_{\omega-\delta, p, q}(\eta, \zeta). \quad (34)$$

Proof By (2.9), we see

$$\begin{aligned} \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}((\xi \oplus_{p, q} \zeta), \eta) \frac{v^\omega}{[\omega]_{p, q}!} &= \frac{2e_{p, q}[(\xi \oplus \zeta)v] \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \\ &= \frac{2e_{p, q}(\xi v)}{e_{p, q}(v) + 1} \times e_{p, q}(\zeta v) \cos_{p, q}(\eta v) \\ &= \left(\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi) \frac{v^\omega}{[\omega]_{p, q}!} \right) \left(\sum_{\omega=0}^{\infty} C_{\omega, p, q}(\eta, \zeta) \frac{v^\omega}{[\omega]_{p, q}!} \right) \\ &= \sum_{\omega=0}^{\infty} \left(\sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} \mathbb{E}_{\delta, p, q}^{(c)}(\xi) C_{\omega-\delta, p, q}(\eta, \zeta) \right) \frac{v^\omega}{[\omega]_{p, q}!}. \end{aligned}$$

Complete proof of (2.19). The proof of (2.20) is similar. \square

Theorem 2.6 For $\omega \geq 0$, we have

$$\mathbb{E}_{\omega, p, q}^{(c)}((\xi \oplus_{p, q} \zeta), \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(c)}(\xi) \zeta^{\omega-\delta}, \quad (35)$$

$$\mathbb{E}_{\omega, p, q}^{(s)}((\xi \oplus_{p, q} \zeta), \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(s)}(\xi) \zeta^{\omega-\delta}. \quad (36)$$

Proof Using (2.9), we have

$$\begin{aligned} \sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}((\xi \oplus_{p, q} \zeta), \eta) \frac{v^{\omega}}{[\omega]_{p, q}!} &= \frac{2e_{p, q}[(\xi \oplus \zeta)v] \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \\ &= \frac{2e_{p, q}(\xi v) \cos_{p, q}(\eta v)}{e_{p, q}(v) + 1} \times e_{p, q}(\zeta v) \\ &= \left(\sum_{\omega=0}^{\infty} \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) \right) \left(\sum_{\omega=0}^{\infty} p^{\binom{\omega}{2}} \zeta^{\omega} \frac{v^{\omega}}{[\omega]_{p, q}!} \right) \\ &= \sum_{\omega=0}^{\infty} \left(\sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(c)}(\xi, \eta) \zeta^{\omega-\delta} \right) \frac{v^{\omega}}{[\omega]_{p, q}!}, \end{aligned}$$

yields the proof of (2.21). The proof of (2.22) is similar. \square

Theorem 2.7 For $\omega \geq 0$, we have

$$\sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(c)}(\xi, \eta) = [2]_{p, q} C_{\omega, p, q}(\xi, \eta), \quad (37)$$

$$\sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(s)}(\xi, \eta) = [2]_{p, q} S_{\omega, p, q}(\xi, \eta). \quad (38)$$

Proof From (2.21), we have

$$\mathbb{E}_{\omega, p, q}^{(c)}((\xi \oplus_{p, q} 1), \eta) + \mathbb{E}_{\omega, p, q}^{(c)}(\xi, \eta) = \sum_{\delta=0}^{\omega} \binom{\omega}{\delta}_{p, q} p^{\binom{\omega-\delta}{2}} \mathbb{E}_{\delta, p, q}^{(c)}(\xi, \eta).$$

Hence, by using (2.17), relation (2.23) is derived. The proof of (2.24) is concluded in a similar way. \square

Theorem 2.8 For $\omega \geq 0$, we have

$$D_{p,q;\xi} \mathbb{E}_{\omega,p,q}^{(c)}(\xi, \eta) = [\omega]_{p,q} \mathbb{E}_{\omega-1,p,q}^{(c)}(p\xi, \eta), \quad (39)$$

$$D_{p,q;\eta} \mathbb{E}_{\omega,p,q}^{(c)}(\xi, \eta) = -[\omega]_{p,q} \mathbb{E}_{\omega-1,p,q}^{(c)}(\xi, p\eta), \quad (40)$$

$$D_{p,q;\xi} \mathbb{E}_{\omega,p,q}^{(s)}(\xi, \eta) = [\omega]_{p,q} \mathbb{E}_{\omega-1,p,q}^{(s)}(p\xi, \eta), \quad (41)$$

$$D_{p,q;\eta} \mathbb{E}_{\omega}^{(s)}(\xi, \eta) = [\omega]_{p,q} \mathbb{E}_{\omega-1,p,q}^{(s)}(\xi, p\eta). \quad (42)$$

Proof Differentiating generating functions (2.9) and (2.10) with respect to ξ and η with the help of equation (1.7) and then simplifying with the help of the Cauchy product rule formulas (2.25), (2.26), (2.27) and (2.28) are obtained. \square

3 Conclusion

In this paper, we introduced (p, q) -Sine Euler polynomials and (p, q) -Cosine Euler polynomials and investigate some properties of these polynomials. By using summation techniques, we established some multifarious formulas and relationships for those polynomials. Furthermore, we derived the (p, q) -derivative operator and (p, q) -representations for the (p, q) -Sine Euler polynomials and (p, q) -Cosine Euler polynomials.

References

1. Corcino RB (2008) On p, q -binomial coefficients. Electron, J Combin Number Theo V.8, #A29
2. Duran U, Acikgoz M (2017) Apostol type (p, q) -Bernoulli, (p, q) -Euler and (p, q) -Genocchi polynomials and numbers. Comput Appl Math 8(1):7–30
3. Khan WA, Muhiuddin G, Duran U (2022) Al-Kadi, D, On (p, q) -Sine and (p, q) -Cosine Fubini polynomials. Symmetry 14(527):1–12
4. Khan WA, Khan IA, Duran U, Acikgoz M (2021) Apostol type (p, q) -Frobenius Eulerian polynomials and numbers. Afrika Matematika 32(1–2):115–130
5. Kang JY, Khan WA (2020) A new class of q -Hermite based Apostol-type Frobenius-Genocchi polynomials. Commun Korean Math Soc 35(3):759–771
6. Khan WA, Nisar KS, Baleanu D (2020) A note on (p, q) -analogue type of Fubini numbers and polynomials. AIMS Math 5(3):2743–2757
7. Khan WA, Khan IA, Ali M (2020) Degenerate Hermite poly-Bernoulli numbers and polynomials with q -parameter. Stud Univ Babes-Bolyai Math 65(1):3–15
8. Khan W (2022) A, A note on q -analogue of degenerate Catalan numbers associated p -adic integral on Z_p . Symmetry J 14(119):1–10. <https://doi.org/10.3390/sym14061119>

9. Khan, WA (2022) A note on q -analogues of degenerate Catalan-Daehee numbers and polynomials. *J Math*, Article ID 9486880, 9 pages
10. Nisar KS, Khan WA (2019) Notes on q -Hermite based unified Apostol type polynomials. *J Interdisc Math* 22(7):1185–1203
11. Milovanovic GV, Gupta V (2016) Malik, N, (p, q) -Beta functions and applications in approximation. *Bol Soc Mat Mex*. <https://doi.org/10.1007/s0590-016-0139-1>
12. Sadjang PN (2018) On the fundamental theorem of (p, q) -calculus and some (p, q) -Taylor formulas. *Results Math* 73:39
13. Sadjang PN (2018) On two (p, q) -analogues of the Laplace transform. *J Diff Eqn Appl* 23:1562–1583
14. Sadjang PN, Duran U (2019) On two bivariate kinds of (p, q) -Bernoulli polynomial., *Miskolc Math Notes* 20(2):1185–1199
15. Sharma SK, Khan WA, Ryoo CS, Duran U (2022) Diverse properties and approximate roots for a novel kinds of the (p, q) -Cosine and (p, q) -Sine Geometric polynomials. *Mathematics* 10, 2709, 1–18. <https://doi.org/10.3390/math10152709>

Different Stages of Watermelon Diseases Detection Using Optimized CNN



Samah Alhazmi

Abstract One of Saudi Arabia's biggest exports to its neighbors is fruits and vegetables. Some of its most productive crops include tomatoes, watermelons, grapes, oranges, onions, and citrus fruits. Watermelons and other fruits are prone to illness. A prompt and correct diagnosis of watermelon infections is necessary to ensure watermelon output. Identifying many stages of watermelon sickness in the environment is the aim of this study. Deep learning has been proposed as a potential tool for disease detection. In this paper, a deep learning approach is used to detect disease in watermelon plants. The convolutional neural network VGG-16 architecture was used in this study to recognize plant diseases and give farmers the tools they need to quickly treat afflicted plants. Unfortunately, because of a network over-fitting problem, the recall using the validation dataset was 0.7576 with the minimum score for true positives of 7%. The over-fitting issue is resolved in this study by contrasting two experiments. Different combinations of various hyperparameters make up each experiment. The initialization of the weights and the optimizer were the two key hyperparameters that were changed to enhance the model's performance. The final results demonstrate a significant advancement over the earlier studies. The enhanced model for detecting persistent objects has a recall of 0.9394 and a minimum true positive score of 98%.

Keywords Deep learning · Leaf diseases · Convolutional neural network · Image processing

1 Introduction

Humans frequently consume watermelon, which is generally available everywhere. It boasts outstanding nutritional qualities, a diverse phytochemical composition, and several asserted medical and health advantages. Lycopene, beta-carotene, phytofluene, phytoene, lutein, and neurosporene are the principal carotenoids found

S. Alhazmi (✉)

College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Kingdom of Saudi Arabia

e-mail: s.alhazmi@seu.edu.sa

in watermelon. The fruit's main bioactive ingredient, lycopene, is thought to have several therapeutic benefits for both people and animals, including anti-inflammatory and anti-cancer properties. Citrulline, an amino acid necessary for the synthesis of arginine, can also be found in watermelon. It is recognized that pre- and postharvest elements such as fruit sampling location, fertilizer application, meteorological elements, and genetic diversity affect the bioactive chemicals and nutrient concentrations of the fruit. One of the most important aspects of locating the disease is rapid detection. Detect plant viruses, bacteria, or fungi quickly in the field throughout the growing season. It is easy to manage [1]. Analyzing plant leaf diseases reveals the main problems and difficulties. Here are a few of them: (1) The leaf images must be of excellent quality, (2) the datasets must be open to the general public, (3) the fresh leaves are affected by noisy data, (4) diseases can be detected by segmentation, but the samples need to go through training and testing, (5) classification presents still another difficulty in the identification of leaf diseases, (6) environmental factors may cause variations in the tone of the leaves, and (7) since different types of plants can exhibit a variety of diseases, disease identification can be challenging.

Watermelon infections are common and can lower yield and quality. Diseases can damage all parts of the plant, from roots to fruit, at any point in the agricultural production cycle. When vines are harmed and died before they are fully ripe, diseases are most severe. However, ailments that affect the leaves decrease the quality of the fruit by subjecting melons to sunburn. Healthy foliage is also necessary for fruit flavor and good ripening. Fruit that has a disease is based on these requirements unmarketable and vulnerable to rot, which causes losses during shipping and transportation. Fortunately, watermelon infections can typically be controlled in order to maintain a regular production and quality. Microorganisms like bacteria, viruses, nematodes, and fungi cause watermelon illnesses. Environmental factors such as soil imbalances, extremes in soil moisture, and chemical damage can lead to abiotic illnesses. This article's goal is to make it easier to spot significant watermelon diseases. The detection of plant diseases is a crucial area for research in the domains of machine learning and computer vision. It is a device that analyzes gathered plant photos to determine whether there are any diseases or pests present [2]. Effective illness management begins with accurate disease detection. Ineffective management practices can be implemented as a result of incorrect identification, which can result in crop failure. (1) ***Anthracnose***: A typical foliar disease of watermelons is anthracnose the entire surface area of the plant is impacted, including the fruit, stem, and leaves. Fruit lesions are highly dangerous because they can develop from little, easily ignored areas just on fruit after harvesting into rotten, leaky melons during shipping. (2) ***Bacterial fruit blotch***: Watermelon can get the occasional illness bacterial fruit blotch. The disease does not do much damage to the leaf, but it badly affects the fruit, making it unusable. There have been claims of up to 50% yield decreases. (3) ***Cercospora leaf spot***: A fungus called Cercospora leaf spot usually has little impact but has the potential to defoliate plants and reduce productivity. (4) ***Downy mildew***: A sporadic but harmful disease that affects watermelons is downy mildew. The illness can cause a quick defoliation, which lowers fruit quality and exposes fruit to sunscald. Most cucurbit crops are impacted, albeit in some regions some crops are more negatively impacted

than others. (5) ***Fusarium wilt***: State-wide occurrence of the soilborne fungus illness fusarium wilt of watermelon. Where vulnerable types are planted, the disease can do significant harm because it kills the entire plant before harvest. (6) ***Gummy stem blight***: It is difficult to control the destructive foliar disease known as “watermelon sticky stem blight.” Fruit quality did drop. (7) ***Powdery mildew***: A foliar disease called powdery mildew is more noticeable on other cucurbits like squash and pumpkins. But the disease has become more significant to watermelon, especially in late plantings. Premature defoliation brought on by the disease can lower fruit quality and plant yield. (8) ***Virus diseases***: Most are potyviruses, which are members of the potato virus Y family of viruses. These involve (a) the watermelon mosaic virus, (b) the papaya ring spot virus, and (c) the yellow mosaic virus of zucchini. Reduced fruit set, decreased plant growth, and irregular fruit development all lead to losses. Early infection infections typically have a negative impact on the majority of crops. There may be other viruses, but their biology and treatment are similar [3]. (9) ***Root-knot nematode***: Nematodes are tiny, circular worm-like organisms that reside in soil. Nematodes that parasitize plants feed on their roots and, when populations are large, can stunt plant growth. Root-knot nematodes make galls and induce swelling in roots. Normal root function is hampered by severe root galling. Watermelons are among the several crop plants that are attacked by different types of root-knot nematodes. The southern underlying cause nematode is especially prevalent in small vegetable fields and gardens. However, commercial agricultural fields occasionally host the nematodes northern root-knot and peanuts root-knot. (10) ***Yellow vine***: The yellow vine poses a severe danger to watermelons, killing crops before they are harvested, although being less susceptible than pumpkins and squash. (11) ***Verticillium wilt***: Cotton and peanuts are also impacted by the verticillium disease. It harms the crops by killing plants before harvest and/or exposes mature fruit to sunscald and has an appearance similar to Fusarium wilt [4].

In this study, we present a method for predicting and categorizing illnesses in a watermelon plant on a big farm at an early stage. The suggested method uses preprocessing, segmentation, feature extraction, and classification as its four major processing approaches. In addition, we focus on applying the CNN technique with hyperparameter turning to segment the images that are derived from of the preprocessed images and categorize them [5].

2 Related Work

In some studies, researchers have developed a method called few-shot learning (FSL) for automatically categorizing numerous pests, plants, and their diseases. This approach offers performance improvements of between 14 and 24% for few-shot photo categorization [6]. Authors in [7] looked at novel CNN-based techniques utilized for UAV-based remote sensing data processing for crop and plant categorization in order to assist researchers and farmers in selecting the optimal algorithms based on the crops being studied and the hardware available. A sophisticated

Internet of Drones (IoD) network with intelligence may monitor farmlands using this method to boost crop productivity with minimal human involvement [7]. The machine learning model for determining the freshness of fruits was created by the researchers in [8]. The fruit's percentage of freshness will be predicted by the model. The created model's accuracy was estimated to be 0.989. The analysis's findings demonstrated that water vapor release and oxygen consumption changed over time in a progressive manner [8]. Moreover, in [9] both explore different plant leaf disorders and analyze them in different ways. After considering the aforementioned arranging strategies, they get to the following conclusion: Perhaps the easiest way for finding the class of a test model is the k-closest neighbor method. The K-NN method's significant negative is how time-consuming prediction requires [9]. Furthermore, researchers in [10] unveiled a brand-new platform for open agricultural data under the name Eden Library. The entire metadata schema, annotating process, and important facets of the platform have been established in order to assist other organizations in addressing the problem of data scarcity in machine learning strategies for agriculture [10].

In response to the authors in [11] work on the use of artificial intelligence and Internet of things for the identification of agricultural diseases, this survey study investigates the different challenges that must be solved as well as potential solutions. Then, a variety of ideas are provided to deal with these issues [11]. A quick and accurate diagnosis of the plant pathogens is provided to the farmer by an automated disease diagnostics system, according to [12]. This has allowed the diagnostic procedure to be sped up so the farmer can collect more harvests from his fields. The authors in [12] examine The Systematic Detection and Surveillance of Plant Disease Using Unmanned Aerial Vehicles. Making management decisions is made easier with the help of UAVs, which are precise and provide a wealth of information about crop status. Yet, there remains a huge market for diagnosing plant diseases [12]. Deep Learning applied Plant disease classification and detection is a topic discussed in the study done by [13], which describes how deep learning technologies have advanced recently for the detection of crop leaf disease. Moreover, researchers in [1][14] advise that use the EfficientNetV2 model to identify illnesses in cardamom plants. The suggested technique employs the U2-Net architecture to get rid of the intricate backdrop while maintaining the quality of the original image. Rather than using the pretrained values for EfficientNet and EfficientNetV2, CNN, EfficientNet, and EfficientNetV2 architectures were trained in this study for classification. According to the experimental findings, the suggested method has a detection performance of 98.26% [14]. A framework for portable devices was provided by [15] for the detection and monitoring of plant diseases. Improvements in rural and agricultural development and the capacity to transform the idea of "preventive activities" will make a difference in the global fight against phytopathogen. A model for detecting watermelon disease based on SSD was proposed by [16]. The accuracy rate of the finalized SSD768 model, according to experiments, is 92.4%. This technique could be used to identify watermelon illnesses naturally [16]. For the automatic identification and categorization of plant leaf diseases, a deep learning model was developed. The suggested model is tested on 13 different species. This model's forecast is essentially accurate [17]. Disease Detection in Fruits Via Deep Learning was proposed by

[18]. The major objective is to offer a quick, affordable way to identify fruit illnesses. For both feature extraction and classification, CNN is utilized. The accuracy of the suggested approach is 89.70%, as well as the loss value of 0.81 [18]. To forecast and categorize illnesses in melons, researchers in [19] proposed utilizing a stacked RNN-based deep learning algorithm. When compared to other existing models, this prediction and categorization of leaf diseases perform at the highest level with great accuracy and computational efficiency [19]. Convolutional neural networks were studied by [20] for the purpose of identifying plant leaf diseases. In this study, 100 of the most pertinent CNN articles on identifying plant leaf diseases were examined. Finally, they draw the conclusion that the best technique for identifying early disease detection is deep convolutional neural networks (DCNN) training on picture data.

Optimizing pretrained models convolutional neural networks in tomato detection of leaf disease was the idea put out by [21]. They test the key contributing factor on two datasets; one that was generated in a lab and one that was self-collected from the field. They discover that all configurations outperform the self-collected data gathered in the field by a wide margin on the laboratory-generated dataset, with performance on many metrics showing variance in the 10–15% range. On both datasets, Inception V3 is found to have the best performance [21]. It has been found [22] that YOLOv4 produces more precise and efficient results when object detection is done on a smartphone with an astounding 98.13% accuracy.

3 Proposed Work—Optimized CNN

While CNNs are comparable toward other neural networks, they use a number of convolutional layers, which adds another level of complexity. An integral part of convolutional neural networks (CNNs) is convolutional layers [23]. Different layers of CNNs are as follows [1]: (a) **Convolutional layer**, which are used to improve a source image, are produced by a set of filters. The convolutional layer produces a feature map, which is a description of the picture pixels with filters applied. It is possible to combine convolutional layers to create more complex systems that can retrieve finer features from images. (b) **Pooling layer**: Deep learning uses convolution layer of a pooling layer variety. The spatial dimension of the input is decreased by pooling layers, which speeds processing and consumes less memory. Pooling both expedites training and helps to reduce the amount of parameters. Max pooling and average pooling are the two primary types of pooling. While average pooling uses the overall average from every feature map, max pooling uses the maximum value. Following convolutional layers, pooling layers are typically used to minimize the number of input before they are transferred into a completely connected layer. (c) **Fully connected layer**: One of the most fundamental varieties of layer in a convolutional network is the fully connected layer. The name “completely connected” refers to the entire connectivity between every neurons in a layer and every other neuron in the layer below [24]. Fully linked layers are frequently used at the end of a CNN when the goal is to utilize the features learned by the prior layers as well as combine

them to create predictions. For example, the final fully integrated layer of a CNN can classify a picture as including an animal, such as a dog, cat, and bird, using the qualities that the prior layers have learned [25].

For image identification and classification applications, CNNs are frequently utilized. CNNs can be utilized, for example, to identify objects in photos or to classify photos as either showing a dog or a cat. CNN models can also be utilized for more challenging tasks, like as describing images or identifying their focus points. Additionally, time-series data, such audio or text data, can be employed with CNNs. CNNs are an effective deep learning technology that have been applied to numerous applications to produce cutting-edge outcomes [26].

In Fig. 1, convolutional neural networks (CNNs) are trained on a dataset for detecting leaf illness, and this dataset is used to construct the general structural architecture of the CNN Model of our suggested system with various layers. There are seven layers overall when the Keras API and Tensor Flow are utilized, the first two of which are convolutional (Conv2D) layers, the third and fourth of which are pooling (MaxPool2D) layers and the flattening layer, and the final two fully connected dense layers, which are simply ANN classifiers. Convolutional layer one is the top layer. Conv2D is the model we employed, and its learnable filters have a size range of 32 for the first filter and 64 for the last. By specifying the kernel size with the help of the kernel filter, each filter changes a specific area of the image. The entire image is subjected to the 5×5 kernel filter matrix. Filters can be thought of as feature map-based image transformations.

Performed several trials in CNN and varied the number of epochs, and discovered that after a predetermined number of epochs, the loss function's findings remained unchanged. In order to create a model that performs at its best, it is crucial to define the number of epochs between set values accurately. Therefore, restrict the network with an early halt to prevent excessive over-fitting. Different sets of hyperparameters are needed depending on the dataset to accurately predict. Figure 2 depicts the Hyperparameter Tuning procedure. Users find it tough to select one hyperparameter because there are so many of them. There is no definitive answer on the optimal number of layers, neurons, or optimizers for each dataset. To create the model from a particular dataset, it is crucial to identify the ideal sets of hyperparameters. The hyperparameters that need to be adjusted are the activation function, optimizer, learning rate, batch

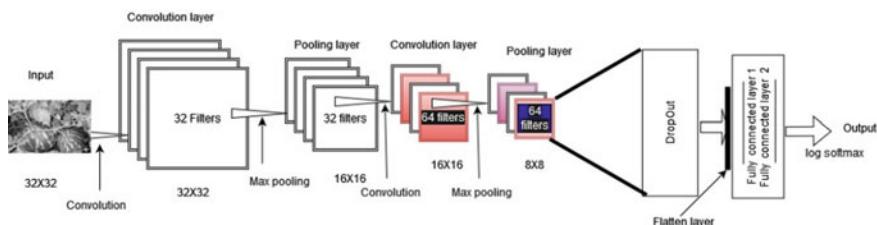


Fig. 1 CNN model [27]

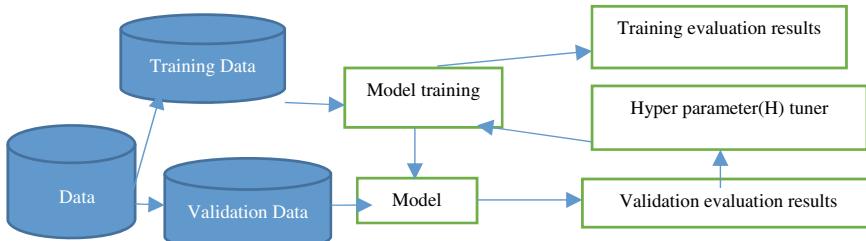


Fig. 2 Optimized CNN (Hyperparameter approach)

size, and epochs. The number of layers must be tuned in the second stage. Other traditional algorithms do not possess this [27]. Hyperparameters directly affect how the training algorithm behaves. These significantly affect how well a model performs. Hyperparameters are variables whose values influence the learning experience and define the model parameter values that a training algorithm ultimately learns [28].

3.1 Dataset and Pre-process

Building a precise diagnosis system requires the collection of adequate training data. For each disease, a tightly regulated field setting was created to solve this problem. To prevent contamination, the targeted disease was spread in a closed setting. Anthracnose, Bacterial fruit blotch, Downy mildew, Fusarium wilt, Gummy stem blight, Powdery mildew, viral infections, Root-knot nematode, Yellow vine, and Verticillium wilt were included in the dataset along with healthy leaves that had been afflicted with Downy mildew. The collection had 11,000 photos of leaves in total (1000 images per class). A leaf has to be present close to the center of the picture in order for it to be captured. Examples of the photos in the dataset are shown in Fig. 3. Images with various aspect ratios and sizes are included in the dataset. As a result, the picture was reduced to a square with its center in it.

Fig. 3 Downy mildew disease on a watermelon leaf



3.2 Initialization of Weights

In neural networks, the likelihood and rate of convergence, as well as the generalization that results, are influenced by the initialization of weights. Given that a network may not converge to the global minimum of the loss function during training, the initialization of weights is crucial. Although there is not currently a way for determining the ideal initialization of weights, in this study we chose several parameters to change the weights' initial value. Obtaining the weights of machine learning algorithms is frequently done using the optimization process known as gradient descent [29].

Use the following equation to find the average gradient of the current batch:

$$\alpha 1_m = H_1 \alpha 1_{m-1} + (1 - H_1) * G_m \quad (1)$$

$$\alpha 2_m = H_2 \alpha 2_{m-1} + (1 - H_2) * G_m^2 \quad (2)$$

where $\alpha 1$ and $\alpha 2$ are averages G is gradient on current batch, and there have been m iterations. The algorithm's hyperparameters are $H1$ and $H2$. Equations 1 and 2 are average of gradient and squared gradient, respectively.

3.3 Epoch

Epochs indicate the number of repetitions a neural network model is applied to a complete dataset. One neural network epoch represents one forward as well as one backward pass of the training dataset. Since when the neural network has not been trained enough, underfitting happens when there are not enough epochs. The training data or epochs must be iterated over several times [30]. We follow 50 and 80 epochs in our experiment. On the other hand, many epochs will cause overfitting, in which the model is able to forecast the data accurately but falls short of doing so for unexpected, novel data. The amount of epochs must be changed for the best results.

3.4 Batch Size

The amount of training samples utilized in a single iteration is known as the batch size. When employing varied batch sizes, the networks do not have to perform back-propagation on all of the samples in a single iteration. Despite some drawbacks, the model performed calculations more quickly, enabling better training and convergence speeds [31]. It is possible that the procedure does not produce the best yields when the batch size is equivalent to 1, as a result of production process variability. Higher computing costs result from iterating the weights and when the batch size is

equal to the number of elements in the dataset. For this reason, a value of around 1 and the total number of elements inside the collection of training data are typically chosen. In our test, the batch size was fixed at 2 and 4.

3.5 Intersection Over Union (IoU)

The two phases that make up the optimized CNN architecture use the IoU. Each object at each place of the images over the feature space is given a binary class name for the purpose of training the suggested regional network. Depending on intersection over union (IoU) with box containing the objects or box of truth, the label's value or score is determined [32]. P is the expected position, and B is the real position of the vessels. The optimum detection performance is when $\text{IoU} = 1$; however, $\text{IoU} > 0.5$ at real detection is still a very good result. The IoU value in our experiment was adjusted at 0.7 and 0.9. Enhancing the IoU would assist decrease object uncertainty and boost detection precision. A threshold value of 0.5–0.9 should be fixed.

3.6 Optimizers

The direction in which weights must be changed to enable convergence to the global or local minimum is generated by the optimizer [33]. The partial derivative of the weights, which is a mathematical approach, will show us the gradient and the direction of the minimum descent. We will be able to improve the weight calculations with the help of the optimizers. Less time will be available to minimize function loss. We utilize the moment estimation algorithm adaptive (ADAM), which is similar to stochastic gradient descent (SGD) and has the benefits of efficient computing and using less memory to update the weights and biases in the network (Table 1).

Table 1 Hyperparameter and metrics for experiment 1 and experiment 2

Hyperparameter	Experiment 1	Experiment 2
Epochs	50	80
Batch size	2	4
Optimizer	ADAM	SGDM
Weight initialization	$\mu = 0; \sigma = 0.05$	By default
Intersection over Union (IoU)	0.7	0.9
Metrics	Experiment 1	Experiment 2
Loss function	0.0437	0.0024
Recall	0.4545	0.9394
Minimum score	96%	98%

4 Result and Discussion

The existing Deep learning-based technique used by the SSD model to detect watermelon disease was 92.4% accurate. Only a few datasets were used in this SSD model. It essentially satisfies the requirements for quick recognition and identification of watermelon illnesses, but because there are fewer samples of leaf blight and the disease's characteristics are similar to those of leaf spot, the detection performance of the disease is rather low [16]. In this paper, we implement optimized CNN to detect disease in watermelon plants. The CNN accuracy in disease detection can be improved by using five sets of hyperparameters that are tested on two sets of experiments. For the case of Batch size = 2, we obtained higher cost function values; on the contrary, with Batch size = 4 we obtained values closer to 0. We observe that the combination $\text{IoU} = 0.9$ and Batch size = 4 reached the best value for the cost function. Looking at the predictions in we observe that only 15 objects were detected correctly, these are the real one's positives (TP) of a total of 33 objects. At the same time, the 18 objects that were not detected correspond to the false negative (FN). Likewise, we appreciate that the minimum score confidence reached 98%. We appreciate that the cost function improved significantly achieving a value of 0.0024. 2 Plot showing the accuracy and loss of the optimized CNN model during the training and validation phases with the dataset split 80%/20%. The model underwent 80 epochs of training. Figure 4 shows training and validation accuracy, and Fig. 5 shows training and validation loss. Each curve that is presented is derived from the history of the Keras model, which computes accuracy and loss for each epoch that the network completes. By contrasting the anticipated class with the actual class, accuracy is determined. The cross-entropy value between the predicted class and the actual class is used to calculate loss. In relation to recall, we observe that did not improve even though the confidence score value was raised. The criteria for choosing the best configuration has achieved good results, not only in the cost function and the confidence score but also in recall [16]. In addition to presenting hyperparameter turning techniques for disease detection in watermelon plants, this work serves as a template for the application of contemporary technology in agriculture, particularly in cash crops. With the aid of a computer vision system, it is possible to manage, analyze, and make decisions from vast amounts of camera-generated data [34].

5 Conclusion

Our proposed approach for diagnosing watermelon plant illnesses scans photos of plant leaves to judge whether a leaf is healthy or ill and identify whether a disease exists. The suggested system makes advantage of optimized CNN. Performance could be improved by adjusting the hyperparameters initialization of weights = 0;

Fig. 4 Training and validation accuracy

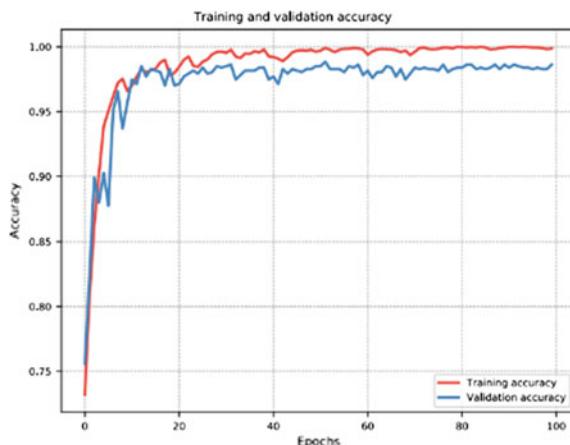
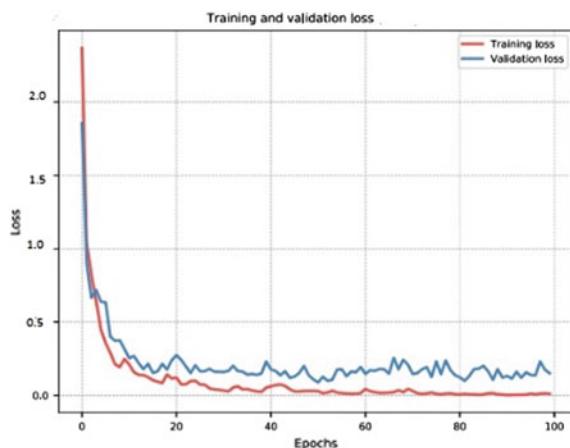


Fig. 5 Training and validation loss



$\sigma = 0.05$ and the SGDM optimizer. Based on the outcomes of experiments, the cost-based value of the trials fell from 0.0437 to 0.0024; the precision in terms of recall increased from 0.4545 to 0.9394; and the minimum confidence score increased from 96 to 98%. When the dataset includes images of leaves, the transfer learning method is useful. The acquired detection model can aid in the invention of conventional object detection algorithms by lowering the incidence of false alarms and information loss, particularly in coastal areas. Our model has certain restrictions because the input photographs must have particular lighting conditions and background clutter because they are obtained from genuine leaves of planted plants, even though it achieves high accuracy with a minimal loss. These conditions pose a challenge for any method used to detect plant diseases while creating a new model or trying to enhance an existing one. Future study should also concentrate on increasing the accuracy of disease detection, particularly by teaching our machine learning software to pinpoint the

exact location of the disease on each leaf, particularly if multiple diseases are found on a single plant leaf. To cover more plant diseases and crop varieties, the dataset for leaf diseases can be expanded.

References

1. Harakkannavar SS, Rudagi JM, Puranikmath VI, Siddiqua A, Pramodhini R (2022) Plant leaf disease detection using computer vision and machine learning algorithms. *Glob Transitions Proc* 3(1):305–310. <https://doi.org/10.1016/j.gtp.2022.03.016>
2. Saleem MH, Potgieter J, Arif KM, Member S (2022) A performance-optimized deep learning-based plant disease detection approach for horticultural crops of New Zealand. *IEEE Access* 10(August):89798–89822. <https://doi.org/10.1109/ACCESS.2022.3201104>
3. Georgia CB, Southern Codod, Severns R, Sparks PM, Srinivasan AN (2022) Characterization of the spatial distribution of the white fly- transmitted virus complex in yellow squash fields in. *Frontiers* (Boulder). Pp 1–21. <https://doi.org/10.3389/fagro.2022.930388>.
4. Dong X, Lian Q, Chen J (2022) The improved biocontrol agent, F1–35, Protects watermelon against fusarium wilt by triggering jasmonic acid and ethylene pathways. *Microorg Artic* 10(1710):1–17. <https://doi.org/10.3390/microorganisms10091710>
5. Zamani AS, Anand L, Rane KP (2022) Performance of machine learning and image processing in plant leaf disease detection. *J Food Qual* 2022:1–7. <https://doi.org/10.1155/2022/1598796>
6. Nuthalapati SV, Tunga A (2021) Multi-domain few-shot learning and dataset for agricultural applications. *Proc IEEE Int Conf Comput Vis* 2021(Octob):1399–1408. <https://doi.org/10.1109/ICCVW54120.2021.00161>
7. Bouguettaya A, Zarzour H, Kechida A, Taberkit AM (2022) Deep learning techniques to classify agricultural crops through UAV imagery: a review. *Neural Comput Appl* 34(12):9511–9536. <https://doi.org/10.1007/s00521-022-07104-9>
8. Jayasinghe PKSC, Sammani S (2022) Detection of freshness of the fruits using machine learning techniques. *SLJoT* 3(01):8–17
9. Gupta HK, Shah HR (2021) A review of different plant leaf diseases and an analysis of different plant leaf diseases identification techniques 10(5)
10. Mylonas N, Malouñas I, Mouseti S, Vali E, Espejo-Garcia B, Fountas S (2022) Eden library: a long-term database for storing agricultural multi-sensor datasets from UAV and proximal platforms. *Smart Agric Technol* 2(November 2021):100028. <https://doi.org/10.1016/j.atech.2021.100028>
11. Orchi H, Sadik M, Khaldoun M (2022) On using artificial intelligence and the internet of things for crop disease detection: a contemporary survey. *Agric* 12(1). <https://doi.org/10.3390/agriculture12010009>
12. Neupane K, Baysal-Gurel F (2021) Automatic identification and monitoring of plant diseases using unmanned aerial vehicles: a review. *Remote Sens* 13(19). <https://doi.org/10.3390/rs13193841>
13. Li L, Zhang S, Wang B (2021) Plant disease detection and classification by deep learning—a review. *IEEE Access* 9(Ccv):56683–56698. <https://doi.org/10.1109/ACCESS.2021.3069646>
14. Sunil CK, Jaidhar CD, Patil N (2022) Cardamom plant disease detection approach using EfficientNetV2. *IEEE Access* 10:789–804. <https://doi.org/10.1109/ACCESS.2021.3138920>
15. Buja I et al (2021) Advances in plant disease detection and monitoring: from traditional assays to in-field diagnostics. *Sensors* 21(6):1–22. <https://doi.org/10.3390/s21062129>
16. He X, Fang K, Qiao B, Zhu X, Chen Y (2021) Watermelon disease detection based on deep learning. *Int J Pattern Recognit Artif Intell* 35(5). <https://doi.org/10.1142/S0218001421520042>
17. Shrestha G, Deepshikha G, Das M, Dey N (2020) Plant disease detection using CNN. *Proc 2020 IEEE Appl Signal Process Conf ASPCON* 2020. 03049:109–113. <https://doi.org/10.1109/ASP CON49795.2020.9276722>

18. Hosakoti R, Kumar SP, Jain P (2021) Disease detection in fruits using deep learning. *J Univ Shanghai Sci Technol* 23(07):309–312. <https://doi.org/10.51201/jusst/21/07125>
19. Jayakumar D, Elakkia A, Rajmohan R, Ramkumar MO (2020) Automatic prediction and classification of diseases in melons using stacked RNN based deep learning model. In: 2020 international conference on system, computation, automation and networking (ICSCAN), 4–8. <https://doi.org/10.1109/ICSCAN49426.2020.9262414>
20. Tugrul B, Elfatimi E, Eryigit R (2022) Convolutional neural networks in detection of plant leaf diseases: a review. *Agriculture* 12(8):1192. <https://doi.org/10.3390/agriculture12081192>
21. Ahmad I, Hamid M, Yousaf S, Shah ST, Ahmad MO (2020) Optimizing pretrained convolutional neural networks for tomato leaf disease detection. *Complexity* 2020. <https://doi.org/10.1155/2020/8812019>
22. Jain S, Sahni R, Khargonkar T, Gupta H, Verma OP, Sharma TK (2022) Automatic rice disease detection and assistance framework using deep learning and a chatbot. *Electron* 11(14). <https://doi.org/10.3390/electronics11142110>
23. Barburiceanu S, Meza S, Orza B (2021) Convolutional neural networks for texture feature extraction. applications to leaf disease classification in precision agriculture. *IEEE Access* 9:160085–160103. <https://doi.org/10.1109/ACCESS.2021.3131002>
24. Oliveira IODE, Laroca R, Menotti D (2021) Vehicle-rear: a new dataset to explore feature fusion for vehicle identification using convolutional neural networks. *Digital Object Identifier* 9:101065–101077. <https://doi.org/10.1109/ACCESS.2021.3097964>
25. Reddy GRM, Sumanth NS, Kumar NSP (2020) Plant leaf disease detection using CNN and raspberry pi. *Int J Adv Sci Res* 5(2):21–25. [Online]. Available: www.ijpam.eu
26. Griggs KN, Ossipova O, Kohlios CP, Baccarini AN, Howson EA, Hayajneh T (2018) Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *J Med Syst* 42(7):1–8. <https://doi.org/10.1007/s10916-018-0982-x>
27. Carvalho BG, Brasileiro SA, Emanuel R, Vargas V, Brasileiro SA Hyperparameter tuning and feature selection for improving flow instability detection in offshore oil wells. In: IEEE 19th international conference on industrial informatics (INDIN), pp 1–6. <https://doi.org/10.1109/INDIN45523.2021.9557415>
28. Cheng M, Hung S, Tsai H, Chou Y (2020) Behavior: a hierarchical approach toward smartphone recycling. *IEEE Trans Eng Manag* 69(5):1–11. <https://doi.org/10.1109/TEM.2020.3007605>
29. Gonzales-martínez R, Machacuay J, Rotta P, Chinguel C (2022) Hyperparameters tuning of faster R-CNN deep learning transfer for persistent object detection in radar images 20(4):677–685
30. Parizad A, Member GS, Hatzidioniu C, Member S (2022) Deep learning algorithms and parallel distributed computing techniques for high-resolution load forecasting applying hyperparameter optimization 16(3):3758–3769
31. Lu G, Zhang W, Member S, Wang Z (2022) Optimizing depthwise separable convolution operations on GPUs. *IEEE Trans Parallel Distrib Syst* 33(1):70–87. <https://doi.org/10.1109/TPDS.2021.3084813>
32. Wang X, Song J (2021) ICIoU: improved loss based on complete intersection over union for bounding box regression. *IEEE Access* 9:105686–105695. <https://doi.org/10.1109/ACCESS.2021.3100414>
33. Ahmad M, Abdullah M, Moon H, Han D (2021) Plant disease detection in imbalanced datasets using efficient convolutional neural networks with stepwise transfer learning. *IEEE Access* 9:140565–140580. <https://doi.org/10.1109/ACCESS.2021.3119655>
34. Bhan Singh D, Kashyap M, Gupta H, Verma OP A deep learning-based food detection and classification system. *Comput Intell Based Solutions Vision Syst*

Machine Learning: An Analytical Approach for Pattern Detection in Diabetes



Ritu Chauhan, Anika Goel, Harleen Kaur, and Bhavya Alankar

Abstract In current times, the amount of data is increasing exponentially but generated data is not transformed into knowledge data which eventually serves no purpose. This issue can be resolved by applying machine learning approaches which prove to play a promising role in the field of prediction. Similarly, in healthcare industries, there is a supply of large amount of data for which it is necessary to extract information for prognosis, diagnosis, and medication development. In this paper, we have focused on diagnosis of diabetes which is a chronic disease with fastest global growth rate. The current study is carried out in two phases where in first phase data is pre-processed to remove the unwanted and missing values and in second phase the data is analyzed by model building using machine learning algorithm. Additionally, we have attempted machine learning techniques in our study such as decision tree classifier and logistic regression which can be used for diagnosis of diabetes, and we attain the accuracy of 75.9% for decision tree classifier and 74.6% for logistic regression.

Keywords Machine learning · Diabetes · Pattern detection · Decision-making

1 Introduction

Machine learning (ML) has changed the arena to discover knowledge from large-scale databases. Moreover, decade has passed away and several application domains have applied ML-based techniques to discover the significant patterns and make future decision-making techniques. Hence, decision-making techniques have overlaid the researchers and scientists through the time discover and automate process to assure better prediction modeling for future knowledge discovery. For example,

R. Chauhan (✉) · A. Goel

Center for Computational Biology and Bioinformatics, Amity University, Noida, Uttar Pradesh, India

e-mail: rchauhan@amity.edu

H. Kaur · B. Alankar

Department of Computer Science and Engineering, Jamia Hamdard, Delhi, India

several designed tools are based on automated technology where healthcare practitioners can manifest the details of the patients and determine the future prediction of diagnosis in disease. But these tools and technologies develop on the ideology of algorithm designing, whereas to measure the accuracy of each model is trivial in real-world application scenario [1, 2].

The current scope of study focuses on measuring the accuracy of these models and drawing the inference from the databases. However, these machine learning models tend to have algorithmic power to solve the problems with better accuracy and prediction rate. Also, we can conceptualize that ML has expanded the vision of automatization from laboratory to practical technology in worldwide commercial use. In addition to this, machine learning under artificial intelligence has become a choice for creating software for computer vision, speech recognition, natural language processing, robot control, and other applications.

In similar context, we can say machine learning provides a whole new dimension for interpreting the complex datasets which can be applied in various fields of biomedical sciences as predictive tool to help in making the decisions better. Moreover, machine learning-based algorithm with robust data analyzing ability has come up as a reliable methodology for clinical decision-making, that on comparing with traditional statistical methods which only relies on fixed equations and variables as model. So, machine learning has the ability to take account of all interaction and can provide new insights within the dataset for providing further update in the algorithms. Therefore, machine learning algorithm can improve pre-treatment predictions and make decision-making easy.

Moreover, these algorithms which enable a computer/machine to learn things over time through experiences are used to train machines which eventually analyze data, find hidden patterns in them, categorize them, and make predictions for the future that make them. All these characteristics make algorithms capable of reading the data and modifying its structure according to the observations made while going through the data. Usually, machine learning techniques are classified on the basis of supervised and unsupervised techniques where supervised learning has been discussed on the basis of labeled data, which can be used to train supervised model. Also, the labeled data will work as an instructor and the model will learn from the labels in the data. In addition, supervised learning is the most extensively utilized machine learning approach which includes e-mail spam classifiers, face recognizers over images, and medical diagnosis systems for patients. The objective is to generate a prediction y^* in response to a query x^* in the form of a group of (x, y) pairings. Simple vectors or more complex things like documents, graphs, DNA sequences, or images could be used as the x inputs. The investigation of various output y types is similar. Significant findings have been obtained [3] by concentrating on the straightforward binary classification problem, where y can take 1 of 2 values. However, unsupervised learning algorithms take the unlabeled data as input, find the common features, and group the data into clusters. These are the techniques which are used when we have a dataset which is non-classified or non-labeled. The technique is designed in a process where it can retrieve hidden patterns or knowledge from the unlabeled data. It does not find the actual output but instead it extracts observations from the dataset to locate the

hidden patterns in the dataset. Lastly, semi-supervised based learning algorithms are the machine learning techniques which lies between supervised learning and unsupervised learning; here, dataset is a mix of labeled and unlabeled data used. Generally, these learning techniques used smaller labeled data and more of unlabeled data.

Certainly, utilization of ML-based technology in real-world application has changed the vision of healthcare domain where different health issues can be prevented, detected, and treated with the aid of machine learning. The finest tools for enhancing the healthcare system are those that use machine learning and data mining. Doctors' manual detection or diagnosis takes a long time and tends to be inaccurate using large databases. Hence, ML-based techniques can help in detection and prognosis of the disease and the issue. Machine learning (ML) can be useful for more than only disease diagnosis and prediction; it can also be useful for behavioral change, drug manufacture, finding new patterns that lead to new medications and treatments, clinical trial research, and smart electronic health records. We can accomplish all of this and more with machine learning.

Furthermore, in the medical industry, a classification algorithm has been used to categorize data into different groups following specified constraints as opposed to using a single classifier. In this article, we have emphasized on detection of diabetes utilizing varied ML-based algorithms. The study was gasped considering the fact that critical illness of diabetes and their related features which inhabited due to medical conditions.

Further, we have evaluated different factors which are responsible for diabetes. To support above approach, database has been trained to perform the task, and based on the training, it can handle the task without being programmed explicitly. The current study is carried out in two phases where in first phase data is pre-processed to remove the unwanted and missing values and in second phase the data is analyzed by model building using machine learning algorithm. To build this whole model, data is divided into two sets, training data and test data in which model learns from the training data and gives output according to test data. At last, the results were analyzed on the basis of different accuracy measures.

Finally, the paper has been overall discussed as follows: Section 1 represents the introduction in relative to ML and diabetes-based terminology; Section 2 discusses the previous suggested studies in the past; Section 3 overlays the methodology discussed and results are elaborated in Section 4; finally, conclusion is discussed.

2 Literature Review

Diabetes is a fast-growing epidemic in India, with 62 million people suffering from diabetes. In 2000, India has the most case of diabetes mellitus followed by China and the United States which comes second and third. According to estimates, 79.4 million people will be diagnosed with diabetes mellitus which is also expected to rise significantly [4]. Many factors influence the disease, and identification of these

factors is necessary to make the required changes in treatment. In recent years, many people from different parts of the world work in predictive analysis and big data analytics. In [5], there is the prediction of Hydrocephalus using image analysis. They took 77 characteristics from the image. SVM, an ML method, was used to analyze the traits of 25 kids. Results showed that three out of four patients required shunts, with responsiveness and accuracy of 75% and 95%, respectively [5]. A prognosis of the type of diabetes, complications, and therapy may be found in [6]. Prediction and the type of therapy were done using the Hadoop map-reduce and predictive analysis technique. [7] presented literature survey over big data and analytics literature. The authors' main objective was to apply ML techniques to industrial power systems and application for failure and power load prediction. In prediction, health care is based on the Naïve Bayes algorithm. Data has been extracted from different disease databases. Firstly, the user shares their health-related problem and then the system predicts the illness based on the Naïve Bayes algorithm. For a long-term diseases like heart disease, ML algorithms are used. The authors suggested a brand-new multimodal illness risk prediction system based on convolutional neural networks. Real-world hospital data for the years 2013–2015 was gathered in central China to evaluate the suggested algorithm [8]. A cerebral infarction experiment was conducted on a chronic condition. The results of the experiments demonstrate that the suggested method, when paired with both structured and text data, outperforms Naïve Bayes for structured data. The authors utilized sepsis mortality as the prediction use case due to the clinical significance of sepsis. For 12 months, data was collected from four emergency rooms. K-mean clustering was employed for data processing and clustering, and the random forest technique was applied for prediction [9]. Traditional models of prediction utilized in an emergency include the logistic regression model and CART. Outcomes indicate that random forest predicts outcomes more accurately than other models. In [10], there is a study on cases of dengue and malaria in Delhi. In [11], there is work on the infertility of women. In [12], the author proposed a model that predicts heart disease. Predicting algorithms also provide knowledge to patients about tests and doctor consultations. This recommendation system works based on a time data series algorithm model to determine whether a patient has diabetes or not has been created based on diabetic risk indicators and a decision tree classifier. Additionally, the functions of bagging and the AdaBoost ensemble machine learning method are discussed [12]. A technique for diabetes prediction is described in [13], which concludes that a patient is experiencing diabetes at a specific age. Utilizing a decision tree classifier, the designed system successfully predicted the onset of diabetes at a specific age [14]. Data for the prediction of diabetes is trained and tested using genetic programming. In comparison with other methodologies, the results from genetic programming fell short of expectations. However, there is a speed improvement because it is quicker than other classifiers. It is also of low cost [15]. There is an application of algorithms that predict the risk of diabetes mellitus. In [15], the author used the four machine learning algorithms namely ANN, DT, LR, and NB. For improving the model, two techniques are used Bagging and Boosting. Results from experiments suggest that the Random forest algorithm produces the best outcome [16–18].

3 Methodology Applied

Diabetes is a condition that impairs the body's ability to build the hormone insulin, results in improper carbohydrate metabolism, and raises blood glucose levels [19]. A person with diabetes typically experiences elevated blood sugar. Increased hunger, frequent urination, and increased thirst are a few signs and symptoms of high blood sugar. If diabetes is not addressed, several complications happen. Diabetic ketoacidosis and non-ketotic hyperosmolar coma are two examples of serious consequences.

For this purpose, in this paper we have used two techniques of machine learning—logistic regression and decision tree to predict the diabetes of a patient. Using these techniques over PIMA database helps to find out whether person is diabetic or non-diabetic. It will help the pre-diabetic patient to avoid diabetes through some changes in lifestyle or at least delay it for some span of time.

3.1 *Dataset Description*

To support the approach, Pima Dataset is collected from Kaggle, which has 768 records and 9 attributes. The NIDDK is the source of the dataset. The dataset of all patients is female and at least twenty-one years old. On the basis of these 9 attributes, paper shows value of one for patients who have diabetes and a value of zero for those who do not. Using a variety of patient diagnostic tests, the database's primary goal is to determine whether a person has diabetes or not. The nine features and their descriptions are shown in Table 1 as count, mean, standard deviation, maximum value, and minimum value.

Table 1 Data statistics

Attributes	Count	Mean	STD	Max	Min
Pregnant	768	3.84	3.36	17.00	0.00
Glucose	768	120.89	31.97	199.00	0.00
BP	768	69.10	19.35	122.00	0.00
Skin	768	20.53	15.95	99.00	0.00
Insulin	768	79.79	115.24	846.00	0.00
BMI	768	31.99	7.88	67.00	27.3
Pedigree	768	0.47	0.33	02.42	0.24
Age	768	33.24	11.76	81.00	24.00
Label	768	0.34	0.47	01.00	0.00

3.2 Data Pre-processing

All the data was collected in excel sheet form so that it can run in ML algorithms. We used the Jupyter notebook (IDE), a python-based platform to build our ML algorithms. To analyze our data using the ML algorithm, we need to import our data into the Jupyter notebook by reading our file using its location address. Once we have the data, we need to do data pre-processing, i.e., removing unwanted data, and filling the null values if any. After doing this, we need to import the libraries to use the required functions in python. We used libraries like pandas, seaborn, matplotlib, pyplot, random, and CSV; these are set of commands which make certain codes functional for the user with actually running those codes individually. Data pre-processing is important as it decreases the variance in the dataset, removes unwanted data, and improves the accuracy of the model which will use this data as it also includes removing the extremes in the dataset. For each machine learning algorithm, we used for the dataset data pre-processing was done and only after going through this step the next step was initiated.

3.3 Algorithm Used

3.3.1 Decision Tree Classifier

To deal with classification problems, a supervised machine learning technique called decision tree is used. Using a decision rule drawn from past data, the decision tree's main objective in this study is to forecast the target class. It uses nodes and internodes for categorization and prediction. Root nodes classify the instances based on several properties. The root nodes may have two or more branches, whereas the leaf nodes reflect classification [3].

3.3.2 Logistic Regression

ML algorithms such as logistic regression are one of the most frequently used in the supervised learning category. Using already known independent factors, it is used to predict the dependent variable. The dependent variable is predicted as an outcome by logistic regression. Instead of giving just 0 or 1, it gives the exact value between 0 and 1.

3.4 Accuracy Measures

Two techniques employed in this experiment are decision trees and logistic regression. Recall, accuracy, precision, and ROC curve are the metrics employed in the classification of this experiment. Using the abbreviations TP for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative in below Eqs. (1, 2, 3):

$$\text{Accuracy} = \frac{((\text{TP} + \text{TN}))}{(\text{Total no of samples})} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \quad (3)$$

4 Results

4.1 Model Building

The likelihood that a person will get diabetes was predicted using two distinct machine learning classifiers (decision tree and logistic regression). The dependent variable for the dataset was “Outcome,” whereas the other features were all independent. The dataset was split into random training and testing subsets, with dependent and independent features segregated. The division ratio was 80:20 for train and test data, respectively. The classification and regression models were trained on training data, and their performance was evaluated on test data. The model was selected for building a prediction algorithm; using both datasets, prediction models were built. After dividing the data into two, for all the algorithms model, data was transformed and fit into the model using the codes for transforming the data and fitting the algorithm on the data. After this, the models were trained using the training data only. Here, the model is shown the variables as well as the result they had. Based on the data, the model learns from the readings in the feature and selects which feature reading impacts the outcome and in what way. The models develop their learning and are ready for prediction. Now, the model is introduced to the rest of the data and this is called the testing data; on this part of the data, the model is checked for its prediction accuracy. As for this part too, the results are known to us, but we want the model to give a prediction for this data and after the model predicts the outcome on the test data, the predicted result, and actual results are compared to know the accuracy of the prediction model. In this way, all the prediction algorithms were executed and the results of the accuracy were noted down which gave us an idea about how

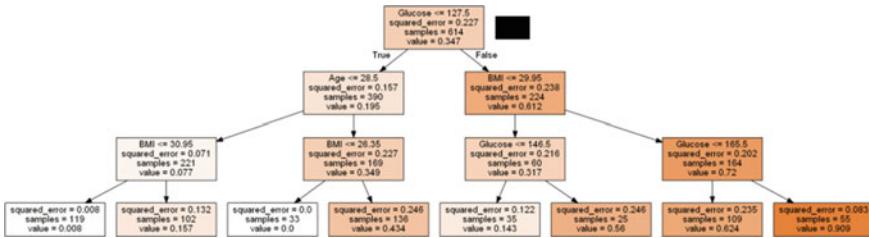


Fig. 1 Decision tree

good the models are with the prediction they are making. The models were built for both datasets and were executed separately. The results of accuracy were noted for further analysis.

4.2 Model Validation

To validate the model, we have calculated accuracy, precision, and recall. Their comparison is given in Fig. 4. The ROC curve has also been made for both classifier and regression to evaluate the effectiveness shown in Figs. 5 and 6.

The R curve is a tool in which ROC is used to validate the accuracy shown by ML algorithms. It divides the result into two by drawing a curve in the graph. Anything under the curve goes as false positive and prediction above the curve counts as true positive. Model validation is required to know which prediction result is more favorable. For the dataset, the decision tree classifier is a better-performing model in comparison with logistic regression with an accuracy of 75.9%. In Fig. 1, decision tree has been depicted. In order to predict, we start from the top and traverse to the deep to make the decision using the tree. Figures 2 and 3 show how the decision tree approach and logistic regression's performance were evaluated using the Confusion Matrix.

5 Conclusion

Predictive analytics' significance in the healthcare sector can change how researchers and professionals in the area evaluate medical data and make decisions. The two well-known computer algorithms are employed. These techniques include decision trees and logistic regression. Diabetes predictions were produced using the 768 records in the PIMA Indian dataset. Nine features were selected for the predictive model's training and testing. Decision trees are effective approaches for predicting diabetes, according to experimental results. Some of the study's limitations are the size of the dataset and the lack of attribute values. To create a diabetes prediction model with a

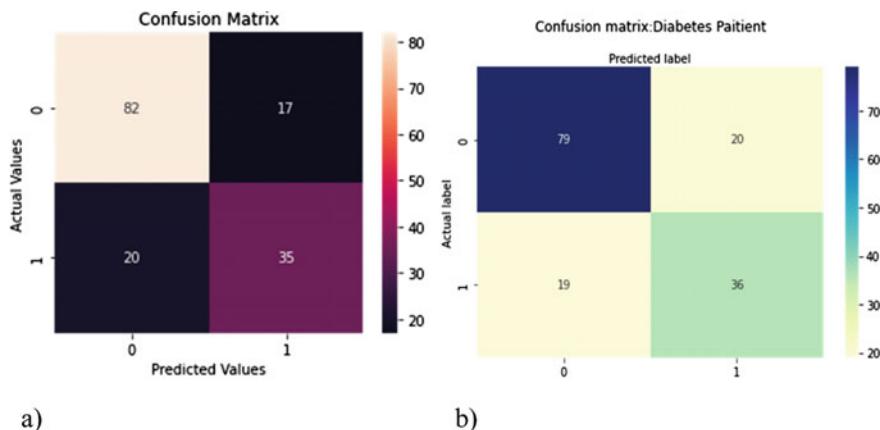


Fig. 2 Confusion matrix of diabetes patient using **a** decision tree **b** logistic regression

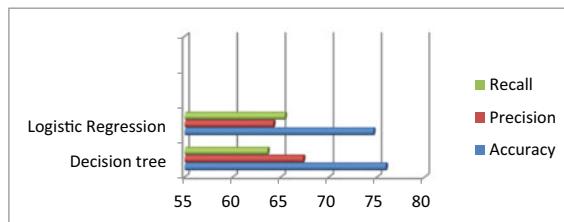


Fig. 3 Comparison between decision tree classifier and logistic regression

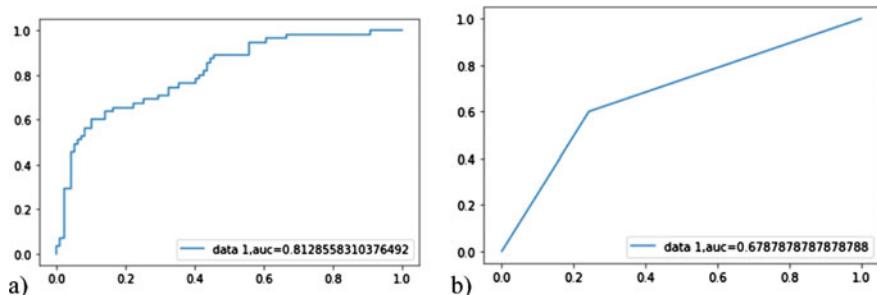


Fig. 4 ROC curve for **a** logistic regression **b** decision tree classifier

99.99% accuracy, we need tens of thousands of records with no missing information. In the future, we will focus on adding more methods to the model to increase the precision of its parameters. Then, new insights and increased predictability will result from testing these models on a big dataset with few to no missing attribute values.

Acknowledgements This research work was catalyzed and supported by the Ministry of Electronics and Information Technology, Govt. of India, New Delhi, India [grant recipient: Dr. Ritu Chauhan and Dr. Harleen Kaur and grant No. 3080229].

References

1. Horvitz E, Mulligan D (2015) Data, privacy, and the greater good. *Science* 349(6245):253–255. <https://doi.org/10.1126/science.aac4520>
2. Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, Zdeborová L (2019) Machine learning and the physical sciences. *Rev Modern Phys* 91(4). <https://doi.org/10.1103/RevModPhys.91.045002>
3. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. *Proc Comput Sci* 132:1578–1585. <https://doi.org/10.1016/j.procs.2018.05.122>
4. Kaul K, Tarr JM, Ahmad SI, Kohner EM, Chibber R (2012) Introduction to diabetes mellitus. *Adv Exp Med Biol* 771:1–11. https://doi.org/10.1007/978-1-4614-5441-0_1 PMID:23393665
5. Kaveeshwar SA, Cornwall J (2014) The current state of diabetes mellitus in India. *Australas Med J* 7(1):45–48. <https://doi.org/10.4066/AMJ.2014.1979>
6. Chiarelli PA, Hauptman JS, Browd SR (2018) Machine learning and the prediction of hydrocephalus. *JAMA Pediatr* 172(2):116
7. Kumar NMS, Eswari T, Sampath P, Lavanya S (2015) Predictive methodology for diabetic data analysis in big data. *Procedia Comput Sci* 50:203–208
8. Chen M, Hao Y, Hwang K, Wang L (2017) Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* 5:8869–8879
9. Taylor RA et al (2016) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data driven, machine learning approach. *Acad Emerg Med* 23(3):269–278
10. Das S, Thakral A (2016) Predictive analysis of dengue and malaria. In: 2016 International conference on computing, communication and automation (ICCCA), pp 172–176
11. Lafta R, Zhang J, Tao X, Li Y, Tseng VS (2015) An intelligent recommender system based on short-term risk prediction for heart disease patients. In: IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), pp 102–105
12. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput Sci* 82:115–121. <https://doi.org/10.1016/j.procs.2016.04.016>
13. Orabi KM, Kamal YM, Rabah TM (2016) Early predictive system for diabetes mellitus disease. In: Industrial conference on datamining, Springer, pp 420–427
14. M P, G R (2014) Design of classifier for detection of diabetes mellitus using genetic programming. *Adv Intell Syst Comput* 1:763–770. <https://doi.org/10.1007/978-3-319-11933-5>
15. Nai-Arun N, Moungmai R (2015) Comparison of Classifiers for the risk of diabetes prediction. *Procedia Comput Sci* 69:132–142. <https://doi.org/10.1016/j.procs.2015.10.014>
16. Chauhan R, Kaur H, Chang V (2017) Advancement and applicability of classifiers for variant exponential model to optimize the accuracy for deep learning. *J Ambient Intell Human Comput* <https://doi.org/10.1007/s12652-017-0561-x>
17. Chauhan R, Kaur H, Alankar B (2021) Air quality forecast using convolutional neural network for sustainable development in urban environments. *Sustain Cities Soc* 75:103239. <https://doi.org/10.1016/j.scs.2021.103239>

18. Kumar N, Chauhan R, Dubey G (2020) Applicability of financial system using deep learning techniques. In: Hu YC, Tiwari S, Trivedi M, Mishra K (eds) Ambient communications and computer systems. Advances in intelligent systems and computing, vol 1097. Springer, Singapore. https://doi.org/10.1007/978-981-15-1518-7_11
19. Sarwar MA, Kamal N, Hamid W, Shah MA (n.d.) Prediction of diabetes using machine learning algorithms in healthcare 1:1–6. [https://doi.org/10.23919/IConAC.2018.8748992.\(2018\)](https://doi.org/10.23919/IConAC.2018.8748992.(2018))

A Dynamic Weighted Federated Learning for Android Malware Classification



Ayushi Chaudhuri , Arijit Nandi , and Buddhadeb Pradhan 

Abstract Android malware attacks are increasing daily at a tremendous volume, making android users more vulnerable to cyber-attacks. Researchers have developed many machine learning (ML)/deep learning (DL) techniques to detect and mitigate android malware attacks. However, due to technological advancement, there is a rise in android mobile devices. Furthermore, the devices are geographically dispersed, resulting in distributed data. In such scenario, traditional ML/DL techniques are infeasible since all of these approaches require the data to be kept in a central system; this may provide a problem for user privacy because of the massive proliferation of android mobile devices; putting the data in a central system creates an overhead. Also, the traditional ML/DL-based android malware classification techniques are not scalable. Researchers have proposed federated learning (FL)-based android malware classification system to solve the privacy preservation and scalability with high classification performance. In traditional FL, federated averaging (FedAvg) is utilized to construct the global model at each round by merging all of the local models obtained from all of the customers that participated in the FL. However, the conventional FedAvg has a disadvantage: if one poor performing local model is included in global model development for each round, it may result in an under-performing global model. Because FedAvg favors all local models equally when averaging. To address this issue, our main objective in this work is to design a dynamic weighted federated averaging (DW-FedAvg) strategy in which the weights for each local model are automatically updated based on their performance at the client. The DW-FedAvg is evaluated using four popular benchmark datasets, Melgenome, Drebin, Kronodroid,

A. Chaudhuri

Department of Computer Science and Engineering, Vellore Institute of Technology (VIT),
Bhopal, India

A. Nandi 

Department of Computer Science, Universitat Politècnica de Catalunya (Barcelona Tech),
Barcelona, Spain
e-mail: jit.ari172@gmail.com

Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain

B. Pradhan

Department of Computer Science and Engineering, University of Engineering
and Management, Kolkata, India

and Tuandromd used in android malware classification research. The results show that our proposed approach is scalable, privacy preserved, and capable of outperforming the traditional FedAvg for android malware classification in terms of accuracy, *F*1-score, AUC score, and FPR score.

Keywords Android malware classification · Federated learning · Android security · Distributed machine learning · Artificial neural network

1 Introduction

Nowadays, android has become one of the most widely and popularly used operating systems [2]. Also, the threats (android malware) in the android operating system have increased at a rapid rate. It has been found that the share of android malware is higher than 46% among all types of mobile malwares and 400% increase in android-based malware since 2010 [13]. Malware is a type of malicious software that targets the mobile devices running on android operating system. Android devices now have new features for installing and using apps compared to traditional computers. Because of this new feature, there is a chance that the android device becomes vulnerable to android malwares [1] and it could be even difficult to safe guard the devices properly from the malwares. Malware infected devices posses serious threats to not only the users but also an organization.

With the advancement of machine learning (ML) or deep learning (DL) algorithms, their application in android malware classification is increasing day by day due to its effectiveness [10]. Furthermore, traditional ML/DL-based malware classification techniques are not suitable and scalable in the current scenario (rapid growth of mobile devices) because of the following reasons:

1. Decentralized data: Users are geographically located so generated data is distributed.
2. Data has sensitive information such as location data and online identifiers (IP address).

Federated cybersecurity is one of the newest and emerging approaches in malware detection and classification [5]. It makes the detection of cyber threats more secure and also use the IoT network system efficiently. This paper focuses on a detailed study of federated models for cybersecurity and machine learning by dividing them into two parts. The first one describes the FL and how it can be applied in cybersecurity for IoT and the second part addresses cybersecurity for FL. This survey mainly focuses on security approaches and also gives importance to performance issues related to FL. Security attacks and preventive measures are summarized and also performance issues in FL for IoT networks are also described proficiently. In [9], authors have proposed a federated learning (FL) architecture along with android malware detection algorithm, known as Fed-IIoT. This architecture consists of two parts, first one is participant side, where the data has been triggered by two dynamic poi-

soning attacks based on generative adversarial network (GAN) and federated GAN; and the second one is server side, which monitors the global model and gives a shape to the robust collaboration training model. This model proposes to avoid anomaly in aggregation by a GAN network defense algorithm to detect the vulnerabilities in the server side and also adjusts and adapts Byzantine defense algorithm on Krum and Medium to increase its effectiveness. By using Fed-IIoT, devices can safely communicate with each other with no privacy issues. These features are employed to classify different malwares by using convolutional neural networks (CNNs). Similar kind of research found in [4], where a new technique called permission maps is developed, which provides combined information of android permissions and their severity levels. The training phase of the Perm-Maps is supported by a federated architecture. A CNN model then is employed to classify different malware families. At last, a feature selection approach is applied to reduce the computational effort and CNN training processes. Authors in [7] has developed ‘Less is More’ (LiM), which is a malware classification framework that uses federated learning to detect and categorize dangerous programs by protecting the privacy of others. LiM employs a safe semi-supervised ensemble learning and FL that maximizes malware classification accuracy. Researchers have proposed federated learning-based approaches for android malware classification [9]. In traditional FL, federated averaging (FedAvg) is utilized to construct the global model at each round by merging all of the local models obtained from all of the customers that participated in the FL. However, the conventional FedAvg has a disadvantage: if one poor performing local model is included in global model development for each round, it may result in an under-performing global model. Because FedAvg favors all local models equally when averaging.

To solve this above mentioned issues, we propose a dynamic weighted federated averaging (DW-FedAvg) strategy in which the weights for each local model are automatically updated based on their performance at the client to classify android malware classification. The main contributions of our paper are as follows:

- We present a dynamic weighted federated averaging (DW-FedAvg) strategy in federated learning framework for android malware classification.
- The proposed DW-FedAvg is capable of delivering a high accuracy global classifier without accessing the distributed data while classifying android malwares.
- The experimental results based on the popular Drebin, Malgenome, Kronodroid, and Tuandromd datasets showed that the DW-FedAvg achieves high accuracy, and it is scalable.

The rest of the paper is structured as follows: Brief introduction to the ideas of android malware and federate learning is presented in Sect. 2. In Sect. 3, the material and methods for DW-FedAvg is presented and in Sect. 4, the experiment results and discussion is provided. Finally, the paper ends with the conclusion in Sect. 5.

2 Preliminaries

Here, in this section, we have provided the brief introduction about android malware and federated learning.

2.1 *Android Malware*

Android malware is a type of malicious software that targets the mobile devices running on android operating system. It has been growing at a significant rate. It has been found that the share of android malware is higher than 46% than that among all types of mobile malwares [13]. There is also 400% increase in android-based malware since 2010 [13]. Android devices now have new features for installing and using apps compared to traditional computers. This new feature makes it even difficult for the android operating system to defend against android malware [1].

2.2 *Federated Learning*

Federated learning (FL [8, 12]) is a promising distributed machine learning approach that enables mobile local clients to build a powerful global model by collaborating with a global server. The mobile devices share a local model developed from the sensitive data accessible to those devices without sharing the sensitive data with other parties. Clearly, in FL, there are two parties involved in the whole approach, global server and local clients (end users or edge user devices) [8].

The main functions of local clients are: (1) access to the local data, develop the model and perform the corresponding learning task (2) share the developed local model to the global server (3) receive the global model and replace the local model with the global model.

The main functions of the global server are: (1) collect the shared local models from local clients; (2) create the global model by model averaging (called federated averaging) of all the collected local models; (3) broadcast the global model to all the local models participated in the federated averaging (FedAvg).

In federated learning, the global server and local clients continue performing the above mentioned functions until a desirable model accuracy is achieved [11]. In Fig. 1, the general framework of FL is presented.

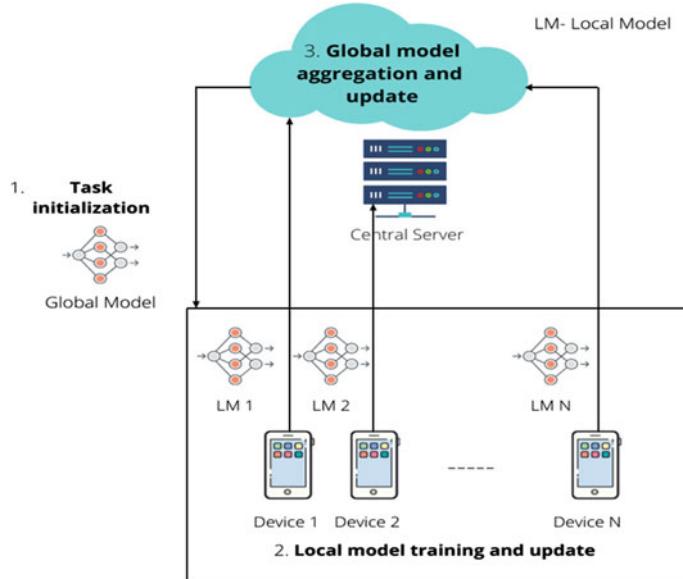


Fig. 1 General framework of federated learning

3 Materials and Methods

In this section, the proposed approach details, the benchmark dataset description, experimental setup, and finally, the performance metric are provided.

3.1 Proposed Approach

At each round of FL, a global model is created by averaging (federated averaging (FedAvg)) all of the local models received from all of the clients that participated in the FL. However, the traditional FedAvg has a drawback: such as if one poor performing local model is included in global model construction for each round, it might result in an under-performing global model. Because FedAvg prioritizes all local models equally while averaging. In this scenario, we can use the weighted averaging approach, in which a weight is assigned to each of the local models. The weight assigned to each local model is very challenging because it is done by trial and error and the overall performance depends on the weight assignment. So, to solve this problem, in this paper, our main motivation is to develop a dynamic weighted federated averaging (DW-FedAvg) approach where the weights for each local model are adjusted automatically based on their performance at the client. The FedAvg is performed at global server based on the following Eq. (1) as follows:

$$w_t^g = \frac{1}{N_c} \sum_1^{N_c} w_{t-1,i}^l \quad (1)$$

Now, the DW-FedAvg is performed based on Eq. (2) at global server:

$$w_t^g = \frac{1}{N_c} \sum_1^{N_c} \beta_i \cdot w_{t-1,i}^l \quad (2)$$

where the w_t^g is the global model created at time t , N_c is the total number of local model received at global server, it also indicates total number of clients participated in the FL $w_{t-1,i}^l$ is the local model received from all the clients at time $t - 1$ and the β is the dynamic weights associated with each local models received. In this paper, we have considered full participation from all the clients in the FedAvg and DW-FedAvg.

The dynamic weight β is adjusted automatically based on each local model's performance at client's end. To do so, the global server first considers each local model has equal powerful, hence, equal priority. The global server creates a global priority index matrix which contains the weights corresponding to each local models. Then, the weights are adjusted automatically based on the local models performance. The dynamic weight changes are based on the following condition:

- **Initialize:** $\beta_{i,t} = \frac{1}{N_c}$, α (the reward/penalty factor) = 0.2.
- $\text{Acc}_p \leftarrow 0$ and $\text{Acc}_c \leftarrow$ all local model's test accuracy.
- If round == 1 then $\text{Acc}_p = \text{Acc}_c$.
- Else:
 - If $\text{Acc}_{ci} > \text{Acc}_{pi}$ then $\beta_{t,i} = \beta_{t,i} + \beta_{t,i} * \alpha$.
 - ElseIf $\text{Acc}_{ci} < \text{Acc}_{pi}$ then $\beta_{t,i} = \beta_{t,i} - \beta_{t,i} * \alpha$.
 - Else do nothing.
- Weight re-scaling: $\beta_{t,i} = \frac{\beta_{t,i}}{\sum(\beta_{t,i})}$.
- repeat everything till all rounds end.

3.2 Dataset Description

We have used four datasets for our approach which are publicly available. The brief description of those considered datasets is as follows:

1. **Malgenome:** This dataset contains features from 3799 app samples where 2539 are benign and 1260 are android malwares from android malware genome project [10]. It contains total features of 215.
2. **Drebin:** This dataset contains features from 15,036 app samples where 9476 are benign and 5560 are android malwares from Drebin project [12]. It also contains 215 features.

Table 1 Dataset description

Dataset name	No. of samples	No. of attributes/features	Class labels
Malgenome	3799	215	Benign (2539) Malware (1260)
Drebin	15,036	215	Benign (9476) Malware (5560)
Tunadromd	4465	241	Benign (903) Malware (3565)
Kronodrid	78,137	463	Benign (36,935) Malware (41,382)

3. **Tunadromd:** This dataset [3] contains features from 4465 app samples where 903 are benign and 3565 are android malwares. It contains total features of 241.
4. **Kronodrid:** This dataset contains features from 78,137 app samples where 36935 are benign and 41,382 are android malwares [6]. It contains total features of 463.

The brief description of those considered benchmark datasets is presented in Table 1.

3.3 Experimental Setup

Herewith the machine setup, software development, experimental environment, and parameter setup are as follows:

- *Machine configuration:* Ubuntu 20.04 64 bit OS, processor core-i7-7700HQ with RAM 24 Gb–2400 MHz and 4Gb-Nvidia GTX-1050 graphics.
- *Software development:* The DW-FedAvg is implemented in Python 3.7 with the help of TensorFlow 2.0 and Keras in the backend.
- *Base classifier and optimizer:* For the base classifier, we have used 4-layer feed-forward network with 1st hidden layer contains 200 neurons 2nd hidden layer contains 100 and the 3rd hidden layer contains 50 neurons. The selection of layers is done by trial-error approach because there is approach to set it automatically. In hidden layers, ReLU is the activation function, and in the output layer, sigmoid is the activation function. In the output layer, sigmoid being the activation function is because we have considered binary classification (Benign or Malware). For training the neural network, we have used Stochastic Gradient Descent (SGD) optimizer because it is popularly used in federated approach.
- *Parameter setup:* The batch size of the DW-FedAvg is 32, and the epoch is 32. We have used 80–20 Hold-Out cross-validation technique to train and test the model and the overall performance comparison. The learning rate of SGD is 0.01.
- *Source-code:* The source code and implementation details of our proposed approach can be found in *Github*.¹

¹ The source code can be found in GitHub at: <https://github.com/officialarijit/DW-FedAvg>.

3.4 Performance Metric

The accuracy, $F1$ -score, area under the ROC curve (AUC), and False Positive Rate (FLR) are metrics for evaluating classification model. For binary classification, the mathematical formulae are as calculated in terms of positives and negatives: Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$, $F1$ -score = $\frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2} * (FP + FN)}$, AUC = $\frac{\Sigma Rank(+) - |+| * \frac{|+| + 1}{2}}{|+| + |-|}$ and FPR = $\frac{FP}{FP + TN}$, where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. $\Sigma Rank(+)$ is the sum the ranks of all the positively classifier samples, $|+|$ is total number of positive samples, and $|-|$ is the total negative samples present in the dataset.

4 Results, Analysis, and Discussion

In this section, we summarize the experimental results of our proposed DW-FedAvg and make a comparison with traditional FedAvg [8] under different number of clients and different number of rounds. For our experiment, we have considered a total of 3 different client scenario (5, 10, and 15). Also, two different rounds are considered such as 10 and 20. The average test average test accuracy, $F1$ -score, AUC, and FPR of the global model are presented in Tables 2 and 3. The better results are in bold in the table.

For the Malgenome dataset, our proposed DW-FedAvg approach has shown almost similar score or greater score in all aspect (accuracy, $F1$ -score, AUC score, and FPR score) compared to the traditional FedAvg approach. In the first comparison Table 2 (where no. of rounds is 10), there are 5 clients. Our proposed DW-FedAvg approach provides an increased accuracy and $F1$ -score of 0.02% and an increased AUC score of 0.05% and a decreased FPR score of 0.21% than the FedAvg approach for those first 5 clients. For 10 clients, it gives almost similar accuracy, $F1$ -score, and AUC score as that of FedAvg approach, whereas the FPR score has been increased to 0.93%. Likewise for 15 clients, it gives a slight decrease or almost similar accuracy, $F1$ -score and AUC score as that of FedAvg approach, whereas the FPR score has been increased to 0.48%. Similarly for round 20 Table 3, it is also tested among 5, 10, and 20 clients. For 5 clients, it gives almost similar accuracy, $F1$ -score, and AUC score as that of FedAvg approach, whereas the FPR score has been increased to 0.36%. For 10 clients, our proposed approach provides an increased accuracy of 0.17% and $F1$ -score of 0.23% and an increased FPR score of 0.1% than the FedAvg approach. For 15 clients, it gives a slight decrease or almost similar accuracy, $F1$ -score and AUC score as that of FedAvg approach, while the FPR score has increased to 0.23%.

Similarly for the Drebin dataset, our DW-FedAvg approach also provides almost similar score or greater score in all aspect compared to the traditional FedAvg approach. This dataset has also been tested through two rounds—10 and 20. In

Table 2 Global model's average test accuracy, $F1$ -score, AUC, and FPR comparison between FedAvg and DW-FedAvg for 10 rounds

Dataset	Number of clients	Accuracy	FedAvg			Dw FedAvg		
			$F1$ -score	AUC	FPR	Accuracy	$F1$ -score	AUC
Drebin	5	0.9826 \pm 0.001	0.9764 \pm 0.001	0.9955 \pm 0.0001	0.0277 \pm 0.002	0.9823\pm0.001	0.9766\pm0.001	0.996\pm0.0002
	10	0.9812 \pm 0.002	0.9744 \pm 0.003	0.9962 \pm 0.0003	0.0269 \pm 0.004	0.9784 \pm 0.003	0.9705 \pm 0.005	0.9956 \pm 0.0004
	15	0.9742 \pm 0.004	0.9648 \pm 0.005	0.9947 \pm 0.001	0.0392 \pm 0.005	0.9734 \pm 0.003	0.9636 \pm 0.005	0.9944 \pm 0.001
Malgenome	5	0.9911 \pm 0.0006	0.9871 \pm 0.0008	0.9998 \pm 0.0004	0.0050 \pm 0.003	0.9943\pm0.001	0.9917\pm0.001	0.9999\pm0.0003
	10	0.9901 \pm 0.003	0.9855 \pm 0.005	0.9994 \pm 0.0002	0.0135 \pm 0.011	0.9896 \pm 0.003	0.9847 \pm 0.004	0.9990 \pm 0.0006
	15	0.9838 \pm 0.001	0.9763 \pm 0.001	0.9988 \pm 0.0006	0.0189 \pm 0.005	0.9875\pm0.005	0.9814\pm0.008	0.9993\pm0.0003
Kronodroid	5	0.9632 \pm 0.003	0.9651 \pm 0.003	0.9898 \pm 0.001	0.0391 \pm 0.005	0.9596 \pm 0.004	0.9618 \pm 0.004	0.9888 \pm 0.001
	10	0.9548 \pm 0.003	0.9569 \pm 0.002	0.9870 \pm 0.001	0.0514 \pm 0.002	0.9516 \pm 0.003	0.9538 \pm 0.002	0.9855 \pm 0.001
	15	0.9498 \pm 0.004	0.9521 \pm 0.003	0.9856 \pm 0.001	0.0566 \pm 0.003	0.9478 \pm 0.004	0.9502 \pm 0.003	0.9844 \pm 0.002
Tuandromd	5	0.9880 \pm 0.002	0.9926 \pm 0.001	0.9988 \pm 0.0004	0.0064 \pm 0.002	0.9861 \pm 0.002	0.9914 \pm 0.001	0.9971 \pm 0.0003
	10	0.9840 \pm 0.004	0.9902 \pm 0.002	0.9983 \pm 0.0004	0.0133 \pm 0.004	0.9857\pm0.005	0.9912\pm0.003	0.9985\pm0.0009
	15	0.9799 \pm 0.008	0.9876 \pm 0.004	0.9977 \pm 0.001	0.0149 \pm 0.006	0.9780 \pm 0.003	0.9864 \pm 0.002	0.9967 \pm 0.001

Table 3 Global model's average accuracy, $F1$ -score, AUC, and FPR comparison between FedAvg and DW-FedAvg for 20 rounds

Dataset	Number of clients	Accuracy	FedAvg			DW FedAvg		
			$F1$ -score	AUC	FPR	Accuracy	$F1$ -score	AUC
Drebin	5	0.9848±0.001	0.9793±0.001	0.9962±0.0009	0.0232±0.001	0.9841±0.002	0.9783±0.003	0.9961±0.0002
	10	0.9808±0.001	0.9740±0.002	0.9957±0.003	0.0258±0.004	0.9825±0.003	0.9763±0.004	0.9957±0.0003
	15	0.9794±0.002	0.9720±0.003	0.9958±0.0003	0.0314±0.002	0.9780±0.005	0.9701±0.006	0.9946±0.0005
Malgenome	5	0.9962±0.001	0.9944±0.002	0.9997±0.0002	0.0054±0.003	0.9923±0.001	0.9887±0.002	0.9998±0.0001
	10	0.9892±0.001	0.9842±0.002	0.9994±0.0002	0.0090±0.006	0.9940±0.003	0.9912±0.005	0.9998±0.0001
	15	0.9921±0.002	0.9884±0.003	0.9997±0.001	0.0088±0.008	0.9897±0.004	0.9849±0.006	0.9992±0.0007
Kronodroid	5	0.9683±0.002	0.9700±0.002	0.9909±0.0007	0.0321±0.004	0.9661±0.003	0.9680±0.003	0.9904±0.001
	10	0.9623±0.002	0.9644±0.002	0.9902±0.001	0.0372±0.003	0.9622±0.003	0.9643±0.003	0.9896±0.001
	15	0.9607±0.004	0.9627±0.004	0.9884±0.001	0.0405±0.007	0.9593±0.005	0.9619±0.005	0.9881±0.001
Tuandromd	5	0.9893±0.001	0.9934±0.001	0.9992±0.0002	0.0050±0.001	0.9870±0.003	0.9920±0.002	0.9975±0.0005
	10	0.9843±0.002	0.9904±0.001	0.9987±0.0003	0.0100±0.002	0.9867±0.005	0.9918±0.003	0.9989±0.0006
	15	0.9839±0.006	0.9901±0.003	0.9960±0.0008	0.0134±0.005	0.9832±0.005	0.9896±0.003	0.9981±0.0008

round 10 Table 2, it is tested among 5, 10, and 20 clients. For 5 clients, the model provides a significant increase of accuracy of 0.32%, $F1$ -score of 0.46%, and AUC score of 0.01% as that of FedAvg approach. For 10 clients, our proposed approach gives almost similar accuracy, $F1$ -score, and AUC score as that of FedAvg approach, whereas the FPR score has increased to 0.19%. For 15 clients, our DW-FedAvg approach provides an increased accuracy of 0.37%, $F1$ -score of 0.51%, AUC score of 0.05%, and an FPR score of 0.42% than the FedAvg approach. Similarly for round 20 Table 3, it is tested among 5, 10, and 20 clients. For 10 clients, a significant increase of accuracy score of 0.48%, $F1$ -score of 0.7%, and AUC score of 0.04% is observed in our approach than the FedAvg approach. For 5 and 15 clients, it gives a slight decrease or almost same accuracy, $F1$ -score, AUC score, and FPR score as that of FedAvg approach.

For the Kronodroid dataset, our proposed DW-FedAvg approach provides almost similar score or greater score as well in all aspect, i.e., accuracy, $F1$ -score, AUC score, and FPR score, in comparison with the traditional FedAvg approach. Likewise, the other datasets, this dataset has also been tested through two rounds—10 and 20. In both the rounds, it is tested among 5, 10, and 20 clients. In round 10 Table 2, for 5 and 10 clients, our proposed approach gives a slight decrease or almost similar accuracy, $F1$ -score, AUC score, and FPR score as that of FedAvg approach. For 15 clients, our proposed approach gives almost similar accuracy, $F1$ -score, and the AUC score, whereas there is a significant increase of the FPR score of 0.31% than the FedAvg approach. In round 20 Table 3, for 20 clients, our proposed approach gives almost similar accuracy, $F1$ -score, AUC score, and FPR score as that of FedAvg approach. For 5 and 10 clients, the accuracy, $F1$ -score, and the AUC score of our approach almost remains same as that of FedAvg approach, while the FPR score has been increased to 0.1% and 0.13%, respectively, than the FedAvg approach. Likewise, for Tuandromd dataset, our DW-FedAvg approach also provides almost similar score or greater score in all aspect in compare to the traditional FedAvg approach. This dataset has also been tested through two rounds—10 and 20. In round 10 Table 2, it is tested among 5, 10, and 20 clients. For 5 and 15 clients, our approach provides almost similar accuracy, $F1$ -score, AUC score, and FPR score to that of FedAvg approach. For 10 clients, there is a significant increase of accuracy by 0.17%, $F1$ -score by 0.1%, and AUC score by 0.02% and a slight decrease of FPR score by 0.07%. In round 20 Table 3, for 5 clients, our approach provides almost similar accuracy, $F1$ -score, AUC score, and FPR score to that of FedAvg approach. For 10 clients, our DW-FedAvg approach provides a significant increase of the accuracy by 0.24%, $F1$ -score by 0.14% and AUC score by 0.02% and a slight decrease of FPR score by 0.07% than the FedAvg approach. For 20 clients, our approach provides an increase in AUC score by 0.21%, whereas the accuracy, $F1$ -score, and the FPR score remain almost same as that of FedAvg approach.

The FedAvg approach takes the average of the local models. Our DW-FedAvg approach adjusts the weight based on the model performance. Our approach rewards best performing models, whereas penalizes the poor performing models. In case of too many local models with poor classifiers, our DW-FedAvg approach dynamically penalizes weights based on their performance, and thus minimum number of models

are getting rewarded. Therefore, the dynamic average overall decreases the accuracy of the model. On the other hand, traditional FedAvg performs average on the local models and does not dynamically adjust the weights of the local models, thus it does not have any effect on accuracy degradation.

In traditional FedAvg approach, simple averaging of the local models is done, and it results into a global model. Thus, equal priority is given to all the models. On the other hand, our DW-FedAvg approach dynamically adjusts the weight of the local models based on their performance. Our approach penalizes the poor performing classifiers, whereas rewards the best performing classifiers. By doing the simple average can create a bad global model for the fact that if a global model performs poorly. Our approach prioritizes the best performing models. As more priority is given to better performing models, the global model gives a better result in the accuracy of our approach.

Finally, from the detailed comparison shows that our approach has outperformed traditional FedAvg for some clients in both the rounds for both the datasets. It also shows the advantage of dynamically adjusting the weights based on the best performing and poor performing models for both the rounds.

5 Conclusion

In this paper, we have proposed a dynamic weighted federated averaging (DW-FedAvg) approach where the weights for each local model are adjusted automatically based on their performance at the client to create a powerful global model in federated learning-based android malware classification. Our proposed DW-FedAvg gives a reward of dynamic weightage to the best performing model and subtracts the dynamic weightage from the poor performing local models. The effectiveness of our proposed approach is evaluated using four benchmark datasets Drebin, Malgenome, Kronodroid, and Tuandromd. The results show that our proposed approach has outperformed the traditional FedAvg under different number of clients and different rounds while classifying android malware's.

As a future work, we have plans to work with semi-supervised approach for classifying android malwares. We want to integrate our approach on different machine learning algorithms other than CNN such as SVM to create a new approach called Federated SVM. Also, deploying this proposed system into real scenario and examine how will it perform.

References

1. Abualola H, Alhwai H, Kadadha M, Otrok H, Mourad A (2016) An android-based trojan spyware to study the notification listener service vulnerability. In: The 7th international conference on ambient systems, networks and technologies (ANT 2016)/the 6th international conference on sustainable energy information technology (SEIT-2016)/affiliated workshops. *Proc Comput Sci* 83:465–471
2. Allix K, Jerome Q, Bissyandé TF, Klein J, State R, Traon YL (2014) A forensic analysis of android malware—how is malware written and how it could be detected? In: 2014 IEEE 38th annual computer software and applications conference, pp 384–393
3. Borah P, Bhattacharyya D, Kalita J (2020) Malware dataset generation and evaluation. In: 2020 IEEE 4th conference on information and communication technology (CICT). IEEE, pp 1–6
4. D'Angelo G, Palmieri F, Robustelli A (2022) A federated approach to android malware classification through perm-maps. *Cluster Comput* 25:2487–2500
5. Ghimire B, Rawat DB (2022) Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. *IEEE Internet Things J* 9(11):8229–8249
6. Guerra-Manzanares A, Bahsi H, Nömm S (2021) KronoDroid: time-based hybrid-featured dataset for effective android malware detection and characterization. *Comput Secur* 110:102399
7. Gálvez R, Moonsamy V, Diaz C (2021) Less is more: a privacy-respecting android malware classifier using federated learning. *Proc Priv Enh Technol* 4:96–116. <https://doi.org/10.2478/popets-2021-0062>
8. Konečný J, McMahan HB, Yu FX, Richtarik P, Suresh AT, Bacon D (2016) Federated learning: strategies for improving communication efficiency. In: NIPS workshop on private multi-party machine learning. <https://arxiv.org/abs/1610.05492>
9. Taheri R, Shojafar M, Alazab M, Tafazolli R (2021) FED-IIoT: a robust federated malware detection architecture in industrial IoT. *IEEE Trans Ind Inf* 17(12):8442–8452
10. Yerima SY, Sezer S (2019) Droidfusion: a novel multilevel classifier fusion approach for android malware detection. *IEEE Trans Cybernet* 49(2):453–466
11. Zhang W, Wang X, Zhou P, Wu W, Zhang X (2021) Client selection for federated learning with non-IID data in mobile edge computing. *IEEE Access* 9:24462–24474
12. Zhang Y, Jiang C, Yue B, Wan J, Guizani M (2022) Information fusion for edge intelligence: a survey. *Inf Fusion* 81:171–186
13. Zhou Y, Jiang X (2012) Dissecting android malware: characterization and evolution. In: 2012 IEEE symposium on security and privacy, pp 95–109

Entropy Measure for the Linguistic q -Rung Orthopair Fuzzy Set



Neelam, Kamal Kumar, and Reeta Bhardwaj

Abstract A linguistic q -rung orthopair fuzzy set (Lq-ROFS) is a valuable tool for conveying the complexity of any qualitative data set, as well as a generalization of the linguistic intuitionistic fuzzy set (LIFS) and the linguistic Pythagorean fuzzy set (LPFS). The primary goal of this research is to develop a novel technique for measuring the uncertainty of Lq-ROFS. We propose an entropy measure for Lq-ROFSs to do this. Several desirable features and criteria of the proposed entropy measure are also explored in order to validate it. Finally, a few numerical examples are used to show how the proposed entropy measure of Lq-ROFSs is better than the ones that are already in use.

Keywords Fuzzy set · Entropy measure · Decision-making · Linguistic fuzzy set · Uncertainty · Qualitative data

1 Introduction

In numerous application domains, decision-making theories have been extensively employed. The depiction of ambiguous or uncertain information is among the tricky problems in real-world decision-making. Fuzzy set (FS) theory was first introduced by [1] in 1965 thereafter, a number of generalization are proposed by the researchers. In 1986, Atanassov [2] proposed the intuitionistic FS (IFS). In 2013, Yager [3] proposed an another extension of FS theory which is known as pythagorean FS (PFS) theory. In order to provide the larger space for decision makers, Yager [4] also introduced the q -rung orthopair FS (q-ROFS) theory. These theories applied by the various researchers [5–9] in different fields. The IFSs, PFSs, and q-ROFSs describe only quantitative information. Sometimes decision makers can not express their pref-

Neelam · K. Kumar · R. Bhardwaj (✉)

Department of Mathematics, Amity School of Applied Sciences, Amity University Haryana, Gurugram, India

e-mail: bhardwajreeta84@gmail.com

K. Kumar
e-mail: kamalkumarrajput92@gmail.com

erences in quantitative information. However, decision-maker(s) can described their assessments in the form of qualitative aspects. For example, instead of numerical values, linguistic terms such as “good”, “average”, and “bad” are commonly used to rate the quality of food. Firstly, Zadeh [10] introduced the concept of linguistic variable (LV). Afterwards, various researcher [11, 12] applied the linguistic approach in decision theories. Xu [13] defined the weighted geometric and weighted averaging aggregation operators (AOs) for LVs. Xu et al. [14] proposed linguistic power AOs with application in MADM issues. After that, Zhang [15] defined linguistic IFS (LIFS) for expressing the qualitative assessment more easily in which membership grade (MG) and non-membership grade (NMG) are in LVs. Chen et al. [16] proposed AOs for aggregating the linguistic intuitionistic fuzzy numbers (LIFNs). Peng et al. [17] defined Heronian mean AOs based on Frank operations. Liu and Wang [18] proposed some enhanced operational laws for LIFNs and AOs based on it. Garg and Kumar [19] proposed AOs for aggregating linguistic connection number and DM method under the LIFNs environment based on the set pair analysis. Li et al. [20] proposed the entropy measure for LIFSs. Kumar et al. [21] introduced the entropy measure for LIFSs and decision-making method based on it. Kumar and Chen [22] defined the distance measure for LIFSs. Kumar and Chen [23] defined the advanced AOs for LIFNs.

Recently, Liu and Liu [24] defined a new extension of the fuzzy set known as linguistic q-ROFS (Lq-ROFS). The LIFS is the particular case of the Lq-ROFS. Verma [25] defined the generalized similarity measure for Lq-ROFSs. Liu and Liu [26] defined the power Muirhead mean operators and entropy measure for Lq-ROFSs. Akram et al. [27] defined a group decision-making method for Lq-ROFSs environment. Lin et al. [28] defined the Heronian mean aggregation operator for Lq-ROFSs environment. Bao and Shi [29] introduced the ELECTRE method-based group decision-making under the Lq-ROFSs context.

However, during mathematical verification, we discovered certain shortcomings in existing Lq-ROFS entropy measure (EM). To circumvent these limitations, we must create a new EM for measuring the uncertainty of Lq-ROFS. We propose a new EM for the Lq-ROFS in this study. To validate it, we specify the proof of several desirable qualities and the validity condition of the suggested EM of Lq-ROFS. The proposed entropy measure can solve the shortcomings of the existing Lq-ROFS entropy measure. The proposed entropy measure is simple and effective for calculating the uncertainty of the Lq-ROFS.

To achieve the aforementioned goals, this paper is organized as follows: In Sect. 2, brief introduction of basic concepts related to this paper is given. In Sect. 3, we have developed an entropy measure for Lq-ROFSs. In Sect. 4, a few advantages of the proposed entropy measure are given. Finally, Sect. 5 concludes the paper.

2 Preliminaries

Definition 1 ([11]) Let us consider a finite odd cardinality linguistic term set (LTS) $S = \{s_t \mid t = 0, 1, 2, \dots, h\}$, where s_t reflects a suitable value for a linguistic variable (LV). For example, while assessing a school's "location", we can consider five linguistic term (LT) as $s_0 = \text{"None"}$, $s_1 = \text{"very bad"}$, $s_2 = \text{"bad"}$, $s_3 = \text{"good"}$ and $s_4 = \text{"very good"}$.

The LTs of LTS S satisfies the following conditions: [11].

- (i) $s_k \leq s_t \Leftrightarrow k \leq t$
- (ii) $\text{Neg}(s_k) = s_{h-k}$
- (iii) $\max(s_k, s_t) = s_k \Leftrightarrow s_k \geq s_t$
- (iv) $\min(s_k, s_t) = s_t \Leftrightarrow s_k \geq s_t$.

Later on, discrete LTS S extended to continuous LTS (CLTS) by Xu [13] as:

$$S_{[0,h]} = \left\{ s_z \mid s_0 \leq s_z \leq s_h \right\},$$

Definition 2 ([24]) A linguistic q -rung orthopair fuzzy set (Lq-ROFS) Q in finite universal set U is defined as:

$$Q = \{ \langle u, s_{\xi(u)}, s_{\varphi(u)} \rangle \mid u \in U \}$$

where $s_{\xi(u)}$, $s_{\varphi(u)}$ indicate the linguistic membership degree (LMG) and linguistic non-membership degree (LNMG) of u to Q respectively. For any $u \in U$, the conditions $s_{\xi(u)}, s_{\varphi(u)} \in S_{[0,h]}$ and $0 \leq \xi(u)^q + \varphi(u)^q \leq h^q$ holds, and in turn, the hesitance of u to Q is defined as $s_{\pi(u)} = s_{(h^q - \xi(u)^q - \varphi(u)^q)^{1/q}}$ where $q \geq 1$.

Definition 3 For a Lq-ROFS $Q = \{ \langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U \}$ and a real $\lambda > 0$, Liu and Liu [24] defined the the Lq-ROFS Q^λ as:

$$Q^\lambda = \left\{ \left\langle u_i, s_{h\left(\frac{\xi(u_i)}{h}\right)^\lambda}, s_{\left(h^q - h^q\left(1 - \frac{\varphi(u_i)^q}{h^q}\right)^\lambda\right)^{1/q}} \right\rangle \mid u_i \in U \right\} \quad (1)$$

Definition 4 ([26]) If $Q = \{ \langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U \}$ be any Lq-ROFS, then a valid entropy measure $E(Q)$ of the Lq-ROFS Q must meet the following characteristics:

- P(1) $E(Q) = 0 \Leftrightarrow Q$ is a linguistic set.
- P(2) $E(Q) = 1 \Leftrightarrow \xi(u_i) = \varphi(u_i), \forall u_i \in U$.
- P(3) $E(Q) = E(Q^c)$.
- P(4) $E(Q_1) \leq E(Q_2)$ if Q_1 is less fuzzy than Q_2 , i.e., $\xi_1(u_i) \leq \xi_2(u_i)$, $\varphi_2(u_i) \leq \varphi_1(u_i)$ for $\xi_2(u_i) \leq \varphi_2(u_i)$ or $\xi_1(u_i) \geq \xi_2(u_i)$, $\varphi_2(u_i) \geq \varphi_1(u_i)$ for $\xi_2(u_i) \geq \varphi_2(u_i) \forall u_i \in U$.

Definition 5 For a Lq-ROFS $Q = \{\langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U\}$, Liu and Liu [26] defined the entropy measure (EM) E_1 as:

$$E_1(Q) = \frac{1}{n} \sum_{i=1}^n \frac{\min\{(\xi(u_i))^q, (\varphi(u_i))^q\} + (\pi_i)^q}{\max\{(\xi(u_i))^q, (\varphi(u_i))^q\} + (\pi_i)^q} \quad (2)$$

3 Proposed Entropy Measure of Linguistic q -rung Orthopair Fuzzy Set

In this section, we propose an entropy measure for linguistic q -rung orthopair fuzzy sets (Lq-ROFSs) and let $\zeta_{[0,h]}$ be the collection of the Lq-ROFSs.

Definition 6 Let a Lq-ROFS $Q = \{\langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U\} \in \zeta_{[0,h]}$, then the proposed entropy measure $E(Q)$ for Lq-ROFS is defined as:

$$E(Q) = \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] \quad (3)$$

where $q \geq 1$.

Theorem 1 The entropy measure for Lq-ROFS $Q = \{\langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U\} \in \zeta_{[0,h]}$ is defined in Eq. (3) satisfies the characteristics given in Definition 4.

Proof Let $Q = \{\langle u_i, s_{\xi(u_i)}, s_{\varphi(u_i)} \rangle \mid u_i \in U\} \in \zeta_{[0,h]}$ be a Lq-ROFS then we have

(P1) We have $E(Q) = 0$

$$\begin{aligned} &\Leftrightarrow \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] = 0 \\ &\Leftrightarrow h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) = 0 \\ &\Leftrightarrow h^{2q} - |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) = 0 \\ &\Leftrightarrow (\xi(u_i))^q = h^q, (\varphi(u_i))^q = 0 \quad \text{or} \quad (\xi(u_i))^q = 0, (\varphi(u_i))^q = h^q, \quad \forall u_i \in U. \end{aligned}$$

(P2) We have $E(Q) = 1$

$$\begin{aligned}
&\Leftrightarrow \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] = 1 \\
&\Leftrightarrow h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) = h^q \\
&\Leftrightarrow \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) = 0 \\
&\Leftrightarrow |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) = 0 \\
&\Leftrightarrow (\xi(u_i))^q = (\varphi(u_i))^q, \quad \forall u_i \in U.
\end{aligned}$$

(P3) $Q^c = \{\langle u, s_\varphi(u), s_\xi(u) \rangle \mid u \in U\}$. We have

$$\begin{aligned}
E(Q) &= \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] \\
&= \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\varphi(u_i))^q - (\xi(u_i))^q| ((\varphi(u_i))^q + (\xi(u_i))^q) \right] \\
&= E(Q^c).
\end{aligned}$$

(P4) Consider the function $f(u, v) = [h^q - \frac{1}{h^q} |u^q - v^q| (u^q + v^q)]$, where, $u, v \in [0, h]$ and $u^q + v^q \leq h^q$. We must demonstrate that when $u \leq v$, the function $f(u, v)$ increases with respect to u and decreases with respect to v . The partial derivatives of $f(u, v)$ with respect to u and v can be defined as:

$$\begin{aligned}
\frac{\partial f(u, v)}{\partial u} &= [h^q - \frac{1}{h^q} |u^q - v^q| (u^q + v^q)] \\
&= -\frac{1}{h^q} [qu^{q-1} (|u^q - v^q| + (u^q + v^q))] \\
\frac{\partial f(u, v)}{\partial v} &= [h^q - \frac{1}{h^q} |u^q - v^q| (u^q + v^q)] \\
&= -\frac{1}{h^q} [qv^{q-1} (|u^q - v^q| - (u^q + v^q))].
\end{aligned}$$

Since $\frac{\partial f(u, v)}{\partial u} \geq 0$ and $\frac{\partial f(u, v)}{\partial v} \leq 0$ for $u \leq v$. Thus, the function $f(u, v)$ is increasing with respect to u and decreasing with respect to v for $u \leq v$. Thus, $f(\xi_1(u_i), \varphi_1(u_i)) \leq f(\xi_2(u_i), \varphi_2(u_i))$ when $\xi_2(u_i) \leq \varphi_2(u_i)$ and $\xi_1(u_i) \leq \xi_2(u_i), \varphi_1(u_i) \geq \varphi_2(u_i)$.

Similarly, $\frac{\partial f(u, v)}{\partial u} \leq 0$ and $\frac{\partial f(u, v)}{\partial v} \geq 0$ for $u \geq v$. Thus, the function $f(u, v)$ is decreasing with respect to u and increasing with respect to v for $u \geq v$. Thus, $f(\xi_1(u_i), \varphi_1(u_i)) \leq f(\xi_2(u_i), \varphi_2(u_i))$ when $\xi_2(u_i) \geq \varphi_2(u_i)$ and $\xi_1(u_i) \geq \xi_2(u_i), \varphi_1(u_i) \leq \varphi_2(u_i)$.

Therefore, if Q_1 is less fuzzy compare to Q_2 then $\frac{1}{n} \sum_{i=1}^n f(\xi_1(u_i), \varphi_1(u_i)) \leq \frac{1}{n} \sum_{i=1}^n f(\xi_2(u_i), \varphi_2(u_i))$. Hence $E(Q_1) \leq E(Q_2)$.

Theorem 2 Let $Q_1 = \{\langle u, S_{\xi_1}(u_i), S_{\varphi_1}(u_i) \rangle \mid u_i \in U\}$ and $Q_2 = \{\langle u, S_{\xi_2}(u_i), S_{\varphi_2}(u_i) \rangle \mid u_i \in U\}$ be two Lq -ROFS, such that either $Q_1 \subseteq Q_2$ or $Q_1 \supseteq Q_2 \forall u \in U$, then

$$E(Q_1 \cup Q_2) + E(Q_1 \cap Q_2) = E(Q_1) + E(Q_2)$$

Proof Consider U_1 and U_2 are the two subsets of U , such that $U_1 = \{u_i \in U \mid Q_1 \subseteq Q_2\}$ and $U_2 = \{u_i \in U \mid Q_1 \supseteq Q_2\}$, i.e., $\forall u_i \in U_1$, we have $\xi_1(u_i) \leq \xi_2(u_i)$, $\varphi_1(u_i) \geq \varphi_2(u_i)$ and $\forall u_i \in U_2$, $\xi_1(u_i) \geq \xi_2(u_i)$, $\varphi_1(u_i) \leq \varphi_2(u_i)$.

Therefore, according to the proposed entropy measure, we have

$$\begin{aligned} E(Q_1 \cup Q_2) &= \frac{1}{nh^q} \sum_{u_i \in U} \left[h^q - \frac{1}{h^q} |(\xi_{Q_1 \cup Q_2}(u_i))^q - (\varphi_{Q_1 \cup Q_2}(u_i))^q| ((\xi_{Q_1 \cup Q_2}(u_i))^q + (\varphi_{Q_1 \cup Q_2}(u_i))^q) \right. \\ &\quad \left. + (\varphi_{Q_1 \cup Q_2}(u_i))^q \right] \\ &= \frac{1}{nh^q} \left[\sum_{u_i \in U_1} \left(h^q - \frac{1}{h^q} |(\xi_{Q_2}(u_i))^q - (\varphi_{Q_2}(u_i))^q| ((\xi_{Q_2}(u_i))^q + (\varphi_{Q_2}(u_i))^q) \right) \right. \\ &\quad \left. + \sum_{u_i \in U_2} \left(h^q - \frac{1}{h^q} |(\xi_{Q_1}(u_i))^q - (\varphi_{Q_1}(u_i))^q| ((\xi_{Q_1}(u_i))^q + (\varphi_{Q_1}(u_i))^q) \right) \right]. \end{aligned}$$

Similarly,

$$\begin{aligned} E(Q_1 \cap Q_2) &= \frac{1}{nh^q} \left[\sum_{u_i \in U_1} \left(h^q - \frac{1}{h^q} |(\xi_{Q_1}(u_i))^q - (\varphi_{Q_1}(u_i))^q| ((\xi_{Q_1}(u_i))^q + (\varphi_{Q_1}(u_i))^q) \right) \right. \\ &\quad \left. + \sum_{u_i \in U_2} \left(h^q - \frac{1}{h^q} |(\xi_{Q_2}(u_i))^q - (\varphi_{Q_2}(u_i))^q| ((\xi_{Q_2}(u_i))^q + (\varphi_{Q_2}(u_i))^q) \right) \right]. \\ E(Q_1 \cup Q_2) + E(Q_1 \cap Q_2) &= \frac{1}{nh^q} \left[\sum_{u_i \in U_1} \left(h^q - \frac{1}{h^q} |(\xi_{Q_1}(u_i))^q - (\varphi_{Q_1}(u_i))^q| ((\xi_{Q_1}(u_i))^q + (\varphi_{Q_1}(u_i))^q) \right) \right. \\ &\quad \left. + \sum_{u_i \in U_2} \left(h^q - \frac{1}{h^q} |(\xi_{Q_1}(u_i))^q - (\varphi_{Q_1}(u_i))^q| ((\xi_{Q_1}(u_i))^q + (\varphi_{Q_1}(u_i))^q) \right) \right] \\ &\quad + \frac{1}{nh^q} \left[\sum_{u_i \in U_1} \left(h^q - \frac{1}{h^q} |(\xi_{Q_2}(u_i))^q - (\varphi_{Q_2}(u_i))^q| ((\xi_{Q_2}(u_i))^q + (\varphi_{Q_2}(u_i))^q) \right) \right. \\ &\quad \left. + \sum_{u_i \in U_2} \left(h^q - \frac{1}{h^q} |(\xi_{Q_2}(u_i))^q - (\varphi_{Q_2}(u_i))^q| ((\xi_{Q_2}(u_i))^q + (\varphi_{Q_2}(u_i))^q) \right) \right] \\ &= E(Q_1) + E(Q_2). \end{aligned}$$

Example 1 Let a Lq-ROFS $Q = \{(u_1, s_1, s_7), (u_2, s_3, s_1), (u_3, s_2, s_0)\} \in \zeta_{[0,h]}$. By utilizing Eq. (3), we conclude the proposed entropy measure $E(Q)$ of the Lq-ROFS Q for $q = 3$ as follows:

$$\begin{aligned}
 E(Q) &= \frac{1}{nh^q} \sum_{i=1}^n \left[h^q - \frac{1}{h^q} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] \\
 &= \frac{1}{3 \times 8^3} \sum_{i=1}^3 \left[8^3 - \frac{1}{8^3} |(\xi(u_i))^q - (\varphi(u_i))^q| ((\xi(u_i))^q + (\varphi(u_i))^q) \right] \\
 &= \frac{1}{(3^*512)} \left[\left(8^3 - \frac{1}{8^3} [|1^3 - 7^3| \times (1^3 + 7^3)] \right) \right. \\
 &\quad + \left(8^3 - \frac{1}{8^3} [|3^3 - 1^3| \times (3^3 + 1^3)] \right) \\
 &\quad \left. + \left(8^3 - \frac{1}{8^3} [|2^3 - 0^3| \times (2^3 + 0^3)] \right) \right] \\
 &= 0.8493
 \end{aligned}$$

4 Advantages of the Proposed Entropy Measure of Lq-ROFSs

Example 2 Let a Lq-ROFS Q “LARGE” over the universal set $U = \{u_1, u_2, u_3, u_4, u_5\}$ as follows:

$$Q = \{(u_1, s_1, s_7), (u_2, s_3, s_1), (u_3, s_2, s_0), (u_4, s_5, s_1), (u_5, s_8, s_0)\} \in \zeta_{[0,8]}.$$

By using the Eq. (1), we generate the following Lq-ROFS:

- (a) $Q^{(1/2)}$ can be viewed as “Less LARGE”;
- (b) Q can be viewed as “LARGE”;
- (c) Q^2 can be viewed as “Very LARGE”;
- (d) Q^3 can be viewed as “Quite very LARGE”;
- (e) Q^4 can be viewed as “Very very LARGE”,

where

$$\begin{aligned}
 Q^{1/2} &= \{(u_1, s_{2.8284}, s_{6.0170}), (u_2, s_{4.8990}, s_{0.7938}), (u_3, s_{4.0000}, s_0), (u_4, s_{6.3246}, s_{0.7938}), \\
 &\quad (u_5, s_{8.0000}, s_0)\} \\
 Q &= \{(u_1, s_{1.0000}, s_{7.0000}), (u_2, s_{3.0000}, s_{1.0000}), (u_3, s_{2.0000}, s_0), (u_5, s_{5.0000}, s_{1.0000}), \\
 &\quad (u_5, s_{8.0000}, s_0)\} \\
 Q^2 &= \{(u_1, s_{0.1250}, s_{7.6982}), (u_2, s_{1.1250}, s_{1.2595}), (u_3, s_{0.5000}, s_0), (u_4, s_{3.1250}, s_{1.2595}), \\
 &\quad (u_5, s_{8.0000}, s_0)\}
 \end{aligned}$$

Table 1 The value of EMs $E_1(\cdot)$ and $E(\cdot)$ for the Lq-ROFSs $Q^{1/2}$, Q , Q^2 , Q^3 and Q^4 given in Example 2

	E_1	E
$Q^{1/2}$	0.5507	0.7017
Q	0.6043	0.6977
Q^2	0.6103	0.6405
Q^3	0.6043	0.6141
Q^4	0.6000	0.6047

$$Q^3 = \{(u_1, s_{0.0156}, s_{7.9029}), (u_2, s_{0.4219}, s_{1.4413}), (u_3, s_{0.1250}, s_0), (u_4, s_{1.9531}, s_{1.4413}), (u_5, s_{8.0000}, s_0)\}$$

$$Q^4 = \{(u_1, s_{0.0020}, s_{7.9682}), (u_2, s_{0.1582}, s_{1.5859}), (u_3, s_{0.0312}, s_0), (u_4, s_{1.2207}, s_{1.5859}), (u_5, s_{8.0000}, s_0)\}.$$

Now, we calculate the Liu and Liu's EM (E_1) and proposed EM (E) given in Eqs. (2) and (3), respectively, for the the above Lq-ROFS $Q^{1/2}$, Q , Q^2 , Q^3 and Q^4 for $q = 3$, and obtained results are summarized in Table 1.

From the perspective of mathematical operations, the EMs (E_1) and (E) for the Lq-ROFSs $Q^{1/2}$, Q , Q^2 , Q^3 and Q^4 must satisfy the following condition [21, 30]:

$$E(Q^{1/2}) > E(Q) > E(Q^2) > E(Q^3) > E(Q^4). \quad (4)$$

From Table 1, we see that

$$E_1(Q^2) > E_1(Q) = E_1(Q^3) > E_1(Q^4) > E_1(Q^{1/2}),$$

$$E(Q^{1/2}) > E(Q) > E(Q^2) > E(Q^3) > E(Q^4).$$

Liu and Liu's EM (E_1) does not satisfy the relation given in Eq. (4). While, the proposed EM (E) given in Eq. (3) satisfies the relation given in Eq. (4). Hence, the performance of the proposed EM (E) is better than the performance of Liu and Liu's EM (E_1).

Example 3 Let a Lq-ROFS Q "LARGE" over the universal set $U = \{u_1, u_2, u_3, u_4, u_5\}$ as follows:

$$Q = \{\langle u_1, s_1, s_7 \rangle, \langle u_2, s_4, s_1 \rangle, \langle u_3, s_2, s_6 \rangle, \langle u_4, s_5, s_2 \rangle, \langle u_5, s_3, s_3 \rangle\} \in \zeta_{[0,8]}. \quad (5)$$

Now, we calculate the Liu and Liu's EM (E_1) and proposed EM (E) given in Eqs. (2) and (3), respectively, for the the above Lq-ROFS $Q^{1/2}$, Q , Q^2 , Q^3 and Q^4 for $q = 3$, and obtained results are summarized in Table 2.

From Table 2, we see that

Table 2 The value of EMs $E_1(\cdot)$ and $E(\cdot)$ for the Lq-ROFSs $Q^{1/2}$, Q , Q^2 , Q^3 and Q^4 given in Example 3

	E_1	E
$Q^{1/2}$	0.6837	0.8716
Q	0.7125	0.8597
Q^2	0.6604	0.7499
Q^3	0.6087	0.6791
Q^4	0.5727	0.6386

$$E_1(Q) > E_1(Q^{1/2}) > E_1(Q^2) > E_1(Q^3) > E_1(Q^4),$$

$$E(Q^{1/2}) > E(Q) > E(Q^2) > E(Q^3) > E(Q^4).$$

Liu and Liu's EM (E_1) does not satisfy the relation given in Eq. (4). While, the proposed EM (E) given in Eq. (3) satisfies the relation given in Eq. (4). Hence, the performance of the proposed EM (E) is better than the performance of Liu and Liu's EM (E_1).

5 Conclusion

Entropy measure (EM) is a powerful technique to measure the uncertainty of any data set. In this paper, we propose an entropy measure for Lq-ROFS. The proposed EM assists decision makers in determining the uncertainty of the Lq-ROFS. Some of the proposed entropy measure's properties have also been discussed. The proposed EM is illustrated with a few numerical examples to validate it. The proposed EM of Lq-ROFS can address the drawbacks of the existing EMs of Lq-ROFSs. From numerical example's results, it has been concluded that the proposed EM is sensible and practicable, and give an easy way to measure the uncertainty of Lq-ROFS. In future, we will use proposed EM to develop the decision-making under the Lq-ROFSs environment.

References

1. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
2. Atanassov KT (1986) Intuitionistic fuzzy sets. Fuzzy Sets Syst 20(1):87–96
3. Yager RR (2013) Pythagorean membership grades in multicriteria decision making. IEEE Trans Fuzzy Syst 22(4):958–965
4. Yager RR (2017) Generalized orthopair fuzzy sets. IEEE Trans Fuzzy Syst 25(5):1222–1230
5. Koundal D, Sharma B, Gandotra E (2017) Spatial intuitionistic fuzzy set based image segmentation. Imag Med 9(4):95–101

6. Kumar K, Chen SM (2022) Group decision making based on q -rung orthopair fuzzy weighted averaging aggregation operator of q -rung orthopair fuzzy numbers. *Inf Sci* 598:1–18
7. Bhalla K, Koundal D, Sharma B, Hu YC, Zagaria A (2022) A fuzzy convolutional neural network for enhancing multi-focus image fusion. *J Vis Commun Image Represent* 84:103485
8. Koundal D, Sharma B, Guo Y (2020) Intuitionistic based segmentation of thyroid nodules in ultrasound images. *Comput Biol Med* 121:103776
9. Dhankhar C, Kumar K (2022) Multi-attribute decision-making based on the advanced possibility degree measure of intuitionistic fuzzy numbers. *Granular Comput* 1–12. <https://doi.org/10.1007/s41066-022-00343-0>
10. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning—I. *Inf Sci* 8(3):199–249
11. Herrera F, Martínez L (2001) A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making. *IEEE Trans Syst Man Cybernet Part B (Cybernet)* 31(2):227–234
12. Dong Y, Xu Y, Li H, Feng B (2010) The OWA-based consensus operator under linguistic representation models using position indexes. *Eur J Oper Res* 203(2):455–463
13. Xu Z (2004) A method based on linguistic aggregation operators for group decision making with linguistic preference relations. *Inf Sci* 166(1):19–30
14. Xu Y, Merigó JM, Wang H (2012) Linguistic power aggregation operators and their application to multiple attribute group decision making. *Appl Math Model* 36(11):5427–5444
15. Zhang H (2014) Linguistic intuitionistic fuzzy sets and application in MAGDM. *J Appl Math* 2014:11. Article ID 432,092
16. Chen Z, Liu P, Pei Z (2015) An approach to multiple attribute group decision making based on linguistic intuitionistic fuzzy numbers. *Int J Comput Intell Syst* 8(4):747–760
17. Peng H, Wang J, Cheng P (2018) A linguistic intuitionistic multi-criteria decision-making method based on the Frank Heronian mean operator and its application in evaluating coal mine safety. *Int J Mach Learn Cybern* 9:1053–1068
18. Liu P, Wang P (2017) Some improved linguistic intuitionistic fuzzy aggregation operators and their applications to multiple-attribute decision making. *Int J Inf Technol Decis Making* 16(03):817–850
19. Garg H, Kumar K (2018) Some aggregation operators for linguistic intuitionistic fuzzy set and its application to group decision-making process using the set pair analysis. *Arab J Sci Eng* 43(6):3213–3227
20. Li Z, Liu P, Qin X (2017) An extended VIKOR method for decision making problem with linguistic intuitionistic fuzzy numbers based on some new operational laws and entropy. *J Intell Fuzzy Syst* 33(3):1919–1931
21. Kumar K, Mani N, Sharma A, Bhardwaj R (2021) 7 A novel entropy measure for linguistic intuitionistic fuzzy sets and their application in decision-making. In: Multi-criteria decision modelling: applicational techniques and case studies, pp 121–138. <https://doi.org/10.1201/9781003125150>
22. Kumar K, Chen SM (2022) Group decision making based on weighted distance measure of linguistic intuitionistic fuzzy sets and the TOPSIS method. *Inf Sci* 611:660–676
23. Kumar K, Chen SM (2022) Multiple attribute group decision making based on advanced linguistic intuitionistic fuzzy weighted averaging aggregation operator of linguistic intuitionistic fuzzy numbers. *Inf Sci* 587:813–824
24. Liu P, Liu W (2019) Multiple-attribute group decision-making based on power Bonferroni operators of linguistic q -rung orthopair fuzzy numbers. *Int J Intell Syst* 34(4):652–689
25. Verma R (2022) Generalized similarity measures under linguistic q -rung orthopair fuzzy environment with application to multiple attribute decision-making. *Granular Comput* 7(2):253–275
26. Liu P, Liu W (2019) Multiple-attribute group decision-making method of linguistic q -rung orthopair fuzzy power Muirhead mean operators based on entropy weight. *Int J Intell Syst* 34(8):1755–1794

27. Akram M, Naz S, Edalatpanah S, Mehreen R (2021) Group decision-making framework under linguistic q -rung orthopair fuzzy Einstein models. *Soft Comput* 25(15):10309–10334
28. Lin M, Li X, Chen L (2020) Linguistic q -rung orthopair fuzzy sets and their interactional partitioned Heronian mean aggregation operators. *Int J Intell Syst* 35(2):217–249
29. Bao H, Shi X (2022) Robot selection using an integrated MAGDM model based on ELECTRE method and linguistic q -rung orthopair fuzzy information. *Math Probl Eng* 2022:13. Article ID 1444486
30. Garg H, Kaur J (2018) A novel (r, s) -norm entropy measure of intuitionistic fuzzy sets and its applications in multi-attribute decision-making. *Mathematics* 6(6):92

Empirical Analysis of Unsupervised Link Prediction Algorithms in Weighted Networks



Ajay Kumar, Shashank Sheshar Singh, and Shivansh Mishra

Abstract Complex relationships in many real-world problems can be represented as networks where nodes represent individuals and relationships among them are represented by links. So, one of the key issues in such networks is the evolution or creation of edges. Link prediction is the solution where it finds the missing links (edges) in a static case (in a given snapshot of a network) or future links in a dynamic case (given several snapshots of networks at different time instants). In this experimental work, we consider the weighted versions of different existing algorithms and showed their performance on several real networks of different domains. We observed that the weighted version of the method path of length 3, i.e. L3-WT outperforms other methods on all four evaluation metrics with some exceptions. LHN1-WT method is the second outperformer on LesMiserables and Netscience datasets.

Keywords Link prediction · Social network analysis · Complex networks · Unsupervised algorithms

In today's scenario, lots of real complex problems can be represented by a network (or graph), e.g. social networks, biological networks, transport networks, etc. Social network analysis has recently attracted lots of attention among researchers due to its wide applicability in capturing social interactions. For example, friend recommendations on Facebook, product recommendations on online shopping websites like Flipkart, protein–protein interaction in biological networks, etc. These operations in networks are examples of link prediction (LP) in networks. Informally, LP can be defined as the finding missing in static networks or finding the likelihood of links that may appear in near future in dynamic networks. Formally, Liben-Nowell and Kleinberg expressed the LP as follows [1]: Given snapshots of networks (graphs) at

A. Kumar (✉)
UPES, Dehradun, India
e-mail: s.lajay007@gmail.com

S. S. Singh
Thapar University, Patiala, Punjab, India

S. Mishra
IIT BHU Varanasi, Varanasi, India

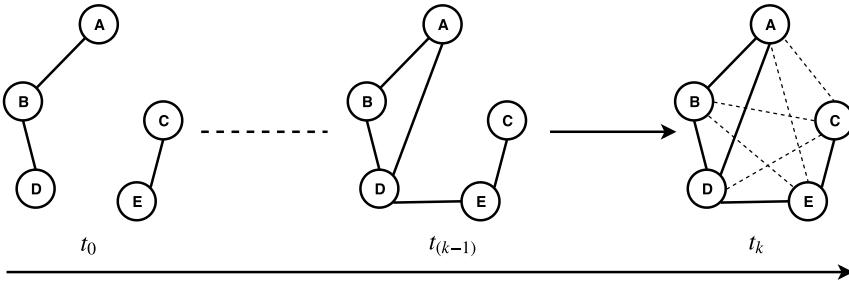


Fig. 1 Illustration of link prediction problem [2]

times $t_0, t_1, \dots, t_{(k-1)}$, LP seeks to predict the links (edges) that will be added to the network during the time interval from $t_{(k-1)}$ to a given future time t_k . The situation is best explained in Fig. 1.

In this figure, we have a network of five nodes. Initially, at t_0 , three links AB , BD , and CE are available; as the time proceeds, the graph evolves and two more links, i.e. AD and DE are generated at time $t_{(k-1)}$. Now, at the next time instant t_k , which links would likely to appear, is the link prediction task. Here, we have considered a simple undirected network, i.e. a graph without parallel edges and self-loops.

Mathematically, LP can be defined as follows. In an undirected graph with n nodes, total possible links; $U = \frac{n(n-1)}{2}$, number of existing links = E and number of non-existing links = $U - E$. LP finds $(U - E)$ number of links in a given network. In general, it tries to understand the association between two nodes and answer interesting questions such as [3]: How does the pattern of association change over time? What are the general factors driving the associations? To what extent can one model the evolution of a network using features intrinsic to the network itself? One can use methods such as perturbation theory to study the extent that the links are predictable in networks [4].

1 Related Work

In this section, we shed a light on a number of existing approaches/methods of link prediction in different types of networks. These methods can be grouped into several categories like similarity-based methods, probabilistic methods, algebraic methods, embedding-based, deep learning-based methods, and so on. Some seminal works [1, 3, 5, 6] are available in the literature that covers these methods comprehensively. In general, these methods assign connection weights $S(u, v)$ to each pair of non-existing nodes (u, v) and arrange them in descending order of their weights. Finally top- l links¹ are selected as the predicted links.

¹ In this article, we use links and edges interchangeably.

Lots of approaches to link prediction in social networks are available in the literature, in which some of the seminal review works are given in the above paragraph. Earlier Newman [7] posed the problem of link prediction where collaboration networks in Physics and Biology are considered. In the article, nodes represent the authors in the network and a link between two authors represents the coauthorship relation, i.e. whenever, any two authors work on a paper, there exists a link between them. Further, Liben-Nowell and Kleinberg [1] simulate link prediction in social networks. They simulate it by mapping a social entity or a person to a node and interaction between two entities to a link between two nodes. This was the first paper that gave the formal definition of link prediction. They computed the similarity scores between each pair of nodes by using the topological structure of the network. Hasan et al. [8] explored more about link prediction and applied supervised learning on it. For this, they simulated link prediction as a binary classification problem where scores were calculated by several structural similarity methods employed as features to the machine and different supervised learning algorithms applied to make predictions. Experimental results showed better accuracy compared to the unsupervised methods or heuristics. Later on, they explored the link prediction problem comprehensively and came up with a seminal survey paper [5]. Social network evolution is somehow related to link prediction because the evolution process of networks consists of the nodes as well as the link evolutions. The research on social network evolution nearly takes after the task of link prediction where several seminal studies have been proposed like Barabasi and Albert work [9] on random network published in Science, Kleinberg work [10] in Nature Communication, [11] in KDD, [1] in EPL, and [12] in Scientific Reports. In recent years, several different types of approaches to link prediction have been proposed that include deep learning methods [13, 14], embedding-based methods [15, 16], probabilistic and maximum likelihood approaches [17–19], and so on.

2 Proposed Work

In this article, we present an experimental work comparing different existing link prediction methods but the weighted versions. The article details a comprehensive analysis of nine existing heuristics (methods) that are unsupervised scoring algorithms. Out of nine methods, six of them belong to local similarity where only local topological (i.e. immediate neighbours, neighbours of neighbours, etc.) information, and two of them belong to quasi-local methods where more than local topological information is considered. The last one belongs to the global similarity approach where the entire topological information of a network is taken under consideration. We have taken these existing methods in weighted network scenarios.

All the unsupervised methods considered here, are similarity-based approaches to link prediction. Five methods (CN-WT, JC-WT, AA-WT, RA-WT, CC-WT) out of nine are based on common neighbours (or mutual friends) and are local similarity approaches. Additionally, two methods (i.e. LPI-WT and L3-WT) are based on path

measures and quasi-local approaches, one method is based on the overall topology of a network and the global similarity approach. The last one, i.e. PA-WT is also a local similarity method but is based on the degree of nodes in a network. Here, we have considered real data from different domains. We split each dataset into the training set and test set. For each method, we calculate the similarity score of all non-existing edges and evaluate it further. During the evaluation, test data is considered. In this article, we split the original data into five different fractions of training and test data. The test sets are of sizes 0.1, 0.2, 0.3, 0.4, 0.5 or 10, 20, 30, 40, 50% and training data split accordingly.

3 Empirical Analysis

3.1 Evaluation Metrics

Existing literature contains several metrics to evaluate the link prediction problem. In this article, we employed some of them viz., area under the precision-recall curve (AUPR) [20], area under the ROC curve (AUC) [21], balanced accuracy (Bal_ACC) [22], and *F*1 score [23].

Area under the precision-recall curve (AUPR) The precision-recall (PR) curve is a widely used measure of link prediction algorithms' performance. Given a threshold, both precision and recall values are computed and plotted on *Y*-axis and *X*-axis respectively, i.e. in this curve, the precision is plotted as a function of the recall. The area bounded by the curve is AUPR values. A higher value of AUPR represents the better predictor or model for a given dataset.

Area under the ROC curve (AUC) Receiver operating characteristic curve (ROC) is also a 2-D depiction of a classifier's performance measure where true positive rate (TPR) and false positive rate (FPR) are plotted on *Y*-axis and *X*-axis, respectively. To compare the performance of two classifiers, AUC, a single scalar value is used. Its value varies from 0 to 1. Best predictors (methods) evaluate their AUC values nearer to 1, i.e. higher the value of the AUC, better the predictor.

Balanced Accuracy (Bal_ACC) Balanced accuracy is a better metric to use with imbalanced data. It accounts for both positive and negative outcome classes and doesn't mislead with imbalanced data. It is simply the arithmetic mean of sensitivity and specificity.

F1 Score The difference between the balanced accuracy score and the *F*1 score is that the first metric is the arithmetic mean of sensitivity and specificity whereas the second metric is the geometric mean of precision and recall. It also handles class imbalance problems.

- (1) If Neg \gg Pos, *F*1 is a better.
- (2) If Pos \gg Neg, balanced accuracy is better.

Table 1 Topological properties of networks

	Nodes (N)	Edges (E)	Avg. degree (K)	Avg. distance (D)	Avg. clustering coeff. (C)
Football	35	118	0	1.301	0.169
SocialWorkJ	36	99	2.75	1	0
World_trade	80	875	10.983	1.718	0.376
Lesmiserables	77	254	3.299	2.400	0.287
StarLink	113	638	5.646	2.647	0.372
Contact_diary	123	753	6.122	4.499	0.332
Celegansneural	297	2345	7.896	3.992	0.169
USAir97	332	2126	12.807	2.738	0.749
Netscience	1589	2742	3.451	5.823	0.878

Clearly, if you can label-switch, both metrics can be used in any of the two imbalance cases above. If not, then depending on the imbalance in the training data, you can select the appropriate metric.

3.2 Dataset Description

We compare all the existing methods mentioned here on real-world network datasets. We have considered these datasets from different domains viz., Netscience (a co-authorships in network science), Celegansneural (a neural network of *C. Elegans*), LesMiserables (a coappearance network of characters in the novel *LesMiserables*), USAir97 (a transport network), and Contact_diary, SocialWorkJ, StarLink and Football (social networks). Some topological properties of these network datasets is given in Table 1. Here, $|N|$ and $|E|$ are the numbers of nodes and edges in a network. The average degree and average path length (distance) of the network are represented by K and D , respectively. C represents the average clustering coefficient of the considered network.

3.3 Methods to Compare

In this article, we have used weighted versions of the existing link prediction methods. Though the authors in [24] also have done the comparison but they employ supervised algorithms for the experimental comparison in weighted networks. Here, we applied unsupervised methods or heuristics namely, Common neighbours [7], Preferential

attachment [7], Jaccard coefficients [25], Adamic Adar index [26], Resource allocation index [27], Clustering coefficient index [28], Local path index [29], L3 index [30], and LHN1 index [31].

- Common Neighbour Index-Weighted (CN-WT)

$$S(u, v) = \sum_{z \in N(u) \cap N(v)} w(u, z) + w(z, v) \quad (1)$$

- Preferential Attachment Index-Weighted (PA-WT)

$$S(u, v) = \sum_{z \in N(u) \cap N(v)} w(u, z) \times w(z, v) \quad (2)$$

- Jaccard Coefficient Index-Weighted (JC-WT)

$$S(u, v) = \frac{\sum_{z \in N(u) \cap N(v)} w(u, z) + w(z, v)}{\sum_{a \in N(u)} w(u, a) + \sum_{b \in N(v)} w(b, v)} \quad (3)$$

- Adamic-Adar Index-Weighted (AA-WT)

$$S(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{w(u, z) + w(z, v)}{\log(1 + \Gamma(z))} \quad (4)$$

- Resource Allocation Index-Weighted (RA-WT)

$$S(u, v) = \sum_{z \in N(u) \cap N(v)} \frac{w(u, z) + w(z, v)}{\Gamma(z)} \quad (5)$$

- Clustering Coefficient Index-Weighted (CC-WT)

$$S(u, v) = CC(u) + CC(v), \quad (6)$$

$$\text{where, } CC(x) = \frac{1}{\Delta(x) * (\Delta(x) - 1)} * \sum_{m, n \in \Delta(x)} \frac{w(x, m) + w(n, x)}{2 * \sum_{z \in \Delta(x)} \frac{w(z, x)}{|\Delta(x)|}}$$

- Local Path Index-Weighted (LPI-WT)

$$S(u, v) = \sum_{z \in N(u) \cap N(v)} w(u, z) + w(z, v) + q, \quad (7)$$

$$\text{where } q = p * \left(\sum_{x, y \in \text{path}(u, x, y, v)} w(u, x) + w(x, y) + w(y, v) \right).$$

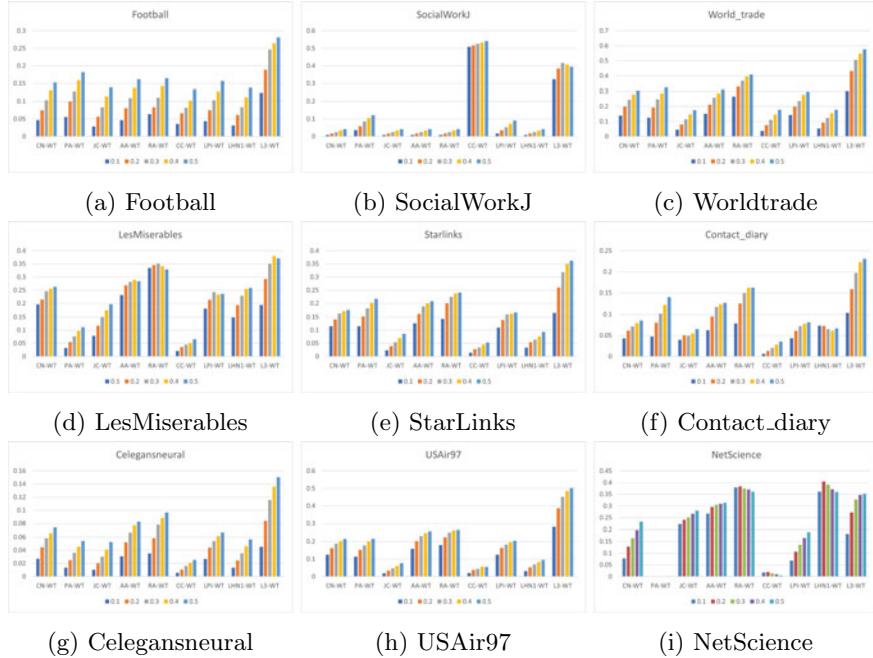


Fig. 2 AUPR results

3.4 Result Analysis

In this section, we analyse the result of the proposed and existing methods with different evaluation metrics. The analysis has been carried out on nine real-world network datasets about which the description has been given in the data description part of the article.

Measuring AUPR Figure 2 shows the AUPR results of nine (9) state-of-the-art methods on nine (9) real networks.² We observe that, L3-WT method (a quasi-local) best performs on all datasets for all fractions³ of test data except LesMiserables, C elegansneural, and Netscience. On LesMiserables data, it performs best on 0.4 and 0.5 fraction of test data, RA-WT is best performer for 0.1 and 0.3 fraction of test data, while CC-WT is best when we consider 0.2 or (20%) of test data. On C elegansneural, L3-WT shows best results on 0.1, 0.2, and 0.5 fraction of test data where as RA-WT is best on 0.3 and 0.4 test data. On the Netscience dataset, L3-WT is best performer on 0.2, 0.3, and 0.4 (or 10, 20, and 40%) fraction of test data while RA-WT is best on the remaining fractions of test data.

² In the article, we use networks and datasets interchangeably.

³ 0.1 Fraction of test data means 10% of the total data is taken as test set.

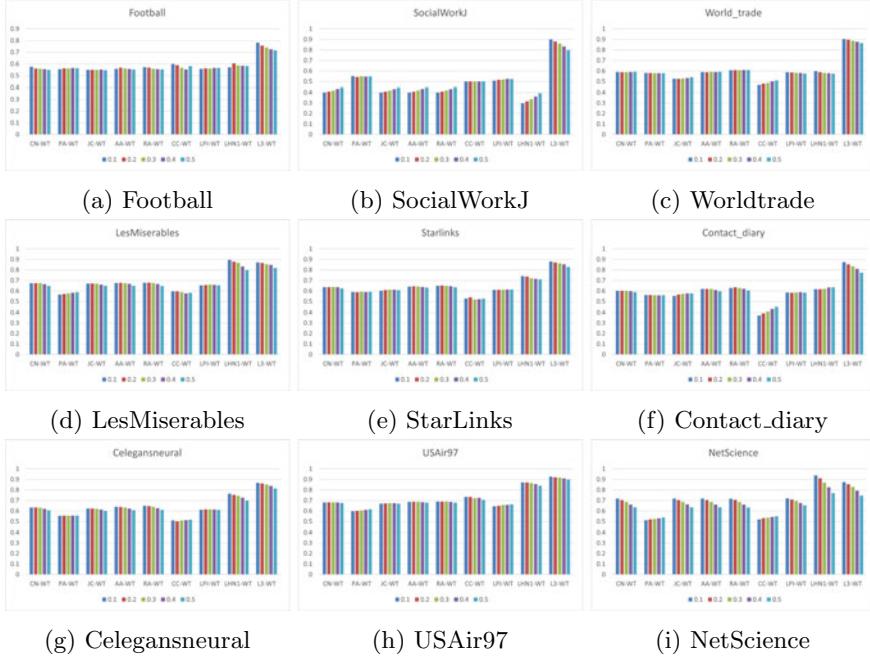


Fig. 3 AUC results

Measuring AUC The area under the ROC curve (AUC results) of the existing methods on the considered datasets is shown in Fig. 3. We observe that the AUC results of the method L3-WT are the best among the considered methods on all fractions of test data for all datasets except LesMiserable and Netscience. It also performs best for 0.4 and 0.5 fractions of test data on LesMiserables dataset. For the remaining fractions, LHN1-WT performs best against all the considered methods. LHN1 is also best-performing method on the Netscience among the other methods for each fraction of test data.

Measuring Balanced Accuracy (Bal_ACC) We present the Bal_ACC results for each method on all datasets in Fig. 4. We observe that these results are even more consistent where L3-WT outperforms state-of-the-arts on each fraction of test set of all datasets except LesMiserables and Netscience. We also observe that the L3-WT results are best with significant margins. On the remaining two datasets (i.e. LesMiserables and Netscience), LHN1-WT method performs best among all the methods considered here. The difference between the LHN1-WT and L3-WT are not of great margins but are of great margins when compared with the other existing methods in the table.

Measuring F1-Score F1 score of the existing methods on each dataset is shown in Fig. 5. Here, we observe that L3-WT outperforms consistently for each fraction of the test set on six datasets viz., USAir97, Contact_diary, World_trade, Starlinks,

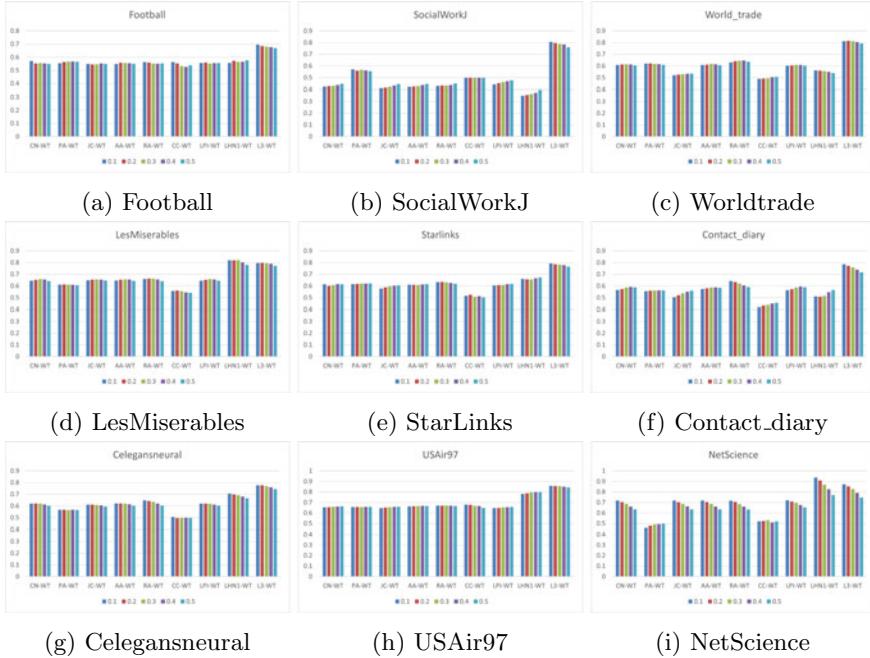


Fig. 4 Balanced accuracy results

SocialWorkJ, and Football networks. On LesMiserables and Netscience, LHN1-WT is best performing method for each fraction of the test set with an exception of 0.1 fraction of test set on Netscience where AA-WT performs best. On Celegansneural dataset, RA-WT outperforms all methods for each fraction of test data. We also observe that the $F1$ score values of CN-WT, JC-WT, AA-WT, RA-WT, and LHN1-WT methods are null (0) for all fraction of SocialWorkJ dataset. On this dataset, L3-WT is best among other methods with great margins.

4 Conclusions

In this article, we present a comprehensive comparison of several existing methods of link prediction in complex networks. The methods under consideration are unsupervised that are based on the topological structure of networks. Here, we consider the weighted versions of all the methods, i.e. these methods are applied to weighted networks. During the comprehensive analysis, we observe that the quasi-local method namely, L3-WT outperforms others in most of the network datasets except LesMiserables, Celegansneural, and Netscience. On these three datasets, the global method namely, LHN1-WT is the best performer. We have made these obser-

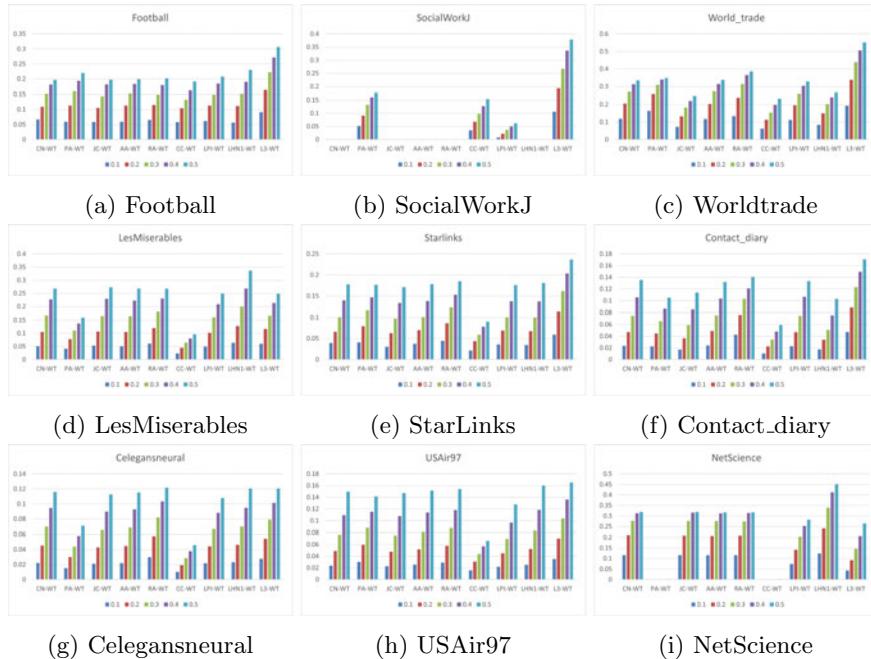


Fig. 5 *F1 score results*

vations on nine real network datasets from different domains. Also, while implementation, we have evaluated these methods on four evaluation approaches viz., the area under the precision-recall curve (AUPR), area under the ROC curve (AUC), balanced accuracy score, and *F1* score.

References

1. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings of the twelfth international conference on information and knowledge management, CIKM'03. ACM, New York, NY, USA, pp 556–559
2. Kumar A, Mishra S, Singh SS, Singh K, Biswas B (2020) Link prediction in complex networks based on significance of higher-order path index (shopi). *Phys A Stat Mech Appl* 545:123790
3. Kumar A, Singh SS, Singh K, Biswas B (2020) Link prediction techniques, applications, and performance: a survey. *Phys A Stat Mech Appl* 553:124289
4. Lü L, Pan L, Zhou T, Zhang Y-C, Stanley HE (2015) Toward link predictability of complex networks. *Proc Natl Acad Sci* 112(8):2325–2330
5. Al Hasan M, Zaki MJ (2011) A survey of link prediction in social networks. Springer US, Boston, MA, pp 243–275
6. Martínez V, Berzal F, Cubero J-F (2017) A survey of link prediction in complex networks. *ACM Comput Surv* 49(4):1–33

7. Newman MEJ (2001) Clustering and preferential attachment in growing networks. *Phys Rev E* 64:025102
8. Al Hasan M, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: Proceedings of SDM'06 workshop on link analysis, counterterrorism and security
9. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512
10. Kleinberg JM (2000) Navigation in a small world. *Nature* 406(6798):845
11. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the Eleventh ACM SIGKDD international conference on knowledge discovery in data mining, KDD'05. ACM, New York, NY, USA, pp 177–187
12. Zhang Q-M, Xiao-Ke X, Zhu Y-X, Zhou T (2015) Measuring multiple evolution mechanisms of complex networks. *Sci Rep* 5:10350
13. Li X, Du N, Li H, Li K, Gao J, Zhang A (2014) A deep learning approach to link prediction in dynamic networks. In: Proceedings of the 2014 SIAM international conference on data mining, Philadelphia, Pennsylvania, USA, 24–26 Apr 2014, pp 289–297
14. Zhang M, Chen Y (2017) Weisfeiler-Lehman neural machine for link prediction. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Halifax, NS, Canada, 13–17 Aug 2017, pp 575–583
15. Grover A, Leskovec J (2016) Node2vec: scalable feature learning for networks. In: Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD'16. ACM, New York, NY, USA, pp 855–864
16. Kipf TN, Welling M (2016) Variational graph auto-encoders. *CoRR*, abs/1611.07308
17. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453(7191):98–101
18. Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. *Proc Natl Acad Sci* 106(52):22073–22078
19. Getoor L, Friedman N, Koller D, Taskar B (2002) Learning probabilistic models of link structure. *J Mach Learn Res* 3:679–707
20. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on machine learning. ICML'06. ACM, New York, NY, USA, pp 233–240
21. Fawcett T (2006) An introduction to roc analysis. *Pattern Recogn Lett* 27(8):861–874
22. Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition, pp 3121–3124
23. Sasaki Y (2007) The truth of the f -measure. *Teach Tutor Mater*
24. de Sá HR, Prudêncio RBC (2011) Supervised link prediction in weighted networks. In: The 2011 international joint conference on neural networks, pp 2281–2288
25. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings of the twelfth international conference on information and knowledge management, CIKM'03. Association for Computing Machinery, New York, NY, USA, pp 556–559
26. Lada A, Adar E (2003) Friends and neighbors on the web. *Soc Netw* 25:211–230
27. Zhou T, Lü L, Zhang Y-C (2009) Predicting missing links via local information. *Eur Phys J B* 71(4):623–630
28. Zhihao W, Lin Y, Wang J, Gregory S (2016) Link prediction with node clustering coefficient. *Phys A Stat Mech Appl* 452:1–8
29. Lü L, Jin C-H, Zhou T (2009) Similarity index based on local paths for link prediction of complex networks. *Phys Rev E* 80:046122
30. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim D-K, Kishore N, Hao T, Vidal MAM, Barabási A-L (2018) Network-based prediction of protein interactions. *bioRxiv*, Calderwood
31. Leicht EA, Holme P, Newman MEJ (2006) Vertex similarity in networks. *Phys Rev E* 73:026120

Detection of Fraudulent Credit Card Transactions Using Deep Neural Network



Kotireddy Yazna Sai, Repalle Venkata Bhavana, and Natarajan Sudha

Abstract The usage of credit cards for online purchases has increased exponentially and so has fraud. In recent years, detecting fraud in credit card transactions has become significantly more difficult. As a result, having effective and accurate methods for detecting fraud in credit card transactions is vital. Although supervised learning algorithms have been shown to be effective in detecting credit card fraud, they have not generated substantial results. Hence, a Deep Neural Network (DNN)-based technique is proposed. The sequential model is used to construct the proposed DNN. However, the parameters of the model, such as the number of nodes and the activation function in the hidden layers, can also have an impact on its output. The proposed technique was evaluated on a credit card transaction dataset that comprised fraudulent and genuine transactions. This technique obtained the accuracy, area under curve (AUC), and precision of 99.93%, 99.99%, and 99.89%, respectively. This technique was compared with various models such as Optimized LightGBM, Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM).

Keywords Deep neural networks · Credit card transactions · Deep learning · Fraud detection

1 Introduction

A credit card is a card issued to a cardholder that allows them to purchase goods and services within their credit limit or withdraw cash in advance. Credit cards provide the cardholder with a time advantage, allowing them to repay later in a certain time frame by rolling it over to the next payment cycle.

K. Yazna Sai · R. Venkata Bhavana (✉) · N. Sudha
SASTRA Deemed University, Thanjavur, India
e-mail: repallebhavana123@gmail.com

K. Yazna Sai
e-mail: yaznasai.16@gmail.com

N. Sudha
e-mail: sudha@cse.sastra.edu

The unlawful use of a debit card or credit card to make payments is referred to as credit card fraud. When we talk about credit card fraud, we mean scenarios in which a payment card is used to make a transaction without the cardholder's knowledge or agreement. Due to the high dependence on technology and the internet, the use of payment cards increased so is the credit card transactions. The rate of credit card fraud is rising because the use of credit cards has become the most common means of payment for both offline and online transactions.

According to the Federal Trade Commission (FTC) statistics, in the year of 2017, there have been almost 1579 intrusions and over 179 million records, with credit card fraud being the most prevalent (133,015 incidents), followed by job or tax-related fraud (82,051 reports), phone fraud (55,045 reports), and bank fraud (50,517 reports) [1].

To address this issue, we need a technique that can cancel a transaction if it detects something strange. Detection of credit card fraud is the technique of detecting fraudulent purchase attempts and denying them instead of performing the order. There are several machine learning and deep learning techniques available nowadays that can assist us in classifying irregular transactions. The sole prerequisite is previous data and an appropriate algorithm that can better match our data.

This research introduces a method for detecting fraud that employs a Deep Neural Network. Using a dataset obtained from Kaggle, we compare the performance of our strategy to that of other standard machine learning techniques.

2 Related Works

For understanding the importance of various supervised and unsupervised learning techniques for detecting credit card frauds, we have studied many research papers. Techniques ranging from Random Forest to LSTM were implemented and reviewed in these papers. Table 1 shows the detailed list of the papers that are studied on the concept credit card fraud detection.

3 Proposed Approach

This paper proposes an approach having three major steps: data preprocessing, DNN model building, and performance evaluation, and these are detailed in the following subsections. Figure 1 presents the entire framework of the proposed approach.

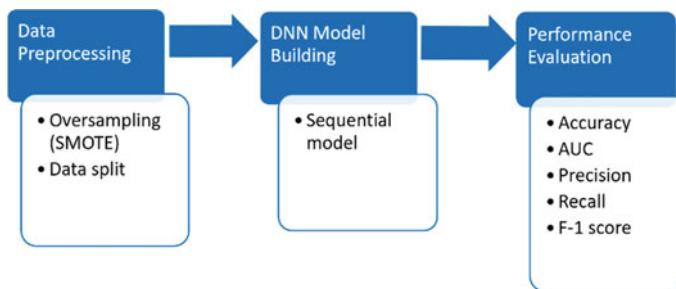
Table 1 Literature survey

Related work	Methodology	Remarks
Esenoglu et al. [1]	LSTM serves as the AdaBoost technique's base learner. As a resampling approach, the SMOTE-ENN method is applied	LSTM outperformed SVM, MLP, DT, and traditional AdaBoost
Asha et al. [2]	Artificial neural networks	ANN gives the best accuracy than SVM and KNN
Chen et al. [3]	Deep convolution neural network scheme using deep learning algorithm	Accuracy obtained is 99%
Pandey et al. [4]	Many algorithms like RF, KNN, Naïve Bayes, SVM, DT, LR, Genetic Algorithm were studied	Imbalanced dataset problem was not resolved
Sanober et al. [5]	Using a ranking strategy where the alert is positioned according to priority, autoencoder is employed in the spark environment to categorize the alert as fake or even allowed	Class imbalance problem is resolved. Achieved better results
Vinutha et al. [6]	SVM, Naïve Bayes, KNN, and random forest	SVM, Naïve Bayes provided the highest accuracy
Priscilla et al. [7]	Optimized XGBoost approach to handle class imbalance in datasets. Randomized search CV hyperparameter optimization technique	Gave higher accuracy for extremely imbalanced data without sampling
Rai et al. [8]	Neural Network (NN)-based unsupervised learning technique	More accuracy than AE, LOF, IF, k-means clustering
Dubey et al. [9]	ANN algorithm and backpropagation	Gave accuracy of 99.96%
Ge et al. [10]	Detection algorithm using LightGBM	Good accuracy and AUC scores but can improve upon optimization
Khatri et al. [11]	Decision tree, KNN, LR, RF and Naïve Bayes are compared	Decision tree gave best results
Taha et al. [12]	Optimized LightGBM and Bayesian-based hyperparameter tuning	Gave better results compared with other algorithms
Varmedja et al. [13]	Random forest, Naïve Bayes, logistic regression, and multilayer perceptron. SMOTE technique used for oversampling of dataset	Random forest gave the best results
Fiore et al. [14]	Generative adversarial networks (GANs) for improving classification effectiveness	GAN achieved an improved sensitivity at the cost of a limited increase in false positives

(continued)

Table 1 (continued)

Related work	Methodology	Remarks
Dornadula et al. [15]	ML algorithms like RF, SVM, LOF, IF, LR, DT were compared and SMOTE technique was used as resampling technique	Matthews correlation coefficient was the better parameter to deal with imbalance dataset. LR, DT, and RF gave better results
Kirkos et al. [16]	Bayesian belief networks for detecting fraud financial statements	Best performance by classifying 90.3% of validation sample than NN and ID3

**Fig. 1** Flow diagram for the proposed approach

3.1 Data Preprocessing

The appropriateness of the proposed methodology is evaluated using a real-world dataset [17]. It includes the credit card transactions made by cardholders over the course of 2 days in the month of September 2013. There are a total of 284,807 transactions, 492 of which are fraudulent, resulting in a 0.172 percent fraud rate. This dataset is extremely imbalanced. Because disclosing a customer's transaction information poses a confidentiality risk, the majority of the dataset's attributes are analyzed using principal component analysis (PCA). All attributes are PCA transformed except 'Time', 'Amount', and 'Class'. Finally, there are 31 attributes present in the European dataset, which are V1, V2,..., V28 as PCA applied attributes, 'Time', 'Amount', and 'Class'. There are no missing data in the whole dataset.

From Fig. 2, with regard to 'Time' attribute, we notice no clear trend for fraudulent and non-fraudulent transactions. As a result, we may remove the column of 'Time' attribute. From Fig. 3, with regard to 'Amount' attribute, we notice that fraudulent transactions are largely concentrated in the lower range of amount, but non-fraudulent transactions are distributed over the low-to-high range of amount. As we can notice some trend for fraudulent and non-fraudulent transactions for 'Amount' attribute, we normalized the column instead of removing it.

An imbalanced dataset is more likely to produce inaccurate outcomes. To minimize overlearning/overfitting, the Synthetic Minority Oversampling Technique

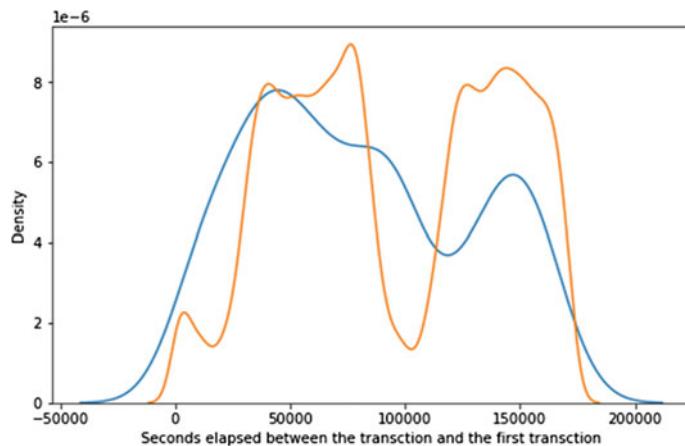


Fig. 2 Distribution of classes with ‘time’

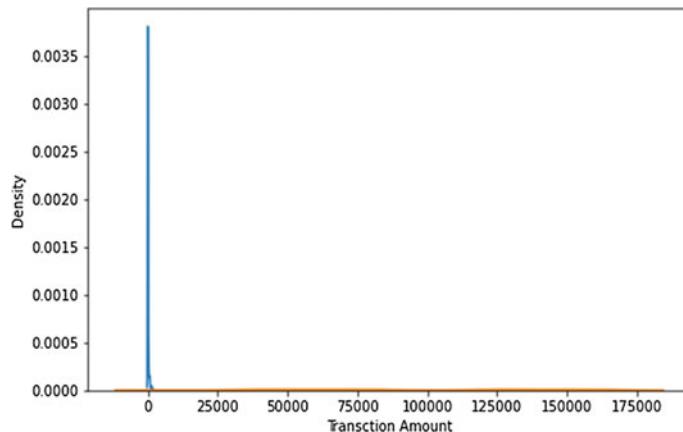


Fig. 3 Distribution of classes with ‘amount’

(SMOTE) is used to evaluate the proposed method. The dataset is split into training set of 80% and testing set of 20%.

3.2 *DNN Model Building*

Deep Neural Networks (DNNs) have got quite a bit of focus in recent years because of their considerable results in a variety of applications. There are many deep learning techniques present for identifying fraudulent credit card transactions. In this paper,

sequential model is used for the detection of fraudulent credit card transactions. Sequential model is a Keras API, that stacks the layers starting from input layer, followed by hidden layers (≥ 1 layer), ending with an output layer. The architecture of the proposed method contains five layers, which is shown in Fig. 4. Each layer used in the model is a dense layer, a layer of nodes that receives the input from its previous layer's nodes.

The first layer is the input layer, which uses 29 units, and ReLU is used as the activation function. The Rectified Linear Activation Unit (ReLU) provides nonlinearity to the network and helps to keep the compute required to run the neural network from growing exponentially. The ReLU function is represented as:

$$f(x) = \max(0, x)$$

where x is an input value.

The next three layers are the hidden layers, which use ReLU as the activation function, but varies in the number of units. Second layer used 24 units, third layer used 20 units, and fourth layer used 15 units. The final layer is the output layer, which uses sigmoid function as its activation function and has 1 unit. Sigmoid function always gives output in the range 0 and 1. Since detecting credit card fraud is a binary classification problem, the best choice is using sigmoid function. The sigmoid function is represented as:

$$S(x) = \frac{1}{1 + e^{-x}}$$

where x is an input value.

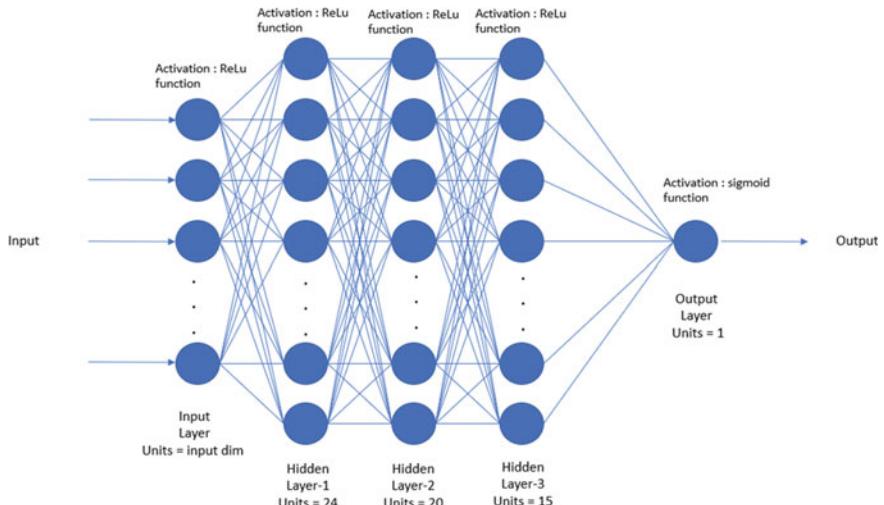


Fig. 4 Architecture of the proposed approach

After building the model, we compiled it by setting a loss function, optimizer, and metric. Loss function is a means of determining how effectively a particular model mimics the provided data. If projections diverge too far from actual outcomes, the loss function will return a big value. Gradually, with the assistance of some optimization function, the loss function learns to minimize prediction error.

We used ‘binary_crossentropy’ as the loss function while compiling, because detecting credit card fraud is a binary classification problem. Optimizers adjust the parameters of a neural network, such as learning rate, bias of neurons, and weights, to decrease loss. We used ‘adam’ optimizer as the optimizer of the model. ‘Adam’ works faster and with more efficiency than other optimizers. The metric used is ‘accuracy’ to evaluate the performance of the model.

3.3 Performance Evaluation

The task of detecting credit card fraud is an imbalanced classification problem: we have two classes to classify as fraud and non-fraud, with one group (non-fraud) accounting for the great majority of data points. The positive class considerably outnumbers the negative class. As a result, accuracy is an inadequate criteria for assessing model performance. Hence, the performance of DNN model is evaluated by calculating confusion matrix, accuracy, precision, recall, F1-score, precision-recall (PR) curve, and ROC curve. Confusion matrix represents True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) in matrix format.

Accuracy is a metric used to measure the performance of the model across all classes. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Precision is a metric which is used to measure the model’s efficiency to detect a data as positive. The precision is calculated by using the equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is a metric which is used to measure the model’s ability to identify positive samples. It is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score is a metric which is used to measure the percentage of correct predictions that a model has made. It is calculated as:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

PR curve is a curve between precision and recall for various threshold values, whereas ROC curve is plotted against True-Positive Rate (TPR) and False-Positive Rate (FPR) for various threshold values. These curves are also important for determining the efficiency of a model and used in comparing with other algorithms.

4 Experimental Results

The proposed approach was implemented in Jupyter Notebook, evaluated with European dataset. The DNN model is built by using five layers, where first four layers use ReLU activation function and last layer (output layer) uses sigmoid activation function. The model is compiled using ‘binary cross-entropy’ loss function and ‘adam’ optimizer. After the compilation, the model is fitted with batch size of 30 and 100 epochs.

Confusion matrix is plotted for testing dataset, which is shown in Fig. 5. Accuracy achieved by the proposed model is 99.95%, whereas recall and precision are 99.99% and 99.89%, respectively. The results obtained after the evaluation of the proposed model are listed in Table 2. ROC and PR curves are plotted and shown in Fig. 6.

Proposed DNN model is compared with algorithms like Optimized LightGBM [12], Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), SVM Linear, SVM RBF, Gaussian Naïve Bayes, KNN, and MLP. DNN model achieved the highest accuracy of 99.93% for the European dataset. From Table 3, we can infer

Fig. 5 Confusion matrix for the proposed DNN model



Table 2 Metrics values of proposed DNN model for the dataset (in %)

Accuracy	AUC	Precision	Recall	F1-score
99.9340	99.9923	99.8961	99.9718	99.9339

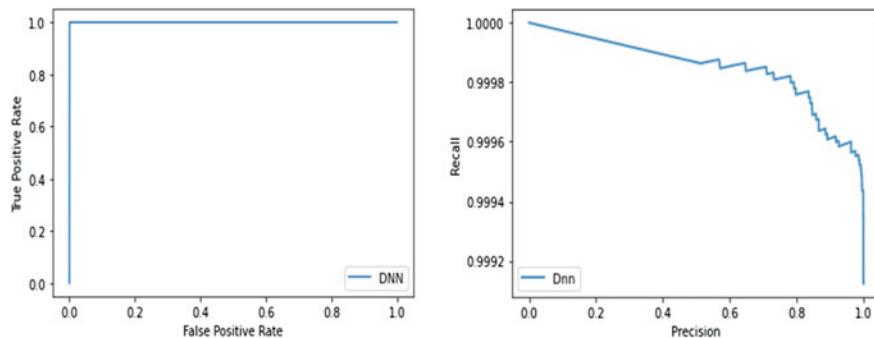


Fig. 6 ROC and PR curves for the proposed DNN model

Table 3 Metric values of proposed DNN model and other techniques for the dataset (in percentages)

Approach	Accuracy	AUC	Precision	Recall	F1-score
DT	90.3410	90.2130	94.2525	85.4153	89.5982
Gaussian Naïve Bayes	91.8680	91.7165	98.2237	84.9071	91.0513
KNN	92.3720	92.2220	94.3952	89.7915	91.9983
RF	93.6450	93.5043	98.5395	88.2540	93.1022
SVM RBF	94.0280	93.9409	97.2438	90.3982	93.6870
LR	94.1554	94.0653	97.5077	90.3696	93.7851
SVM Linear	94.1562	94.1115	96.4877	91.4679	93.8976
Optimized LightGBM	99.6019	90.8022	65.1456	81.9719	67.2296
MLP	99.6157	99.6150	99.2869	99.7419	99.2741
Proposed DNN model	99.9340	99.9923	99.8961	99.9718	99.9339

that the proposed DNN model (highlighted in bold) outperforms other algorithms and thus proves to be a better approach. Figure 7 shows the comparison of ROC curves and PR curves for all the mentioned algorithms.

The performance of the proposed approach is compared in terms of accuracy with Taha et al. [12] with accuracy of 98.4% and Pranali Shenvi et al. [18] with the accuracy of 99.72%, and the comparison shows that the proposed approach performed better with the accuracy of 99.93%.

5 Conclusion

The rise in credit card frauds is troubling many companies and individuals. Detecting these fraudulent transactions is necessary to prevent any losses. In this paper, a deep learning technique, that is a sequential DNN model, incorporated with five dense layers, is proposed to detect credit card frauds more accurately. The highly

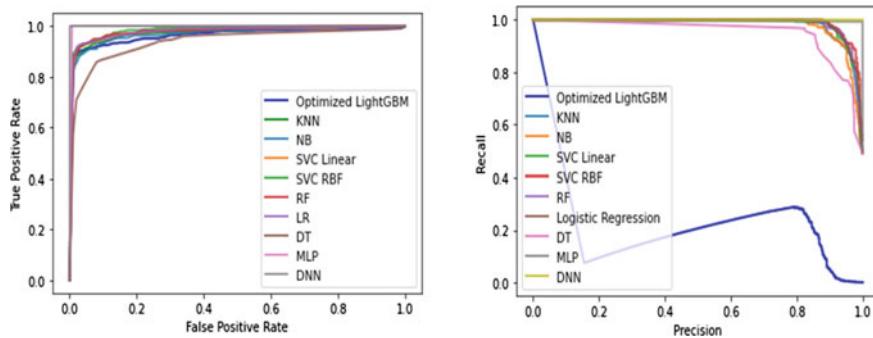


Fig. 7 ROC and PR curves for proposed DNN model and other algorithms

imbalanced dataset problem was tackled by using SMOTE, a type of oversampling technique.

The proposed model is evaluated based on accuracy, AUC, precision, recall, F1-score, confusion matrix, ROC curve, and PR curve. The proposed DNN model produced the best results compared to many supervised and unsupervised learning algorithms. More sophisticated deep learning techniques might be used in future research to improve the performance of the model described in this paper.

References

1. Esenogho E, Mienye ID, Swart TG, Aruleba K, Obaido G (2022) A neural network ensemble with feature engineering for improved credit card fraud detection. *IEEE Access* 10:16400–16407
2. Asha RB, KR SK (2021) Credit card fraud detection using artificial neural network. *Global Transitions Proc* 2(1):35–41
3. Chen JIZ, Lai KL (2021) Deep convolution neural network model for credit-card fraud detection and alert. *J Artif Intell* 3(02):101–112
4. Pandey K, Sachan P, Ganpatrao NG (2021) A review of credit card fraud detection techniques. In: 2021 5th international conference on computing methodologies and communication (ICCMC), pp 1645–1653. IEEE
5. Sanober S, Alam I, Pande S, Arslan F, Rane KP, Singh BK, Khamparia A, Shabaz M (2021) An enhanced secure deep learning algorithm for fraud detection in wireless communication. *Wirel Commun Mobile Comput*
6. Vinutha H, Joyson A, Apoorva J, Ashitha GR, Tejashwini B (2021) Credit card fraud identification using machine learning algorithm. *J Contemp Issues Bus Gov* 27(3)
7. Priscilla CV, Prabha DP (2020) Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection. In: 2020 third international conference on smart systems and inventive technology (ICSSIT), pp 1309–1315. IEEE
8. Rai AK, Dwivedi RK (2020) Fraud detection in credit card data using unsupervised machine learning based scheme. In: 2020 international conference on electronics and sustainable communication systems (ICESC), pp 421–426. IEEE

9. Dubey SC, Mundhe KS, Kadam AA (2020) Credit card fraud detection using artificial neural network and backpropagation. In: 2020 4th international conference on intelligent computing and control systems (ICICCS), pp 268–273. IEEE
10. Ge D, Gu J, Chang S, Cai J (2020) Credit card fraud detection using LightGBM model. In: 2020 international conference on e-commerce and internet technology (ECIT), pp 232–236. IEEE
11. Khatri S, Arora A, Agrawal AP (2020) Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 2020 10th international conference on cloud computing, data science & engineering (confluence), pp 680–683. IEEE
12. Taha AA, Malebary SJ (2020) An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access* 8:25579–25587
13. Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A (2019) Credit card fraud detection-machine learning methods. In: 2019 18th international symposium INFOTEH-JAHORINA (INFOTEH), pp 1–5. IEEE
14. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F (2019) Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Inf Sci* 479:448–455
15. Dornadula VN, Geetha S (2019) Credit card fraud detection using machine learning algorithms. *Procedia Comput Sci* 165:631–641
16. Kirkos E, Spathis C, Manolopoulos Y (2007) Data mining techniques for the detection of fraudulent financial statements. *Expert Syst Appl* 32(4):995–1003
17. Credit Card Fraud Dataset. Available <https://www.kaggle.com/mlg-ulb/creditcardfraud/data>. Last accessed 4 Sept 2019
18. Shenvi P, Samant N, Kumar S, Kulkarni V (2020) Implementation of Interpolation in credit card fraud detection. *Soft Comput Signal Process* 1118:125

Recommendation System and Its Techniques in Machine Learning: A Survey



Neeru Banwala, Gurpreet Singh, Jaspreet Singh,
Vishwajeet Shankar Goswami, and Aashima Bagnia

Abstract Machine learning-based recommendation systems are formidable tools that target specific customers with tailored product and content recommendations based on their user data and behavioural patterns. Due to the abundance of information available today, it is now challenging to separate out and provide the user with the most pertinent information. Therefore, using recommendation systems to reduce time and costs for both the business and the user is now necessary to address the issue of information overload. Based on a user's interest and past preferences, these systems might suggest suitable products to them, thus increasing revenue. This paper seeks to clarify the idea of the recommendation systems built upon, the many methods used to create these systems and the necessity of using these systems in the current and future environments.

Keywords Content filtering · Collaborative filtering · Hybrid filtering · Machine learning · Recommendation systems

1 Introduction

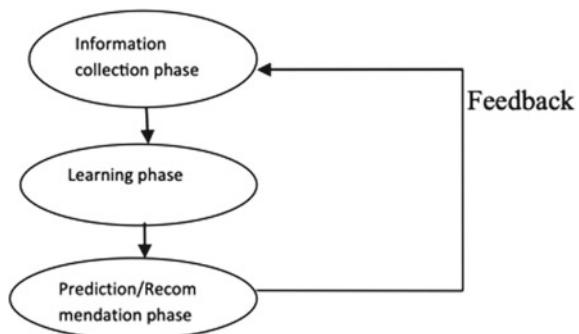
In today's era of this digital world, there is a huge rise in the number of individuals using internet, which ultimately have resulted in the rapid increase in the volume of digital information available, and hence, it has created the problem of information overload which prevents the user from timely access of the relevant information. Some big techs like Google, DevilFinder, and Altavista are examples of information retrieval systems that have largely solved this challenge although prioritisation and

N. Banwala (✉)
Data Science, Department of Mathematics, Chandigarh University, Mohali, India
e-mail: neerubanwala9510@gmail.com

G. Singh · J. Singh
Department of Computer Science and Engineering, Chandigarh University, Mohali, India

V. S. Goswami · A. Bagnia
Department of Mathematics, Chandigarh University, Mohali, India
e-mail: vishwajeet.e9858@cumail.in

Fig. 1 Phases in recommendation process



personalisation of information are still lacking. This problem of information overload and delayed access to the relevant information has raised the demand of recommender systems [1].

Recommender systems solve the above problems by filtering huge amount of constantly created data and selecting out only the relevant information based on the user's choices, interests, or observed behaviour with regard to the particular item [2]. With the constant interaction of the user, these systems learn user's behaviour, interests, choices, needs, etc. and create a user-based profile, based on which there is a high probability of getting these recommendations accurate that whether the user would like that recommended product or not. These recommendation systems suggest the products to users based on a variety of parameters depending upon their choices, interests, search history, likes, previous orders, etc. As these are highly developed systems, mostly used and developed by large corporations which are very difficult to crack, hence are also termed as "black box." As these recommendations are the results based on users interest or need, but sometimes the user is unaware of the he/she needs until recommended by these systems [2, 3]. Both service providers and users benefit from the recommender systems. They reduce the expenses incurred by a company. While purchasing a certain thing on an e-commerce application, they reduce the expenses and time consumed in searching and selecting the required products. These systems have also helped in increasing the quality and speed of selecting and purchasing a product, hence increasing the overall decision-making process [4]. There are various techniques which are followed to filter the information based on the type of content, based on the similar tastes of different users termed as collaborative filtering, and the combination of these two termed as hybrid filtering.

2 Phases of Recommendation Process

A recommendation system goes through three phases as shown in Fig. 1 [5]. It collects information and gather data about the user in the first phase. Then, that data about the user are processed through various learning algorithms, where the systems learns

about the users choices and needs and creates a user profile, and then, finally based on that user profile created in the learning phase, the recommendation system suggests items or things of user interest which the user might like [6].

2.1 Information Collection Phase

The recommendation system is a complex software that provides information to the user about the products or services that would be suitable for them, without them having to actively search for them. The system takes the user's preferences and backgrounds and uses this information to provide them with suggestions that are most likely to appeal to them. This has the effect of improving the user experience and has the potential to save users considerable time and energy. The recommendation system has the potential to improve a wide range of user experiences, such as shopping, entertainment, social interactions, and many others. The data collection process in a recommendation system involves first building an initial dataset and then using a variety of techniques to improve the quality of the dataset [7].

2.2 Learning Phase

In a recommendation system, the user's preferences, background information, and other data are used to make recommendations about products or services that are likely to appeal to them. The user's initial input is used to build a dataset, which is used to train the system and improve the quality of the recommendations it provides. The user then interacts with the system to obtain feedback about the recommendations, and the system can be updated based on their feedback. The user continues to interact with the system, obtaining feedback and improving their experience [8].

2.3 Recommendation Phase

The recommendation phase of a recommendation system is the stage where user behaviour is analysed and the best recommendations are generated for the user. This is the most user-focused phase of a recommendation system and revolves around understanding user preferences and behaviour to generate the best recommendations for the user. This has wide-ranging applications in areas such as e-commerce and advertising, where it can be used to generate targeted recommendations for the user. The recommendation phase of a recommendation system consists of two primary areas of focus: user behaviour analysis and feature engineering. In Sect. 3, various recommendation techniques have been discussed [9].

3 Recommendation Filtering Techniques

Each model uses its own type of mechanism to filter out the information and suggest recommendations. It varies from company to company to select a particular filtering technique based on the type of data used in their work [8, 9]. Figure 2 contains chart of types of recommendation filtering techniques.

3.1 Content-Based Filtering Technique

Content-based recommender systems are those that use text data to make recommendations for other content. Most of today's recommendation systems use a broad range of signals and try to find the most relevant content for the user. However, these systems are not always effective because they cannot capture the subtle nuances of the content. For example, a system that recommends movies for a user who loves action movies might not recommend a movie for a user who loves silent movies [10]. One of the most common uses of content-based recommender systems is to provide users with a list of products similar to the product they are browsing. For example, if a user is browsing a shoe store, a recommender system might come up with a list

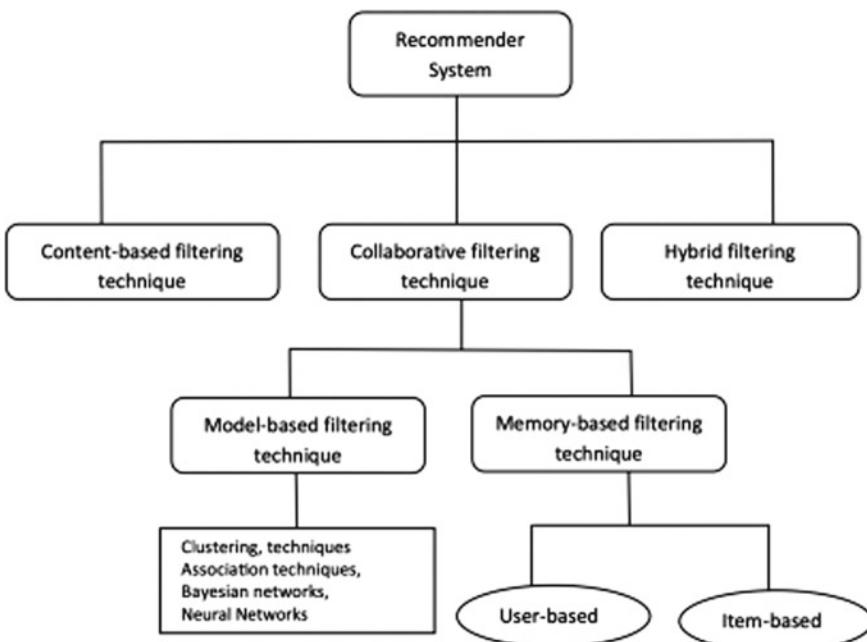


Fig. 2 Recommendation filtering techniques

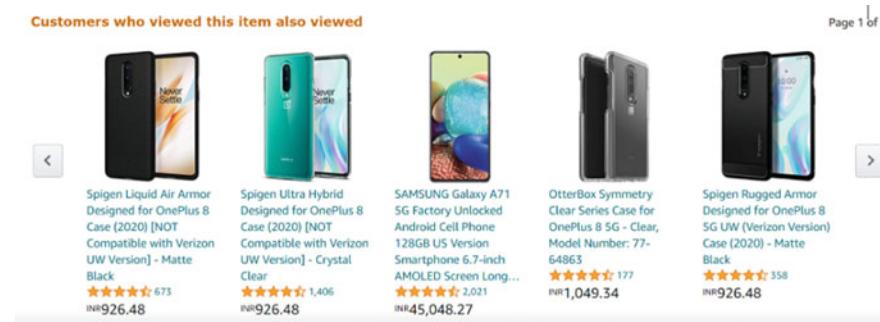


Fig. 3 Recommendations from content-based system on Amazon

of websites that sell running shoes, such as Nike, Adidas, Reebok, and Puma, and might provide the user with a link to one of those websites. The recommender system would use the text from the user's query and the text from the website to create a recommendation, which is a list of web pages that are similar to the original page. These recommender systems are useful for building a large-scale, customised index [11] here as shown in Fig. 3. Amazon uses the search history and recently viewed items to suggest similar items to the user.

Example of Content-based filtering: Assume I am a huge fan of the Harry Potter series and exclusively watch movies about it on the internet. When my information is acquired from Google or Wikipedia, it will be discovered that I am a fantasy film enthusiast. As a result, my list will be dominated by fantasy films, and following that list, best fantasy films will be suggested to me by these tech giants. If there are two films, one of which is Fantastic Beasts, and the other is Shawshank Redemption, then Fantastic Beasts will be recommended to me based on my preference for fantasy films.

3.2 Collaborative Filtering Technique

Collaborative filtering, or crowd-sourcing, is a technique in which users are asked to provide opinions and feedback on the content and quality of a particular piece of information, such as a piece of news or a product. Through this technique, a system can be built that is able to leverage user feedback to provide a more relevant experience for the user [12, 13]. Instead of relying on a single expert to make a decision, collaborative filtering uses a network of users to build a recommendation. The goal is to find the best answer to a question by combining the feedback of multiple users. These predictions are computed by analysing the similarity in tastes of other people who have purchased the same or similar items in the past. The basic idea is that if two people have similar interests and past purchases, then one might be likely to make similar choices in future too.

Customers Who Bought This Item Also Bought

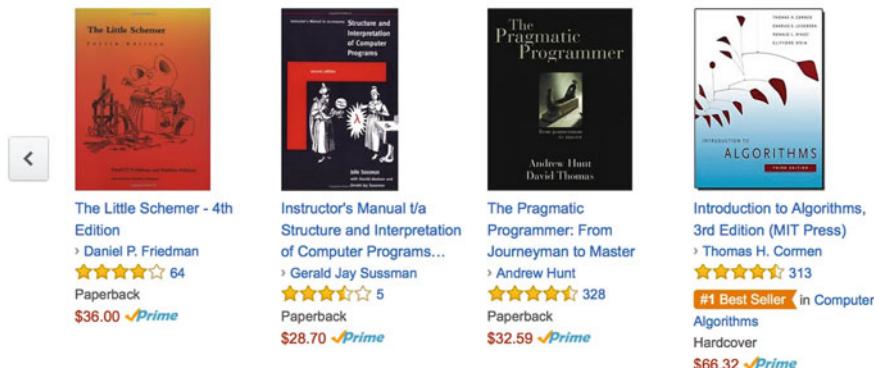


Fig. 4 Collaborative filtering technique used by Amazon e-commerce application

Collaborative filtering makes customised recommendations for users with similar preferences (refer Fig. 4) by filtering data from user reviews using algorithms [14]. It is either memory based or model based [13, 14].

Memory-based technique Here, already reviewed products by the customer play an important role in his/her search for a neighbour who shares common interests with him/her. Once a user's neighbour has been identified, several algorithms can be used to integrate the preferences of the user's neighbours in order to provide suggestions [15, 16]. Both the item-based and user-based techniques can be implemented to create collaborative-based filtering [17].

Model-based technique It is a model-based technique which uses the existing reviews of that product and increases the accuracy of collaborative filtering technique [18, 19]. Because they use a pre-computed model, these strategies may swiftly recommend a list of items, and these models are taught to give recommendations based on the results of neighbourhood-based recommendation techniques [20, 21].

Examples of Collaborative-Based Filtering: (i) YouTube recommends its users videos based on other users who have similar subscriptions or have watched similar videos. (ii) CourseEra course recommendations are based on the results of other people who have completed the same course you have.

Hybrid filtering technique The challenge with collaborative filtering is the huge amount of data required to build a user profile. In order to reduce the amount of data needed, hybrid filtering was developed. This technique combines the predictions made by collaborative filtering with traditional content-based filtering [22]. The power of collaborative filtering is that it is able to make predictions on what a user might want to buy, given their interests and behaviours. This means that it is able to leverage the social data that are available to make more accurate recommendations. However, the limitation of collaborative filtering is that it is unable to take into account the individual preferences of a user. This is because it relies on the user's

social connections to make predictions [23]. So, by using hybrid systems, it covers the shortcoming of collaborative-based model with the help of content-based filtering approach and is able to provide accurate recommendations by using content-based filtering technique when there are not much users of similar interests or choices [23].

Example of Hybrid filtering: Netflix is a corporation that employs a hybrid recommendation system, in which viewers are given recommendations based on the watch and search history of other users, i.e. using of both techniques at the same time, i.e. filtering the movies recommended by collaborative technique with help of content-based approach to provide more filtered and accurate recommendation to the user [24].

4 Best Recommendation

It is a challenge for many businesses to identify what constitutes a good recommendation system. By the term “good” in recommendation refers to evaluate the recommendation system one has created. There are many methods which are used to evaluate these recommendations systems and one way to do so is by calculating the accuracy percentage, as shown below:

$$\text{Precision} = \text{Correctly recommendations item} / \text{Total recommendations item}$$

The method of evaluation of the recommendation system is determined by the type of data and strategy used to generate the suggestions/recommendations [25, 26]. And in an ideal situation, every company wants maximum active user base, so they evaluate their recommendation system and try to monitor how the user reacts to the recommendations provided by their systems and try to improve their systems [27, 28].

5 Conclusion

By solving the problem of information overload, recommender systems have saved lot of time and expenses of both the companies as well as users by immediately suggesting the users their required products or services. Various types of recommendation systems are used by the companies depending upon the type of information filtering technique used. The benefits and problems of two traditional recommendation techniques, content and collaborative, were addressed in this review paper, along with little briefing on hybrid filtering-based recommendation system utilised to improve their efficacy. Taking consideration of the fact that in future the speed with which the volume of data is going to be created, we are going to require systems which will provide recommendations taking in considerations of the dynamic data created at the same time when the request is made from the user. This type of systems

can be built by incorporating the data received from cookies, feedbacks of the newly added and recommended products, including the latest search history, constantly updating the created user-based profile, overall this will increase the reach and fields of the recommender system and enhancing its efficacy.

References

1. Bhatt B, Premal JP, Gaudani H (2014) A review paper on machine learning based recommendation system
2. Debnath S (2008) Machine learning based recommendation system. Master's thesis, Department of Computer Science and Engineering, Indian Institute of Technology
3. Sharda S, Josan GS (2021) Machine learning based recommendation system: a review. *Int J Next-Gener Comput* 12(2)
4. Isinkaye FO, Folajimi YO, Ojokoh BA (2015) Recommendation systems: principles, methods and evaluation. *Egypt Inform* J 16(3):261–273
5. Gallego D, Barra E, Aguirre S, Huecas G (2012) A model for generating proactive context-aware recommendations in e-learning systems. In: 2012 frontiers in education conference proceedings. IEEE, pp 1–6
6. Singh J, Duhan B, Gupta D, Sharma N (2020) Cloud resource management optimization: taxonomy and research challenges. In: 2020 8th international conference on reliability, Infocom technologies and optimization (trends and future directions) (ICRITO). IEEE, pp 1133–1138
7. Van Meteren R, Van Someren M (2000) Using content-based filtering for recommendation. In: Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop, vol 30, pp 47–56
8. Singh J, Gupta D (2017) Energy efficient heuristic base job scheduling algorithms in cloud computing. *IOSR J Comput Eng (IOSR-JCE)*. e-ISSN: 2278-0661
9. Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. *Advances in artificial intelligence*
10. Singh J, Gupta D, Sharma N (2019) Cloud load balancing algorithms: a comparative assessment. *J Comput Theor Nanosci* 16(9):3989–3994
11. Najafabadi MK, Mahrin MNR (2016) A systematic literature review on the state of research and practice of collaborative filtering technique and implicit feedback. *Artif Intell Rev* 45(2):167–201
12. Raghuvanshi SK, Pateriya RK (2019) Collaborative filtering techniques in recommendation systems. Data, engineering and applications. Springer, Singapore, pp 11–21
13. Al-Bashiri H, Abdulgabber MA, Romli A, Kahtan H (2018) An improved memory-based collaborative filtering method based on the TOPSIS technique. *PLoS ONE* 13(10):e0204434
14. Muskan GS, Singh J, Prabha C (2022) Data visualization and its key fundamentals: a comprehensive survey. In: 2022 7th international conference on communication and electronics systems (ICCES), 2022, pp 1710–1714. <https://doi.org/10.1109/ICCES54183.2022.9835803>
15. Stephen SC, Xie H, Rai S (2017) Measures of similarity in memory-based collaborative filtering recommender system: a comparison. In: Proceedings of the 4th multidisciplinary international social networks conference, pp 1–8
16. Singh J, Singh G, Bhati BS (2022) The implication of data lake in enterprises: a deeper analytics. In: 2022 8th international conference on advanced computing and communication systems (ICACCS), vol 1. IEEE, pp 530–534
17. Do MPT, Nguyen DV, Nguyen L (2010) Model-based approach for collaborative filtering. In: 6th international conference on information technology for education, pp 217–228
18. Singh J (2022) Genetic approach based optimized load balancing in cloud computing: a performance perspective. In: 2022 9th international conference on computing for sustainable global development (INDIACOM). IEEE, pp 814–819

19. Thorat PB, Goudar RM, Barve S (2015) Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *Int J Comput Appl* 110(4):31–36
20. Singh J, Singh G, Verma A (2022) The anatomy of big data: concepts, principles and challenges. In: 2022 8th international conference on advanced computing and communication systems (ICACCS), pp 986–990. Conference series, vol 1000, No 1, p 012101. IOP Publishing. <https://doi.org/10.1109/ICACCS54159.2022.9785082>
21. Steck H (2013) Evaluation of recommendations: rating-prediction and ranking. In: Proceedings of the 7th ACM conference on recommender systems, pp 213–220
22. Tintarev N, Masthoff J (2015) Explaining recommendations: design and evaluation. *Recommender systems handbook*. Springer, Boston, MA, pp 353–382
23. Cremonesi P, Garzotto F, Negro S, Papadopoulos AV, Turrin R (2011) Looking for “good” recommendations: a comparative evaluation of recommender systems. In: IFIP conference on human-computer interaction. Springer, Berlin, Heidelberg, pp 152–168
24. Pawar L, Singh J, Bajaj R, Singh G, Rana S (2022) Optimized ensembled machine learning model for IRIS plant classification. In: 2022 6th international conference on trends in electronics and informatics (ICOEI), pp 1442–1446. <https://doi.org/10.1109/ICOEI53556.2022.9776724>
25. Fanca A, Puscasiu A, Gota DI, Valean H (2020) Recommendation systems with machine learning. In: 2020 21th international Carpathian control conference (ICCC). IEEE, pp 1–6
26. Portugal I, Alencar P, Cowan D (2018) The use of machine learning algorithms in recommender systems: a systematic review. *Expert Syst Appl* 97:205–227
27. Goswami A, Dwivedi P, Kant V (2018) Trust-enhanced multi-criteria recommender system. *Soft computing: theories and applications*. Springer, Singapore, pp 439–448
28. Kumar J, Taterh S, Kamnthaia D (2018) Study and comparative analysis of various image spamming techniques. *Soft computing: theories and applications*. Springer, Singapore, pp 351–365

Approximation of Signal Belongs to $W'(L^p, \xi(t))$ Class by Generalized Nörlund-Cesáro Product Means



Smita Sonker and Paramjeet Sangwan

Abstract This paper aims to establish a theorem for approximation of the signal which belongs to conjugate Fourier series using product means. The infinite series, which is not summable by the left linear operator or the right linear operator alone, can be summable by their product. With this motivation, a theorem for approximation of signal (function) belongs to $W'(L^p, \xi(t))$, ($p \geq 1$), ($t > 0$) class by using (N, p_m, q_m) (C, α, η) product means of conjugate Fourier series is established. The existing results are extended and generalized to similar results. The results motivate the scientists to work in the field of approximation theory.

Keywords Conjugate Fourier series · Weighted Lipschitz class · Hölder's inequality · Lebesgue integral · Nörlund-Cesàro product mean

1 Introduction

Let $\sum u_m$ be a given infinite series with the sequence of its m th partial sum $\{s_m\}$. Suppose $p = \{p_m\}$ and $q = \{q_m\}$ are non-increasing, monotonic and positive sequences and given by

$$P_m = \sum_{w=0}^m p_w \rightarrow \infty, \text{ as } m \rightarrow \infty. \quad (1)$$

$$Q_m = \sum_{w=0}^m q_w \rightarrow \infty, \text{ as } m \rightarrow \infty. \quad (2)$$

For all $i \geq 1$, $P_{-1} = p_{-1} = 0$ and $Q_{-1} = q_{-1} = 0$

S. Sonker (✉) · P. Sangwan

National Institute of Technology Kurukshetra, Kurukshetra 136119, India
e-mail: smita.sonker@gmail.com

The convolution product $\{R_m\}$ can be defined as

$$R_m = (p \times q)_m = \sum_{w=0}^m p_{m-w} q_w \rightarrow \infty, \text{ as } m \rightarrow \infty \quad (3)$$

The transformation from sequence-to-sequence

$$t_m^N = \frac{1}{R_m} \sum_{w=0}^m p_{m-w} q_w s_w \quad (4)$$

defines the sequence $\{t_m\}$ of the (N, p_m, q_m) mean of $\{s_m\}$ [1].

If $t_m^N \rightarrow s$, as $m \rightarrow \infty$, then $\sum u_m$ is (N, p_m, q_m) summable to s .

The m th Cesàro means of order (α, η) is denoted by $C_m^{(\alpha, \eta)}$ with $\alpha + \eta > -1$ of the sequence $\{s_m\}$, i.e., (see [2])

$$C_m^{(\alpha, \eta)} = \frac{1}{A_m^{\alpha+\eta}} \sum_{h=0}^m A_{m-h}^{\alpha-1} A_h^\eta s_h, \quad (5)$$

where $A_m^{\alpha+\eta} = O(m^{\alpha+\eta})$, $\alpha + \eta > -1$ and $A_0^{\alpha+\eta} = 1$.

The series $\sum u_m$ becomes (C, α, η) summable to 's' if

$$C_m^{(\alpha, \eta)} = \frac{1}{A_m^{\alpha+\eta}} \sum_{h=0}^m A_{m-h}^{\alpha-1} A_h^\eta s_h \rightarrow s, \text{ as } m \rightarrow \infty. \quad (6)$$

The (N, p_m, q_m) transform of the (C, α, η) transform defines $(N, p_m, q_m) (C, \alpha, \eta)$ transform and we shall denote it by $(NC)_m^{p,q;\alpha,\eta}$. Moreover, if

$$(NC)_m^{p,q;\alpha,\eta} = \frac{1}{R_m} \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \rightarrow s, \text{ as } m \rightarrow \infty. \quad (7)$$

Then we can say that $\sum u_m$ is $(N, p_m, q_m) (C, \alpha, \eta)$ summable to a definite number 's'.

The sufficient and necessary conditions for the (N, p_m, q_m) mean to be regular are

$$\sum_{h=0}^m |p_{m-h} q_h| = O(R_m) \quad (8)$$

and $p_{m-h} = o(R_m)$, as $m \rightarrow \infty$, for every $h > 0$, for which $q_h \neq 0$.

Particular Cases

1. $(N, p_m, q_m)(C, \alpha, \eta)$ becomes $(N, p_m)(C, 1)$ means for $q_m = 1, \alpha = 1, \eta = 0 \forall m$.
2. $(N, p_m, q_m)(C, \alpha, \eta)$ becomes $(\overline{N}, q_m)(C, 1)$ means for $p_m = 1, \alpha = 1, \eta = 0 \forall m$.
3. $(N, p_m, q_m)(C, \alpha, \eta)$ becomes $(C, \delta)(C, 1)$ means for $p_m = \binom{m + \delta - 1}{\delta - 1}, \delta > 0, q_m = 1, \alpha = 1, \eta = 0 \forall m$.
4. $(N, p_m, q_m)(C, \alpha, \eta)$ becomes $(N, p_m, q_m)(C, 1)$ means for $\alpha = 1, \eta = 0 \forall m$.
5. $(N, p_m, q_m)(C, \alpha, \eta)$ becomes $(N, p_m, q_m)(C, 1, \eta)$ for $\alpha = 1 \forall m$.

Suppose a signal denoted by g having 2π as periodic time and is integrable in the same way as of Lebesgue for the limit $(-\pi, \pi)$. Suppose

$$s_m(g; x) = \frac{a_0}{2} + \sum_{h=1}^m (a_h \cos hx) + \sum_{h=1}^m (b_h \sin hx) \quad (9)$$

denotes the $(m + 1)$ th partial sum and is known as trigonometric polynomial of degree m of the Fourier series of g at x and

$$\sum_{h=1}^{\infty} (b_h \cos hx) - \sum_{h=1}^{\infty} (a_h \sin hx) \quad (10)$$

is called its conjugate series and the m th partial sum is given as

$$\overline{s_m}(g; x) = \sum_{h=1}^m (b_h \cos hx) - \sum_{h=1}^m (a_h \sin hx) \quad (11)$$

Definition 1 For g , the L_{∞} -norm and L_p -norm are generally denoted by $\|g\|_{\infty}$ and $\|g\|_p$ -norm, respectively. These norms can be given as

$$\|g\|_{\infty} = \sup \{|g(x)| : x \in R\} \quad (12)$$

$$\|g\|_p = \left\{ \int_0^{2\pi} |g(x)|^p dx \right\}^{1/p}, \quad p \geq 1. \quad (13)$$

Definition 2 The degree of approximation of ' g ' by $t_m(x)$ of order m for $\|\cdot\|_{\infty}$ is given by [3] with

$$\|t_m - g\|_{\infty} = \sup \{|t_m(x) - g(x)| : x \in R\} \quad (14)$$

and $E_m(g)$ of $g \in L_p$ is defined as

$$E_m(g) = \min_m \|t_m(g; x) - g(x)\|_p \quad (15)$$

Definition 3 A signal g of Lip class generally denoted by $g \in \text{Lip}\beta$ if

$$|g(x+t) - g(x)| = O(|t|^\beta), \quad 0 < \beta \leq 1, \quad t > 0$$

and $g \in \text{Lip}(\beta, p)$ if

$$\begin{aligned} \omega_p(t; g) &= \left(\int_0^{2\pi} |g(x+t) - g(x)|^p dx \right)^{\frac{1}{p}} \\ &= O(|t|^\beta) \text{ for } 0 < \beta \leq 1, \quad p \geq 1, \quad t > 0. \end{aligned} \quad (16)$$

For $\xi(t)$, $g \in \text{Lip}(\xi(t), p)$ if

$$\begin{aligned} \omega_p(t; g) &= \left(\int_0^{2\pi} |g(x+t) - g(x)|^p dx \right)^{\frac{1}{p}} \\ &= O(\xi(t)) \text{ for } p \geq 1, \quad t > 0. \end{aligned} \quad (17)$$

And a real valued signal $g \in W'(L^p, \xi(t))$, if

$$\begin{aligned} \omega_p(t; g) &= \left(\int_0^{2\pi} |g(x+t) - g(x)|^p \sin^{\gamma p}(x) dx \right)^{\frac{1}{p}} \\ &= O(\xi(t)) \text{ for } \gamma \geq 0, \quad p \geq 1, \quad t > 0. \end{aligned} \quad (18)$$

We redefine weighted class for our comfort to evaluate $I_{1,2}$ without error as given below

$$\begin{aligned} \omega_p(t; g) &= \left(\int_0^{2\pi} |g(x+t) - g(x)|^p \sin^{\gamma p}\left(\frac{x}{2}\right) dx \right)^{\frac{1}{p}} \\ &= O(\xi(t)) \text{ for } \gamma \geq 0, \quad t > 0. \end{aligned} \quad (19)$$

Notations:

$$\begin{aligned} \psi_x(t) &= g(x+t) - g(x-t) \\ (\overline{NC})_m^{p,q;\alpha,\eta}(t) &= \frac{1}{\pi R_m} \left[\sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{\cos(j+\frac{1}{2})t}{2 \sin \frac{t}{2}} \right] \end{aligned}$$

Summability theory plays an important role in pure and applied mathematics, signal analysis, digital signal processing, vibration analysis, Fuzzy number sequences, etc. The signals are represented by a one-variable function, while two-variable functions represent images. Weierstrass theorem was used in starting to originate the approximation theory, and further, this study was carried out using trigonometric polynomials. Zygmund [3] developed the method of trigonometric approximation of signals for the periodic series. Several researchers studied the approximation of signals or functions by various summability methods like Cesàro, Euler and Nörlund-mean, etc. Various investigators like Rhoades et. al. [4], Bor [5], Lal and Nigam [6], Sonker and Singh [7], Mishra et. al [8], Krasniqi [9] carried out their work on estimation of signals belong to different Lipschitz classes. Giri et. al [10] carried out their study on neural network-based approximation method. Dass and Srivastava [11] highlighted that the complexity and number of rules can be reduced for approximation of higher order systems by using Recurrent Fuzzy System. Mittal et. al [12, 13] obtained various new theorems on approximations of functions by using weighted Lipschitz classes and generalizes the results of various researchers. Sonker and Sangwan [14, 15] recently worked on triple product means. Krasniqi and Deepmala [16], Rathore and Shrivastava [17] worked on approximation of functions by Nörlund-Euler product summability. Rhoades [18] worked on approximation of conjugate function, removed the monotonicity condition and generalized the results of Lal [19]. In the present study, we determine a new theorem for signal approximation which belongs to $W'(L^p, \xi(t))$, $(p \geq 1)$, $(t > 0)$ by $(N, p_m, q_m)(C, \alpha, \eta)$ product operator of conjugate Fourier series that generalizes the result of Rhoades [18], Rathore and Shrivastava [17].

2 Main Theorem

Theorem 1 *If a signal \bar{g} , conjugate to a 2π periodic function, Lebesgue integral and of class $W'(L^p, \xi(t))$, $(p \geq 1)$, $(t > 0)$, then its approximation using $(\bar{N}\bar{C})^{p,q;\alpha,\eta}$ product means of series (10) is*

$$\|t_m^{(\bar{N}\bar{C})^{p,q;\alpha,\eta}}(g; x) - \bar{g}\|_p = O \left((1+m)^{\gamma + \frac{1}{p}} \xi((1+m)^{-1}) \right) \quad (20)$$

$\{\xi(t) \cdot t^{-1}\}$ is a non-increasing sequence,

$$\left(\int_0^{\frac{\pi}{(1+m)}} \left(\frac{|\psi_x(t)|}{\xi(t)} \right)^p \sin^{\gamma p} \left(\frac{t}{2} \right) dt \right)^{\frac{1}{p}} = O(1) \quad (21)$$

$$\left(\int_{\frac{\pi}{(1+m)}}^{\pi} \left(\frac{|\psi_x(t)|}{\xi(t) \cdot t^\delta} \right)^p dt \right)^{\frac{1}{p}} = O \left(\frac{1}{(1+m)^{-\delta}} \right) \quad (22)$$

where δ is a number which is arbitrary such that $(1 - \delta)q - 1 > 0$, $\frac{1}{p} + \frac{1}{q} = 1$, $1 \leq p \leq \infty$, conditions (21) and (22) hold uniformly in x and $(\overline{NC})^{p,q;\alpha,\eta}$ is $(N, p_m, q_m)(C, \alpha, \eta)$ summable of the series (10) and $\overline{g}(x)$ is defined for almost every x by

$$2\pi \overline{g}(x) = - \lim_{\epsilon \rightarrow 0} \int_{\epsilon}^{\pi} \psi_x(t) \cot\left(\frac{t}{2}\right) dt. \quad (23)$$

3 Lemmas

Lemmas required to prove the main theorem are as follows:

Lemma 1 $|\overline{NC}_m^{p,q;\alpha,\eta}(t)| = O\left[\frac{1}{t}\right]$, for $0 < t \leq \frac{\pi}{(1+m)}$; $t \leq \pi \sin\left(\frac{t}{2}\right)$ and $|\cos(mt)| \leq 1$.

Proof

$$\begin{aligned} |\overline{NC}_m^{p,q;\alpha,\eta}(t)| &\leq \frac{1}{2\pi \cdot R_m} \left| \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{\cos\left(j + \frac{1}{2}\right) t}{\sin \frac{t}{2}} \right| \\ &\leq \frac{1}{2\pi \cdot R_m} \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{|\cos\left(j + \frac{1}{2}\right) t|}{|\sin \frac{t}{2}|} \\ &\leq \frac{1}{2\pi \cdot R_m} \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{1}{\frac{t}{\pi}} \\ &= \frac{1}{2t \cdot R_m} \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \\ &= \frac{1}{2t \cdot R_m} \sum_{h=0}^m p_{m-h} q_h \left\{ as \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta = A_h^{\alpha+\eta} \right\} \\ &= O\left[\frac{1}{t}\right]. \end{aligned}$$

Lemma 2 $|\overline{(NC)}_m^{p,q;\alpha,\eta}(t)| = O\left[\frac{1}{t}\right]$, for $0 < \frac{\pi}{(1+m)} \leq t \leq \pi$; $t \leq \pi \sin\left(\frac{t}{2}\right)$.

Proof

$$\begin{aligned}
|\overline{(NC)}_m^{p,q;\alpha,\eta}(t)| &\leq \frac{1}{2\pi R_m} \left| \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{\cos(j + \frac{1}{2})t}{\sin \frac{t}{2}} \right| \\
&\leq \frac{1}{2\pi R_m} \left| \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{\cos(j + \frac{1}{2})t}{\frac{t}{\pi}} \right| \\
&\leq \frac{1}{2t R_m} \left| \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{i(j + \frac{1}{2})t} \right| \\
&\leq \frac{1}{2t R_m} \left| \sum_{h=0}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} |e^{i\frac{t}{2}}| \right| \\
&\leq \frac{1}{2t R_m} \left| \sum_{h=0}^{\tau-1} p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} \right| \\
&\quad + \frac{1}{2t R_m} \left| \sum_{h=\tau}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} \right|. \tag{24}
\end{aligned}$$

Now, consider first term of (24), we have

$$\begin{aligned}
&\frac{1}{2t R_m} \left| \sum_{h=0}^{\tau-1} p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} \right| \\
&\leq \frac{1}{2t R_m} \left| \sum_{h=0}^{\tau-1} p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta |e^{ijt}| \right| \\
&= \frac{1}{2t R_m} \left| \sum_{h=0}^{\tau-1} p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \right| \\
&= \frac{1}{2t R_m} \left| \sum_{h=0}^{\tau-1} p_{m-h} q_h \right| \left\{ \text{as } \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta = A_h^{\alpha+\eta} \right\} \\
&= O\left[\frac{1}{t}\right]. \tag{25}
\end{aligned}$$

Considering second term of (24)

$$\begin{aligned}
& \frac{1}{2t \cdot R_m} \left| \sum_{h=\tau}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \operatorname{Re} \left\{ \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} \right\} \right| \\
& \leq \frac{1}{2t \cdot R_m} \sum_{h=\tau}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \max_{0 \leq k \leq h} \left| \sum_{j=0}^k A_{h-j}^{\alpha-1} A_j^\eta e^{ijt} \right| \\
& \leq \frac{1}{2t \cdot R_m} \sum_{h=\tau}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \max_{0 \leq k \leq h} \sum_{j=0}^k A_{h-j}^{\alpha-1} A_j^\eta |e^{ijt}| \\
& = \frac{1}{2t \cdot R_m} \sum_{h=\tau}^m p_{m-h} q_h \frac{1}{A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \\
& = \frac{1}{2t \cdot R_m} \sum_{h=\tau}^m p_{m-h} q_h \left\{ \text{as } \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta = A_h^{\alpha+\eta} \right\} \\
& = O \left[\frac{1}{t} \right]. \tag{26}
\end{aligned}$$

Collecting (25) and (26), we get

$$|(\overline{NC})_m^{p,q;\alpha,\eta}(t)| = O \left[\frac{1}{t} \right]. \tag{27}$$

4 Proof of the Theorem

According to Zygmund [3], suppose $\overline{s_m}(g; x)$ is partial sum of the series (10) and is written as

$$|\overline{s_m}(g; x) - \overline{g}(x)| = \frac{1}{2\pi} \int_0^\pi \psi_x(t) \frac{\cos(m + \frac{1}{2})t}{\sin \frac{t}{2}} dt. \tag{28}$$

The $(\overline{NC})_m^{p,q;\alpha,\eta}$ transform of $\overline{s_m}(g; x)$ is given by

$$\begin{aligned}
|t_m^{(\overline{NC})^{p,q;\alpha,\eta}}(g; x) - \overline{g}| &= \frac{1}{R_m} \left| \sum_{h=0}^m p_{m-h} q_h \int_0^\pi \frac{\psi_x(t)}{\pi A_h^{\alpha+\eta}} \sum_{j=0}^h A_{h-j}^{\alpha-1} A_j^\eta \frac{\cos(j + \frac{1}{2})t}{2 \sin \frac{t}{2}} dt \right| \\
&= \int_0^\pi |\psi_x(t)| \cdot |(\overline{NC})_m^{p,q;\alpha,\eta}(t)| dt. \tag{29}
\end{aligned}$$

Using assumptions of the theorem,

$$\int_0^\pi |\psi_x(t)| \|\overline{NC}\|_m^{p,q;\alpha,\eta}(t) dt = O\left((1+m)^{\gamma+\frac{1}{p}} \xi((1+m)^{-1})\right). \quad (30)$$

Now,

$$\begin{aligned} |t_m^{(\overline{NC})^{p,q;\alpha,\eta}}(g; x) - \bar{g}| &= \int_0^\pi |\psi_x(t)| \|\overline{NC}\|_m^{p,q;\alpha,\eta}(t) dt \\ &= \left[\int_0^{\frac{\pi}{(1+m)}} |\psi_x(t)| + \int_{\frac{\pi}{(1+m)}}^\pi |\psi_x(t)| \right] \|\overline{NC}\|_m^{p,q;\alpha,\eta}(t) dt \\ &= |I_{1,1}| + |I_{1,2}| \text{ (say).} \end{aligned} \quad (31)$$

Using Hölder's inequality, lemma 1, condition (21) and $(\sin t/2)^{-1} \leq \frac{\pi}{t}$, for $0 < t \leq \pi$,

$$\begin{aligned} |I_{1,1}| &\leq \int_0^{\frac{\pi}{(1+m)}} |\psi_x(t)| \|\overline{NC}\|_m^{p,q;\alpha,\eta}(t) dt \\ |I_{1,1}| &\leq \left(\int_0^{\frac{\pi}{(1+m)}} \left(\frac{|\psi_x(t)|}{\xi(t)} \sin^\gamma(t/2) \right)^p dt \right)^{\frac{1}{p}} \left[\int_0^{\frac{\pi}{(1+m)}} \left(\frac{\xi(t) \|\overline{NC}\|_m^{p,q;\alpha,\eta}(t)}{\sin^\gamma(t/2)} \right)^q dt \right]^{\frac{1}{q}} \\ &= O(1) \operatorname{ess\,sup}_{0 < t \leq \frac{\pi}{(1+m)}} [\xi(t)^q]^{\frac{1}{q}} \left[\int_0^{\frac{\pi}{(1+m)}} (t^{-1-\gamma})^q dt \right]^{\frac{1}{q}} \\ &= O\left(\xi\left(\frac{\pi}{1+m}\right)\right) \operatorname{ess\,sup}_{0 < t \leq \frac{\pi}{(1+m)}} \left[\int_0^{\frac{\pi}{(1+m)}} (t^{-1-\gamma})^q dt \right]^{\frac{1}{q}} \\ &= O\left(\xi\left((1+m)^{-1}\right)\right) \left[\lim_{\epsilon \rightarrow 0} \int_{\epsilon}^{\frac{\pi}{(1+m)}} (t^{-1-\gamma})^q dt \right]^{\frac{1}{q}} \\ &= O\left[\xi\left((1+m)^{-1}\right)(1+m)^{\gamma+1-\frac{1}{q}}\right] \\ &= O\left[(1+m)^{\gamma+\frac{1}{p}} \xi\left((1+m)^{-1}\right)\right] \left\{ \text{as } \frac{1}{p} + \frac{1}{q} = 1, 1 \leq p \leq \infty. \right\} \end{aligned} \quad (32)$$

Using lemma 2, condition (22), $(\sin t/2)^{-1} \leq \frac{\pi}{t}$, for $0 < t \leq \pi$ and $|\sin t/2| \leq 1$,

$$\begin{aligned}
|I_{1.2}| &\leq \int_{\frac{\pi}{(1+m)}}^{\pi} |\psi_x(t)| \left| (\overline{NC})_m^{p,q;\alpha,\eta}(t) \right| dt \\
&\leq \left[\int_{\frac{\pi}{(1+m)}}^{\pi} \left(\frac{\sin^\gamma(t/2) |\psi_x(t)|}{\xi(t) \cdot t^\delta} \right)^p dt \right]^{\frac{1}{p}} \left[\int_{\frac{\pi}{(1+m)}}^{\pi} \left(\frac{\xi(t) \cdot t^\delta |(\overline{NC})_m^{p,q;\alpha,\eta}(t)|}{\sin^\gamma(t/2)} \right)^q dt \right]^{\frac{1}{q}} \\
&\leq \left[\int_{\frac{\pi}{(1+m)}}^{\pi} \left(\frac{|\psi_x(t)|}{\xi(t) \cdot t^\delta} \right)^p dt \right]^{\frac{1}{p}} \left[\int_{\frac{\pi}{(1+m)}}^{\pi} \left(\frac{\xi(t) \cdot t^{\delta-1}}{\sin^\gamma(t/2)} \right)^q dt \right]^{\frac{1}{q}} \\
&= O\left(\frac{1}{(1+m)^{-\delta}}\right) \left[\int_{\frac{1}{\pi}}^{\frac{(1+m)}{\pi}} \left(\xi\left(\frac{1}{z}\right) \cdot z^{-\delta+1+\gamma} \right)^q \frac{dz}{z^2} \right]^{\frac{1}{q}} \left\{ \text{Putting } t = \frac{1}{z} \right\} \\
&= O\left(\frac{1}{(1+m)^{-\delta}} \xi\left(\frac{\pi}{1+m}\right)\right) \left[\int_{\frac{1}{\pi}}^{\frac{(1+m)}{\pi}} z^{(-\delta+1+\gamma)q-2} dz \right]^{\frac{1}{q}} \\
&= O\left(\frac{1}{(1+m)^{-\delta}} \xi\left((1+m)^{-1}\right)\right) \left[(1+m)^{(-\delta+1+\gamma)-\frac{1}{q}} \right] \\
&= O\left(\xi\left((1+m)^{-1}\right)\right) \left[(1+m)^{\gamma+1-\frac{1}{q}} \right] \\
&= O\left((1+m)^{\gamma+\frac{1}{p}} \xi\left((1+m)^{-1}\right)\right) \cdot \left\{ \text{as } \frac{1}{p} + \frac{1}{q} = 1, 1 \leq p \leq \infty. \right\} \tag{33}
\end{aligned}$$

Putting Eqs. (32), (33) in (31), we get the proof of theorem

$$|t_m^{(\overline{NC})^{p,q;\alpha,\eta}}(g; x) - \bar{g}| = O\left((1+m)^{\gamma+\frac{1}{p}} \xi\left((1+m)^{-1}\right)\right).$$

Thus

$$\begin{aligned}
\|t_m^{(\overline{NC})^{p,q;\alpha,\eta}}(g; x) - \bar{g}\|_p &= \left(\int_0^{2\pi} |t_m^{(\overline{NC})^{p,q;\alpha,\eta}}(g; x) - \bar{g}|^p dx \right)^{\frac{1}{p}} \\
&= O\left(\int_0^{2\pi} \left((1+m)^{\gamma+\frac{1}{p}} \xi\left((1+m)^{-1}\right) \right)^p dx \right)^{\frac{1}{p}} \\
&= O\left((1+m)^{\gamma+\frac{1}{p}} \xi\left((1+m)^{-1}\right)\right).
\end{aligned}$$

5 Corollaries

Corollary 1 In theorem 1, if $\gamma = 0$, $W'(L^p, \xi(t))$, ($p \geq 1$), ($t > 0$) reduces to $Lip(\xi(t), p)$ then its approximation of signal is given by

$$\|t_m^{(NC)}\|^{p,q;\alpha,\eta}(g; x) - \bar{g}\|_p = O\left((1+m)^{\frac{1}{p}}\xi\left((1+m)^{-1}\right)\right). \quad (34)$$

Corollary 2 In theorem 1, if $\gamma = 0$ and $\xi(t) = t^\beta$, $0 < \beta \leq 1$ then its approximation of signal $g \in Lip(\beta, p)$, $\frac{1}{p} \leq \beta \leq 1$ is given by

$$\|t_m^{(NC)}\|^{p,q;\alpha,\eta}(g; x) - \bar{g}\|_p = O\left((1+m)^{-\beta+\frac{1}{p}}\right). \quad (35)$$

Corollary 3 If $\gamma = 0$, $\xi(t) = t^\beta$ for $0 < \beta < 1$ and if $p \rightarrow \infty$ in corollary 2, $g \in Lip(\beta, p)$ reduces to $Lip\beta$, then its approximation of signal $g \in Lip\beta$ is given by

$$\|t_m^{(NC)}\|^{p,q;\alpha,\eta}(g; x) - \bar{g}\|_\infty = O\left((1+m)^{-\beta}\right). \quad (36)$$

6 Applications

Summability theory has applications in a variety of fields, including pure and applied mathematics, neural networks, signal analysis, signal theory (as a double digital filter), digital signal processing, vibration analysis, fuzzy number sequences and so on. The Fourier transform is a useful function for approximating other functions, similarly neural networks can be thought of a function approximation approach.

7 Conclusion

Theory of approximation can be applied to deal with many problems in various disciplines, such as computer science, mathematical physics, electronics, mathematical analysis and other branches of engineering. The infinite series, which is not summable by the linear operator alone, become summable by the product of the linear operators. In this study, a quite new theorem on approximation of signal belongs to weighted Lipschitz class by Nörlund-Cesàro product means is established. The main results are generalized and extended to similar results.

Acknowledgements The authors are thankful to the Science and Engineering Research Board.

References

1. Borwein D (1958) On product of sequences. *J London Math Soc* 33:352–357. <https://doi.org/10.1112/jlms/s1-33.3.352>
2. Hardy GH (1949) Divergent series, 1st edn. Oxford University Press
3. Zygmund A (1959) Trigonometric series, 2nd edn. Cambridge Univ, Press
4. Rhoades BE, Ozkoklu, Albayrak KI (2011) On degree of approximation to functions belonging to the Lipschitz class by Hausdroff means of its Fourier series. *Appl Math Comput* 217(1):6868–6871. <https://doi.org/10.1016/j.amc.2011.01.034>
5. Bor H (1985) On two summability methods. *Math Proc Camb Philos Soc* 97:147–149. <https://doi.org/10.1017/S030500410006268X>
6. Lal S, Nigam HK (2001) Degree of approximation of conjugate of a function belonging to $Lip(\xi(t), p)$ class by matrix summability means of conjugate Fourier series. *Int J Math Math Sci* 27:555–563. <https://doi.org/10.1155/S0161171201005737>
7. Sonker S, Singh U (2012) Degree of approximation of the conjugate of signals (functions) belonging to $Lip(\alpha, r)$ -class by $(C, 1)(E, q)$ means of conjugate trigonometric Fourier series. *J Inequalities Appl* 2012(1):278. <https://doi.org/10.1186/1029-242X-2012-278>
8. Mishra VN, Khatri K, Mishra LN (2012) Approximation of functions belonging to $Lip(\xi(t), r)$ class by $(N, p_n)(E, q)$ summability of conjugate series of Fourier series. *J Inequal Appl* 2012(296):1–10. <https://doi.org/10.1186/1029-242X-2012-296>
9. Krasniqi XZ (2015) On the degree of approximation of functions belonging to the Lipschitz class by $(E, q)(C, \alpha, \beta)$ means. *Khayyam J Math* 1(2):243–252. <https://doi.org/10.22034/kjm.2015.13168>
10. Giri JP, Giri PJ, Chadge R (2018) Neural network-based prediction of productivity parameters. In: Pant M, Ray K, Sharma T, Rawat S, Bandyopadhyay A (eds) Soft computing: theories and applications. Advances in intelligent systems and computing, vol 583. Springer, Singapore. https://doi.org/10.1007/978-981-10-5687-1_8
11. Dass A, Srivastava S (2018) On comparing performance of conventional fuzzy system with recurrent fuzzy system. In: Pant M, Ray K, Sharma T, Rawat S, Bandyopadhyay A (eds) Soft computing: theories and applications. Advances in intelligent systems and computing, vol 583. Springer, Singapore. https://doi.org/10.1007/978-981-10-5687-1_35
12. Mittal ML, Rhoades BE, Mishra VN (2006) Approximation of signals (functions) belonging to the weighted $W(L_{tp}, \xi(t), (p \geq 1))$ class by linear operators. *Int J Math Math Sci* 5353:1–10. <https://doi.org/10.1155/IJMMS/2006/53538>
13. Mittal ML, Rhoades BE, Sonker S, Singh U (2011) Approximation of signals of class $Lip(\alpha, p)$ by linear operator. *Appl Math Comput* 217(9):4483–4489. <https://doi.org/10.1016/j.amc.2010.10.051>
14. Sonker S, Sangwan P (2021) Approximation of Fourier and its conjugate series by triple Euler product summability. *J Phys: Conf Ser* 1770(012003):1–10. <https://doi.org/10.1088/1742-6596/1770/1/012003>
15. Sonker S, Sangwan P (2021) Approximation of signals by harmonic-Euler triple product means. *J Indian Math Soc* 88(1–2):176–186. <https://doi.org/10.18311/jims/2021/26084>
16. Krasniqi XZ, Deepmala (2020) On approximation of functions belonging to some classes of functions by $(N, p_n, q_n)(E, \theta)$ means of conjugate series of its Fourier series. *Khayyam J Math* 6(1):73–86. <https://doi.org/10.22034/kjm.2019.97173>
17. Rathore HL, Shrivastava UK (2012) Approximation of function belonging to $W(L_p, \xi(t))$ class by $(E, q)(N, p_0 n)$ means of its Fourier series. *Int J Sci Res Publ* 6(2):1–10
18. Rhoades BE (2002) On the degree of approximation of the conjugate of a function belonging to the class by matrix means of the conjugate series of a Fourier series. *Tamkang J Math* 33(4):365–370. <https://doi.org/10.5556/j.tkjm.33.2002.285>
19. Lal S (2000) On the degree of approximation of conjugate of a function belonging to weighted class $W(L_p, \xi(t))$ by matrix summability means of conjugate series of a Fourier series. *Tamkang J Math* 31:279–288. <https://doi.org/10.5556/j.tkjm.31.2000.385>

Dignet: A Deep Learning-Based Efficient Digit Recognition System



Debashish Mondal, Narinder Kumar, and Ravinder Kaur

Abstract Character recognition through handwritten documents' images has got more interest in the pattern recognition community, due to its associated real-time applications. For handwritten digit recognition, depending on a classification technique, character recognition and extraction of features are performed. For the recognition of handwritten digits, previous techniques lacked high accuracy and processing speed. The proposed model aims to make the route to digitization more obvious by delivering high accuracy and speed. The system of digit recognition is implemented using a convolutional neural network (CNN) with a rectified linear unit (ReLU) activation function. The proposed CNN based architecture is well developed for MNIST digit classification accuracy with appropriate parameters. The system is also evaluated for different layers of CNN. Handwritten digits are recognized using computational methods. For handwritten digit recognition, the current study used a neural network using convolutions as a classifier, MNIST as a set of data with appropriate training and assessment criteria, and an ensemble model in combination with data augmentation technique. The approach achieves a level of accuracy of 99.70%, which is greater than previously presented techniques.

Keywords Handwritten digit recognition · CNN · Adam optimizer · Data augmentation

1 Introduction

Artificial Intellect (AI) is the emulation of human intelligence by machines, and these cognitive processes are linked to the ability to learn, reasoning, identity, and detect [1]. Handwritten digit recognition is a difficult subject that academics have been studying with the use of machine learning methods. HDR is designed to accept and understand handwritten data. In the form of images or documentation extraction of text from genuine photographs, but in the other side, a complex task, because of

D. Mondal · N. Kumar · R. Kaur (✉)
Department of CSE, Chandigarh University, Mohali, India
e-mail: dravinder2920@gmail.com

the vast differences—font size and shape, surface, and color background knowledge, among other things. Handwritten digit recognition is frequently working in a range of innovation domains, including the finance company actual evidence, fully automated registration plate identification, the mailing address confirmation from the inside of mailers, and identity card and postal codes' recognition [2]. MNIST dataset has contributed in the improvement of pattern matching investigation. Another of them may be MNIST, which is extensively used as a standards for pattern recognition applications. On the MNIST dataset, several classifiers are being used; they are basically restricted Boltzmann machines (RBMs) and neural networks attempted earlier [3]. Deep learning is a rapidly expanding field among various machine learning models for achieving improved performance in the domain of character identification and pattern recognition [4]. It employs a layered design, each neuron is a mathematical function, and each layer is a mathematical function. A weight-bearing device between input and output, there are layers. Layer that is not visible is those that can be modified based on your preferences. Concentrating on the task convolutional neural network (CNN) is used to classify images [5].

Several algorithms are presented to label one of the most back-breaking problems in the domain of digit recognition; because numbers written through hands may be written in a variety of practice and adaptations, analyzers encounter various obstacles in self-acting the detection of handwritten digit [6]. The MNIST dataset of handwritten drawings was used for feature extraction, training, and classification. Arora [7] Feedforward and convolutional neural networks were employed as architectures. Activation refers to the input image and the value of each neurons' carries. The output layer is the final layer comprising brain cells which have a large range of target classes; in the instance of handwritten digit recognition, there are ten categories with probability values ranging from 0 to 9 [8]. When it refers to handwritten digit recognition, the result indicated that CNN outperforms FFNN. The digit recognition precision for CNN is 95.63%, while the digit classification accuracy of FWNN is 90%. On the MNIST dataset, Ghosh et al. [8] conducted a comparison of deep neural networks (DNNs), deep belief networks (DBNs), and CNN. CNN's categorized digit accuracy is $> 98\%$, according to work, with minor mistake rates. Rani [9] focused on investigation into the recognition of Kannada, one among the most popular extensively South Indian scripts which was utilized. One of the deep CNN models is utilized in the training of character image samples with an accuracy of 92%. This paper aims to train the model on CNN with a random filter maps at convolutional layer max pool which is used to reduce feature map.

2 Literature Review

Younis and Alkhateeb [11] proposed deep neural network (DNN) models for detecting and recognizing handwritten character recognition using a very popular and benchmarked dataset MNIST dataset. Models were able to extract features with a high 98.46% accuracy without preprocessing of data. Another work on MNIST

dataset was taken by Dutt and Dutt [12]. They showed that multilayer CNN utilizing Keras and Theano libraries was able to achieve 98.7% accuracy. Another article by Ghosh and Maghari [13] compared three neural network methods and concluded that DNN produced the most accurate one with 98.08% accuracy. Abu Ghosh [14] also evaluate neural network for OCR recognition and compared CNN, DNN, and DBN techniques. As per their findings, DNN accuracy outperforms other neural networks; however, it lags behind CNN in terms of execution time. Hou [22] the experimental findings reveal that the combined depth network's recognition impact is clearly superior to that of a single network. The integrated network achieves a more accurate recognition result of 99.55%. It was further discovered that CNN is the best classifier if compared with SVM, KNN, RFC for HDR. The work of HDR is performed in this research by the application of CNN, with RELU activation function that is capable to capture nonlinearity in the data. It provide excellent precision and a short computation time. Using convolution neural network techniques, we hope to finish this. In experiments, the network was trained using the MNIST dataset which consists of 70000 handwritten images of fixed size (28×28) picture, and the digits are centered. These methods are used to figure out how accurate these numbers are categorized. In terms of high accuracy, the CNN classification presented for handwritten digit recognition appears to be superior to previous algorithms utilized for handwritten digit identification.

Higher and more exact results were produced using this method (CNN-based ensemble model for HDR). Though the initial objective is to construct a model that can detect numbers, the model may be expanded to letter and subsequently to the handwriting of a person.

3 Convolutional Neural Network Structure

A CNN is an artificial neural network with an input layer, an output layer, and a number of hidden layers. The system is of repeating levels of convolutional and streaming that make up the hidden layer which eventually leads to one or more fully linked layers. The CNN's architecture is discussed on depth in the subsequent sections, as well as in Fig. 1.

3.1 Convolutional Network Layer

It has been the most basic layer of a CNN that retains its characteristics at an input picture during scanning by using vertical and horizontal slide filters to cover the whole region. It then applies a bias to each zone before evaluating both the filter parameters, and the image sections are combined into a scalar product. Rectified linear unit is used to apply component-wise activation functions that is maximum (0, x), sigmoid, and tan-h as a result of this layer's output for optimal value.

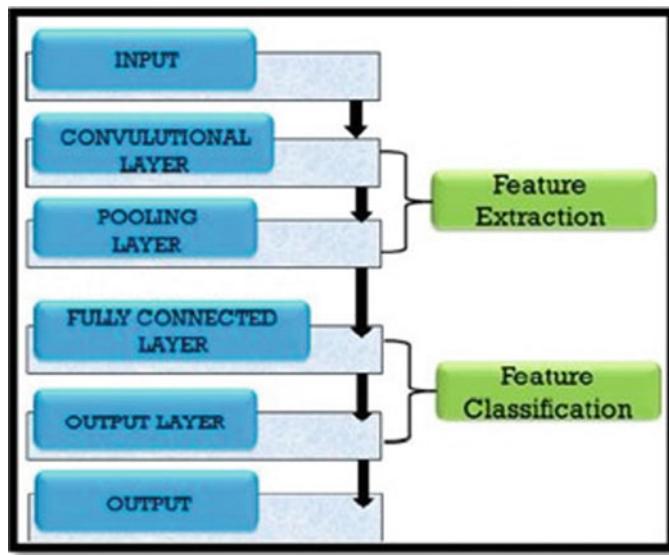


Fig. 1 Brief flow graph of the CNN

3.2 Pooling Layer

The layer is also known as the max pooling layer. The volume of data shrinks at the pooling layer, allowing for easier and quicker network computing. The major tools for implementing pooling are maximum and average pooling. This layer decreases the volume of data by obtaining the greatest by using vertical and lateral sliding filters on the original image, you may get a quality and an for each segment of the input data, the average value as illustrated in Fig. 2.

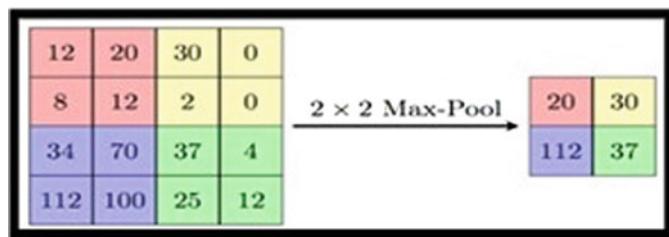


Fig. 2 Reducing the dimension of the image in pooling layer

3.3 Dense Layer

Finally, in a normal neural network with n numbers of neurons, there is a completely linked layer after convolution and pooling, where ‘ n ’ belongs to number of predicted classes. In the digit character classification issue, for example, there are ten neurons for ten classes (0–9).

4 Methodology

The features’ extraction approach is accompanied by a character recognition algorithm based on classification on the features, which is a key justification for OCR from images. For OCR, numerous techniques for feature categorization and extraction were previously used, but with the CNN in deep learning, requirement of different traditional algorithms has come to an end. The DNN architecture is nonlinear system with lots of hidden layers and having large number of links and parameters requiring small number of data modeling challenging process. In CNN, as there is requirement of a small number of parameters, it is easy to train the system. CNN is the only technique capable of accurately mapping datasets for both input and output by altering the trainable parameters and number of hidden layers [14]. As a result, the CNN architecture with ZFNet which is having ten hidden layers in combination with data augmentation approach is deemed the best match for character identification from handwritten digit pictures in this study. The normalized standard MNIST dataset is used for the tests and verification of the system’s performance (Fig. 3).

4.1 MNIST Dataset

A 28 by 28 pixel images are input to the network of CNN; hence, the input layer has 784 neurons. The input pixels are grayscale values, with 0 representing white and 1 representing black, and a variety of monochromatic input images over the range of zero to one are assigned to expose the photographs to shadowiness. The system’s

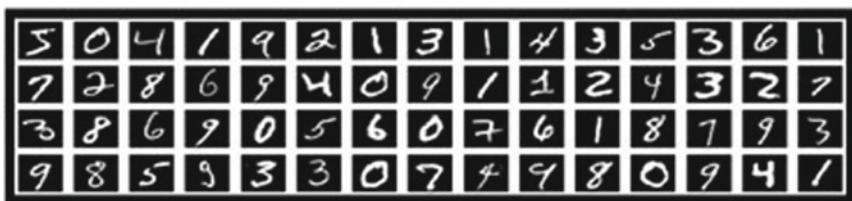


Fig. 3 Sample handwritten digit recognition dataset

output layer consists of ten neurons that represent the digits zero (0) through nine (9). Because the number of output neurons ranges from 0 to 9, the digit is determined by the neuron with the highest activity value.

4.2 *Design for CNN Structure*

The parameters employed in the network determine a CNN's performance in a specific application. Basically, they are feature extraction and feature classification. Every layer of the network in the feature extraction unit gathers the classification unit and, on the other hand, develops the expected outputs by taking the last layer's output, as well as forwarding it to the next layer as inputs. As demonstrated in Fig. 4, for MNIST digit recognition, a CNN architecture with convolutional layers is utilized. The recommended network's general architecture is listed below. Initially, a convolutional layer with a filter map of size 5 receives (28×28) one picture as input and outputs a feature map of form $(24 \times 24)20$. Following that, a pooling layer is used. The resolution of features is reduced by the pooling layers. The downsampling procedure decreases the output size from $(24 \times 24)20$ to $(12 \times 12)20$ by employing a pool size of (2×2) with a stride of 2. Following that, ReLU activation was carried out. The activation function was the ReLU or rectified linear unit. There are many different activation functions to select from whenever neural network models are being trained. Most commonly, the activations' sigmoid, tan-h, ReLU, and Softmax are commonly utilized. Hence, it (a) facilitates the development of bigger NN and (b) is a rapid and effective activity that enables reduce problems of vanishing gradients in neural networks. It eliminates major drawbacks' numbers based on the output, which ensures that the level sizes in the input and output are the same. A filter map of size 5 is employed in the nonlinearity function. It takes an image (12×12) as input and outputs a feature map with shape (6×6) . The max pooling function is used on a layer in conjunction with the ReLU layer, which assists in the establishment of feature assumptions while reducing over fitting and training time. The ReLU's output is sent to the max pooling layer, which gradually lowers the dimensions of the feature map depictions, lowering the overall amount of inputs and calculations as in system. With it aid of it max filter, max pooling conducts overfitting on the linearized convolved results. Data augmentation is the method utilized to solve this challenge. Data augmentation is a method of modifying data while keeping the data's essence. In this investigation, data augmentation is required. Finally, to obtain network results for ten digits (0–9), other completely linked ten-fold layer (10), the extracted features are employed.

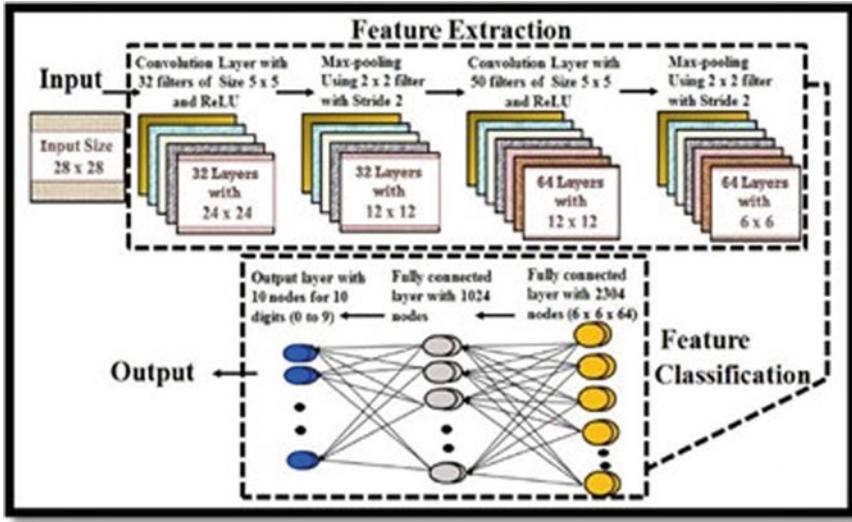


Fig. 4 Proposed methodology

4.3 Mathematical Computation

Every input node in the convolution layer's convolution operation extracts input characteristics pictures. The input nodes, in the max pooling layer's average or maximum operation the image's features should be abstracted. Final results of the S-1st level is used as that of the nth layer's insight, and also, the inputs are then passed via a set of kernels before being sent through the nonlinear function ReLU (f). Consider X^{s-1} inputs from the 1st layer of $s-1$. Consider X^{s-1} are the kernel of layer 1. The convolution operation may therefore be expressed as follows:

$$X_j^8 = f\left(X_i^{s-1} * X_{i,j}^{s-1}\right) + b_j^8 \quad (1)$$

In pooling layer, a $2 * 2$ downsampling kernel is used, with halving each output dimension. It is given as:

$$X_j^8 = \text{down}(X_i^{s-1}) \quad (2)$$

In contrast to classic neural networks, CNN propagates lower-level layer characteristics to create higher-level layer features (NN). Feature propagation reduced the dimension of the feature, based on the size of the convolution and pooling masks. However, for improved classification accuracy, High feature encoding. Appropriate characteristics of the input photos can be selected. As inputs, the result of CNN's last layer is used by the fully connected layers. To provide classification results, the Softmax technique is commonly used.

5 Results

All tests are run on a 2.2 GHz processor of Intel having a RAM of 8 GB, and the programming language utilized is Python. The MNIST dataset was used in all of the tests in this paper. The system was trained on 60,000 photos into the MNIST learning database, as well as it was evaluated on 10000 images from the testing dataset. Table 1 shows how many test photos every digit's (0–9) were utilized in the recognition process (Fig. 5).

Table 1 Number of test photos used and the accuracy with which they were recognized (percent)

Digits	Testing data	Incorrect classified images	Correctly classified images	Accuracy (%)
0	974	6	968	99.38
1	1126	9	1117	99.20
2	1009	23	986	97.72
3	974	35	939	96.40
4	961	21	940	98.00
5	878	14	864	98.40
6	937	21	916	97.75
7	1003	25	978	97.50
8	916	38	878	96.0
9	974	35	939	96.40

```

Model: "sequential"
-----  

Layer (type)          Output Shape       Param #
conv2d (Conv2D)      (None, 28, 28, 10)  260
conv2d_1 (Conv2D)     (None, 28, 28, 10)  2510
max_pooling2d (MaxPooling2D) (None, 14, 14, 10)  0
flatten (Flatten)     (None, 1960)        0
dense (Dense)         (None, 64)          125504
dropout (Dropout)     (None, 64)          0
dense_1 (Dense)       (None, 10)          650
-----  

Total params: 128,924
Trainable params: 128,924
Non-trainable params: 0
  
```

Fig. 5 Model summary

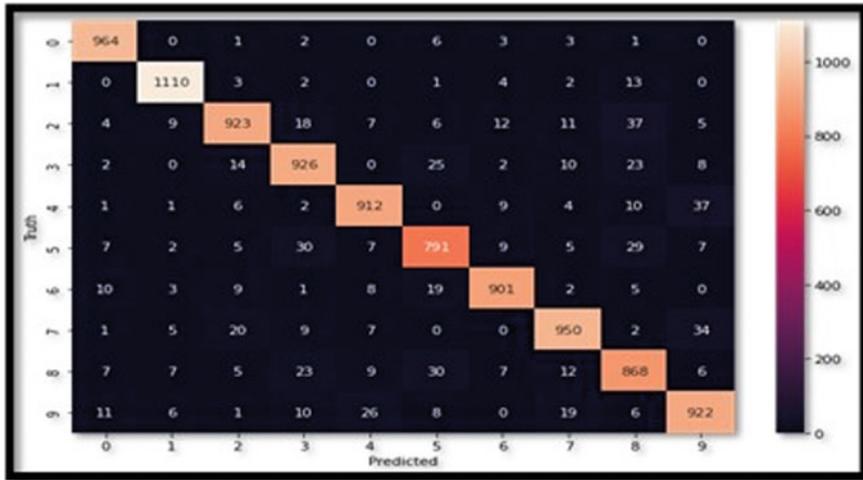


Fig. 6 Heat map with less number of hidden layer

The explanation for this variation is that when collecting the testing dataset, we used varied writing styles and digit sizes. As a result, the recommended methods achieved an accuracy of 99.70%. A larger dataset is recommended to boost the accuracy rate and lower the error. These are in contrast to findings, and the researchers predicted that expanding CNN's many layers will result in high rates of accuracy for digits created by hand. As a consequence, the hypothesis was tested using ten CNN layers. In addition to such findings, the researchers expected that escalating CNN layers will yield in higher precision rates. As a result, ten CNN layers were used to test the hypothesis.

In Fig. 6, the heat map shows the prediction of getting the number guessed correctly or not with using less number of hidden layer. As we can see that number 2 got predicted wrongly as 8, 37 times out of 923. In Fig. 7 of the heat map, we have reduce the error rate of being prediction of 2 as 8 from 34 to only 4 by applying more number of hidden layers, which also increases our accuracy. From Table 2, the author presented a comparison table with the existing works with the proposed model. Clearly can say that the proposed methodology by the author is giving higher accuracy in comparison with other author's results (Fig. 8).

6 Conclusion

The study of artificial intelligence and computer vision relies heavily on character recognition. The proposed model aims to make the route to digitization more obvious by delivering high accuracy and speed. The system of digit recognition is implemented using a convolutional neural network (CNN) with a rectified linear unit

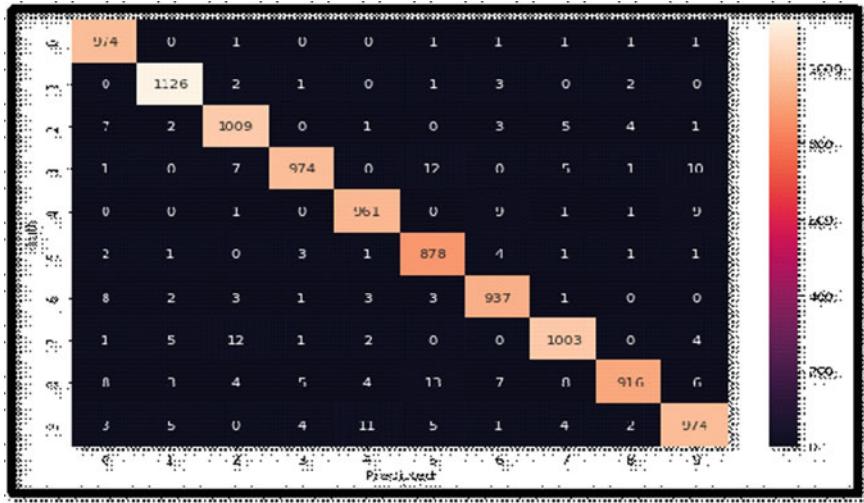


Fig. 7 Heat map with hidden layer = 10

Table 2 Accuracy comparison with other authors with the proposed Dignet model

Author(s)	Method	Accuracy (%)
Younis and Alkhateeb [12]	CNN	98.46
Anuj Dutt, Aashi Dutt [13]	CNN (Keras with Theano)	98.7
Ghosh MMA and Maghari [14]	DNN	98.08
Lee et al. [18]	LeNet-4-CNN	98.9
Sahuand et al. [19]	SVM	98.6
Katiyar and Mehfuz [16]	Adaptive MLP	98.30
Siddique et al. [21]	ANN	97.32
Hossain and Ali [23]	CNN	99.15
Enriquez et al. [24]	CNN	98
Chakraborty et al. [11]	CNN	99
Zhao and Liu [22]	CNN and multi-level fusion	98
Ali and Shaukat [20]	CNN + DL4J	99.21
Proposed work	CNN and ZFNet	99.70

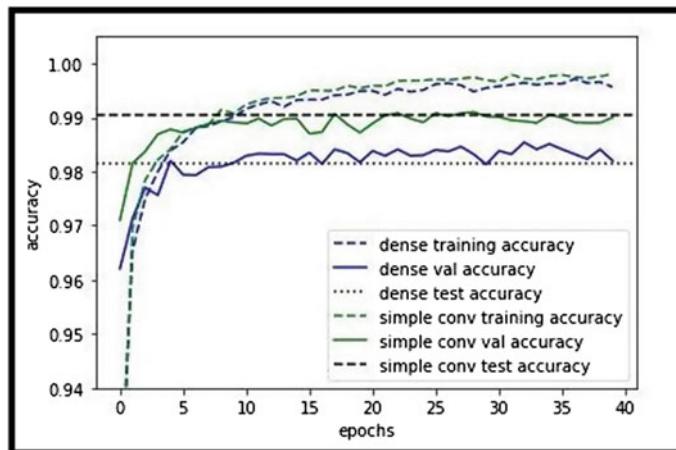


Fig. 8 Accuracy prediction curve

(ReLU) activation function. The proposed CNN-based architecture is well developed for MNIST digit classification accuracy with appropriate parameters. The system is also evaluated for different layers of CNN. Handwritten digits are recognized using computational methods. It is worth noting that the CNN architecture is made up of convolutional layers with two kernels, one with 32 filters and a window size of 5×5 and the other with 64 filters and a window size of 7×7 . When compared to earlier presented systems, the experimental results show that the proposed model CNN architecture for the MNIST dataset performs well in terms of speed and accuracy. With this result, handwritten numerals are accurately identified with an accuracy of 99.70%.

References

1. Yin Y, Zhang W, Hong S, Yang J, Xiong J, Gui G (2019) Deep learning-aided OCR techniques for Chinese uppercase characters in the application of internet of things. *IEEE Access* 7:47043–47049
2. Bušta M, Neumann L, Matas J (2017) Deep TextSpotter: an end-to-end trainable scene text localization and recognition framework. In: IEEE international conference on computer vision (ICCV). Venice, 22–29 Oct 2017, pp 2223–2231
3. Lee S, Son K, Kim H, Park J (2017) Car plate recognition based on CNN using embedded system with GPU. In: 10th international conference on human system interactions (HSI). Ulsan, 17–19 July 2017, pp 239–241
4. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions, 1, 3. [arXiv:1409.4842](https://arxiv.org/abs/1409.4842)
5. Cubuk ED, Zoph B, Shlens J, Le QV Randaugment: practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*
6. Boukharouba ABA (2017) Novel feature extraction technique for the recognition of handwritten digits. *Appl Comput Inform* 13(1):19–26

7. Arora S, Bhatia MPS (2018) Handwriting recognition using deep learning. In: Keras, international conference on advances in computing, communication control and net- working (ICACCCN2018), vol 18, pp 142–145
8. Vasanthi R, Jayavadiel R, Prasad K, Vellingiri J, kilarasu GA, Sudhakar S, Balasubramaniam PM (2020) A novel user interaction middleware component system for ubiquitous soft computing environment by using the fuzzy agent computing system. *J Ambient Intell Humanized Comput* (2020), Springer. <https://doi.org/10.1007/s12652-020-01893-4>
9. Ghosh MMA, Maghari AY (2017) A comparative study on handwriting digit recognition using neural networks. In: IEEE, (37s)
10. Shobha Rani N, Chandan N, Sajan Jain A, Kiran HR (2018) Deformed character recognition using convolutional neural networks. *Int J Eng Technol* 7(3):1599–1604
11. Chakraborty S, Paul S, Sarkar R, Nasipuri M (2019) Feature map reduction in CNN for handwritten digit recognition. Springer, Singapore, pp 143–148(40s)
12. Younis KS, Alkhateeb AA (2017) A new implementation of deep neural networks for optical character recognition and face recognition. In: Proceedings of the new trends in information technology, Jordan, Apr 2017, pp 157–162
13. Dutt A, Dutt A (2017) Handwritten digit recognition using deep learning. *Int J Adv Res Comput Eng Technol* 6(7):990–997
14. Ghosh MMA, Maghari AY (2017) A comparative study on handwriting digit recognition using neural networks. In: International conference on promising electronic technologies, pp 77–81
15. Ghosh MMA, Maghari AY (2017) A comparative study on handwriting digit recognition using neural networks. In: 2017 international conference on promising electronic technologies (ICPET), Deir El-Balah, pp 77–81
16. Katiyar G, Mehfuz S (2016) A hybrid recognition system for offline handwritten characters. Springerplus 5:357
17. Aly S, Mohamed A (2019) Unknown-length handwritten numeral string recognition using cascade of PCA-SVMNet classifiers. *IEEE Access* 7:52024–52034
18. Lee S-W Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network. *IEEE Trans Pattern Anal Mach Intell* 18(6):648–652
19. Sahuand MK, Dewangan NK (2017) A survey on handwritten character recognition. *Int Adv Res Sci Eng Technol.* 10.17148/IARJS ET.2017.4120
20. Ali S, Shaukat Z (2019) An efficient and improved scheme for handwritten digit recognition based on convolutional neural network
21. Siddique MAB, Khan MMR, Arif RB, Ashrafi Z (2018) Study and observation of the variations of accuracies for handwritten digits recognition with various hid- den layers and epochs using neural network algorithm. In: 4th international confer- ence on electrical engineering and information and communication technology, pp 118–123
22. ICIBMS 2017, Track2: Artificial intelligence, robotics and human-computer interaction, Okinawa, Japan Handwritten digit recognition based on depth neural network Yawei Hou School of electrical and electronic engineering Shanghai insti tute of technology Shanghai
23. Hossain MA, Ali MM (2019) Recognition of handwritten digit using convolutional neural network (CNN). *Glob J Comput Sci Technol D Neural Artif Intell* 19:27–33
24. Enriquez EA, Gordillo N, Bergasa LM, Romera E, Huelamo CG (2019) Convolutional neural network vs traditional methods for offline recognition of handwritten digits, vol 855. Springer, Basel, pp 87–99

Voice Command Automation System (VCAS) for Controlling Electrical Devices Using Arduino



Maliha Rahman, Abdullah Al Farabe, Md. Rayhan Al Islam, Moshiur Rahman, Md. Rezyuan, and Ghalib Ashraf

Abstract Time has changed the nature of the human race, and people need their movement to be exceptionally quick, resolute, capable, and rehashed. A person cannot act faultlessly, that is why machines and robots came to give this combination and desire a lively face. Controlling machinery through voice-controlled commands took the century to its peak. This paper will reflect on the facts of home automation using Arduino through voice control. Here, the proposed model of ours is satisfactorily enabled to operate according to the voice commands throughout the experimental sessions. Before that, it could verify the human presence in a room which is a developed detection model that controls the home automation system (HAS) and whether electric devices need to be turned off or not. The presence of a human in a room will work as an activator of the device to take any voice command to switch on any electrical devices in that room. Results of the sessions show that in 80.51% of situations, it received the given voice-controlled instructions, and the device also works at a distance of 25 m of range of the user.

Keywords Arduino · Bluetooth module · Electrical devices · Smartphone · Voice control · Sonar sensor

1 Introduction

In twenty-first century, we live a life that is dependent on technology. Ordinary new innovations are made to make our lives less demanding, calm, and more agreeable. Voice recognition is the process where voice commands are converted into a machine-understandable format. The main objective of innovation has been to expand proficiency and lessen exertion. The voice command system is one of the inventive methods which simplify our daily life and improve our standard of living. Speech recognition devices allow us to give more focus on the work which has to be

M. Rahman · A. Al Farabe (✉) · Md. R. Al Islam · M. Rahman · Md. Rezyuan · G. Ashraf
Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka
1212, Bangladesh
e-mail: abdullah.farabe@gmail.com

done by our hand [1]. The voice-controlled system is an independent and less costly method that is perhaps used to control electrical machines. Nowadays, disabled people are increasing because of accidents and different kinds of chemical effects. This automation not only helps the working inhabitants but also the handicapped and elderly people.

We put a detection sensor named sonar sensor. Its usage is to verify in a room that is there a person remains in the room or not. If not then, automatically, all electric devices will be turned off. If someone does not want to detect anything, then we have also given an option to turn it off. From Google Play Store, we also need to download an app called “BT Voice Control for Arduino.” Basically, we are connecting our mobile device with the Arduino through Bluetooth so that we could give our voice as a string to the Arduino.

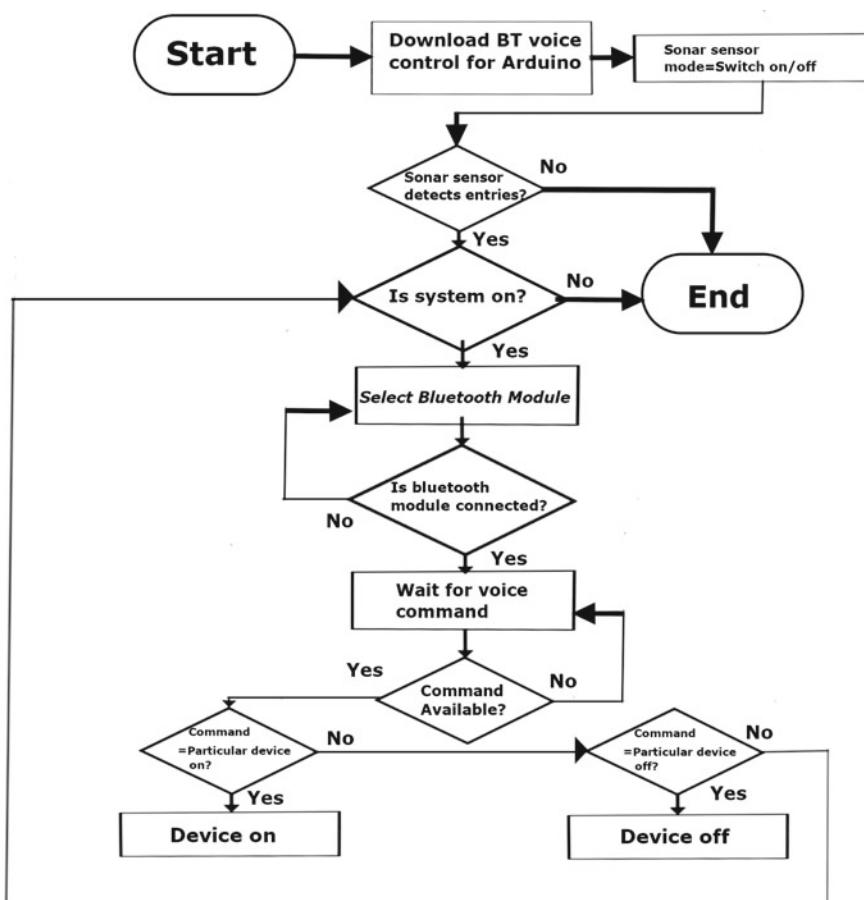


Fig. 1 Flowchart of the entire system

Our system can turn on and off any electrical device like a fan, light, or television via Bluetooth by using the android application to capture and process the voice via smartphones. After processing [2] the voice, the data is sent to Arduino using Bluetooth, and the vacant hardware translates the command to control the electrical devices. Based on a user's words, it will send to the Arduino as strings through Bluetooth. A string works as a resemblance that put restrictive expressions to turn something turn on/off. So, it can be used in both homes and hospitals. Figure 1 shows the whole application of the system.

The rest of the paper is about the background study of voice recognition technology, the proposed model, the experimental environment we set up, and the steps of how we reached our final model after several modifications followed by the concluding statement. Our motivation is to give priority to disabled patients, decrease the bills of electricity from schools/colleges, and build up home automation. For disabled patients, our device reduces their suffering by only giving voice commands to put switch on/off. Auto-presence detection could be beneficial by auto-switching off the electrical devices if anyone forgets to switch them off which reduces the electricity bill of the proprietor. The whole project could be useful at a cheap rate.

2 Literature Review

Thinking about the advancement of voice recognition protocol is progressing from the baby-talk level of remembrance of single digits to building thousands of vocabulary words. The first articulate recognition structure that could acknowledge only digits was designed by Bell Laboratories in 1952 the "AUDREY" system [3]. Voice command system made their major signs of progress in the 1970s, and one of the main reasons was Carnegie Mellon's "HARPY" voice-understanding system which could understand 1011 words. Moreover, the system turns Toward Prediction Over the next period. The vocabulary increased to several thousand words, and the reason behind this change was the new statistical model named the hidden Markov model [4]. After the invention of faster processors, voice recognition systems became more practical as it could distinguish uninterrupted speech. Therefore, the arrival of the Google Voice Search Application added a new change in history [5]. Google search application added 'personalized recognition' on android phones for voice search. Thus, we used the android voice recognition application to control the electrical devices where humans can pass voice commands [3].

In the existing literature, Agustin et al. [6] have invented a process that could be helpful for physical disability patients. This experimentation serves to develop voice recognition which is home automation, and it will be applied to hospitals. A disabled patient room is made to simulate the whole system of a room. The microphone receives the patient's voice, and for the voice recognition process, a module is used named V3 voice recognition. Arduino Uno is used to be handled to control the whole recognition system. The process will start after the recording of the patient's voice and gives an accuracy of 75%.

A paper of bt Aripin and Othman [7] represent home appliances development based on the android commander voice recognition system. It is established to help disabled and elder people at home. For voice recognition, the Google application has been used to capture voice as input from the smart device. A Bluetooth module has used in the Arduino Uno which takes the voice commander from a person to the Arduino to start the whole system, and it will give a signal to control the light and fan. The proposed system illustrates control of up to 20m of range via Bluetooth.

In a research paper, written by Suresh et al. [8] have found an issue that most of the students and faculties are adapted not to switch off the lights, fan, aircon, etc., which effect to a negative impact on organization for receiving a huge bill of overuse of electric energy. Challenging of change habits so they have come up with an idea to develop an automatic system using efficient energy in the classroom to control them in a specific region based on the presence of humans using relay control as opposed to the ceiling-mounted control and whether people are present in a room regardless of their position. Additionally, they have offered mobility and remote command execution for the relay control by utilizing the Android mobile app through Bluetooth and voice command.

In research, Ramya and Nandan [9] are concerned for two major reasons based on security and privacy for home automation. The discussion of the procedures is to ensure the security of the user data. They have used a PIR sensor and triggering circuit in their machine to detect intruders through the surveillance system.

From studying the above research in our paper, we tried to develop the same process so that we could get better accuracy with a bigger distance for voice recognition. It came to our mind to support disabled patients, classrooms, and home automation. Disabled patients suffer the most they could not walk to turn the button on/off. In the classroom, most of the students or faculties leave the classroom without switching off lights or fans. Moreover, distance limitation is another problem to control through mobile for voice recognition. We are able to range up to 25 m. In addition, we have used a Sonar sensor which also could detect a person in a room so that in an empty room any sort of device will turn off automatically. We have used a Sonar sensor rather than using PIR sensor. Detecting obstacles is not affected by as many factors because ultrasonic sensors (Sonar sensor) work using sound waves. Reliability is a must to select sensors among all odds, and ultrasonic sensors are more reliable than IR sensors at a cheap cost [10]. Our developed process could give around 80.5% accuracy of work and is very easy to use for any person. Voice command technology can hold a great part in simplifying the lives of human beings by controlling electrical inventions. This system can reduce the need for the training process of new technology which can also save time. Moreover, the use of voice commands can increase the demand and usage of technology among ordinary people, and it can be used for disabled people for leading an obstacle-free life and also very good for market value as well.

3 Proposed Model

In the voice recognition system for controlling electrical devices using Arduino and android, the components shown in Fig. 2 are used. They are—android-based application, solderless breadboard, Arduino Uno, HC-05 serial Bluetooth module, jumper cables, 4 channel 5v relay module

3.1 Android-Based Phone

Android provides a unified approach to managing application development for mobile devices; developers must create for android and ensure that their applications can function on a variety of android-powered devices [11]. For doing our project, we have used an android application, and we have to connect it with Arduino via a Bluetooth module. In this model, we have used BT voice control for Arduino, which is shown in Fig. 3a.

The application work by perceiving human voice charge. Figure 3b shows the voice command application where it is waiting for the voice command from a user. It will then show the words that have been talked and then send information/strings to the Arduino by means of Bluetooth. A string resembles a word; user can put restrictive expressions out of it [ex: if (voice == “*fan on”) { //turn Pin #2 on}]. The “voice” is the user’s string; “==” is user condition (implies equivalent to); “*computer on” is user order, and the code inside the wavy supports “{ }” are the codes to be executed once the string matches the charge condition [11].

The application sends strings in this configuration *command#; the reference bullet (*) demonstrates the beginning of another charge, and the hashtag (#) shows the finish of an order.

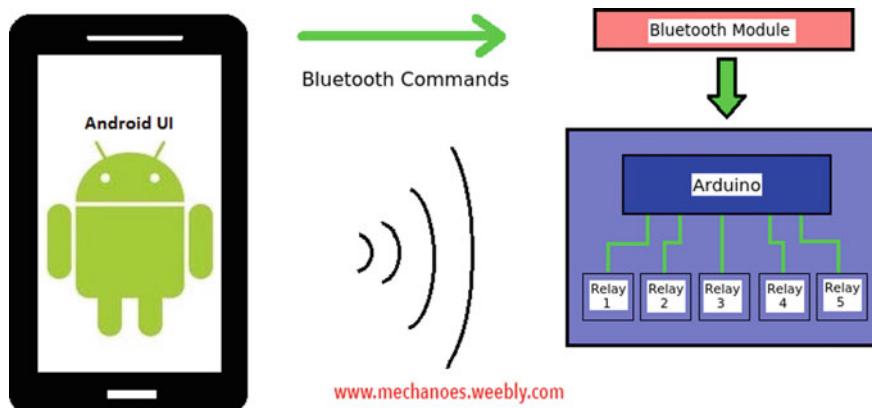
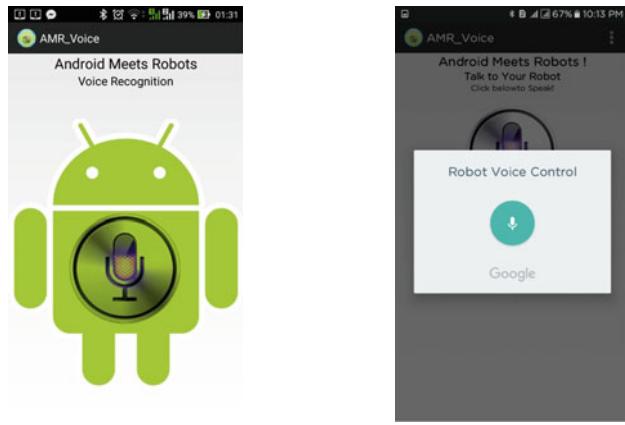


Fig. 2 Block diagram of the system



(a) BT Voice Control For Arduino

(b) Waiting for Voice Command

Fig. 3 Voice command application

3.2 *Arduino Uno (R3)*

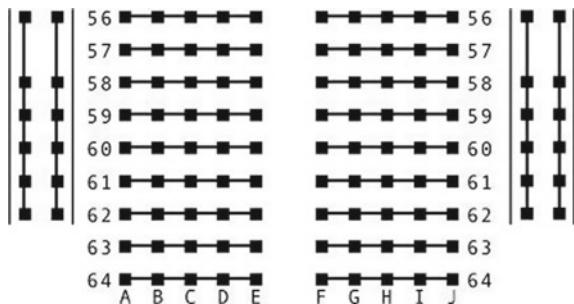
In Uno, 14 advanced input/output pins, including six that can also be used as PWM outputs, six simple inputs, a USB connection, an energy jack, and a reset button is required. It stores everything required to support those microcontrollers, effectively connects it to a computer for a USB connection or powers it using an AC-to-DC connector or battery [12]. First of all, there is a power USB which is pointing by number 1, and the barrel jack is by 2. There are several pins such as 5v, 3.3v, GND, analog, digital, PWM, and AREF which are numbered followed by 4, 5, 3, 6, 7, 8, 9. Moreover, there is a reset button and power LED indicator which is labeled by numbers 10 and 11. Figure number 12 represents TX RX LEDS. Now, the main IC part is major in Arduino, and for Arduino Uno (R3), this one is labeled by 13. Finally, the voltage regulator is indicated by the number 14 [13].

3.3 *Breadboard*

In a breadboard, gaps are conductively associated unquestionably. The lines named with numbers (from 1 to 64) are conductive, with a non-conductive score amidst the board. A breadboard is a board that we might utilize in a model with distinctive circuits. Numerous breadboards can be associated together to structure bigger prototyping zones [14].

The breadboard's structure is outlined in Fig. 4. The breadboard has a chip using columns E and F; this chip has been put on rows 5, 6, 7, and 8.

Fig. 4 Structure of breadboard



On the off chance that an association is made to any point in position A5, B5, C5, D5, this is the same as interfacing straightforwardly to the chip leg that is sitting in E5, in light of the fact that the line A5–E5 is electrically associated [14]. Because the line from position A5 to position E5 is electrically connected and connecting to any location in positions A5, B5, C5, or D5 is equivalent to connecting directly to the chip leg in position E5 [14]. The two legs of the chip on line 5 are not connected along these lines because there is a physical hole between lines E and F. It merely takes associating with any location in position G5, H5, I5, or J5 to interface with the chip's top right leg. Vertical transporters are located at the breadboard's outside left and right sides where vertical electrical connections exist between these edges. Transporters are typically used for a power source's positive and ground terminals, such as a battery, because we may need to make numerous associations to other devices [14, 15].

3.4 *Bluetooth Module HC-05 and 4 Channel 5V Relay Module 10A*

In place to control our model, we associated a Bluetooth HC-05 chip with our Arduino gadget. This Bluetooth module can undoubtedly attain serial remote information transmission. This module will be configured for a serial interface, which is straightforward to use and streamlines the whole design process [16]. The HC-05 can operate with a supply voltage ranging from 3.6 to 6 VDC although the RXD pin's logic level is 3.3 V, and it is not 5 V tolerant. Connecting it straight to a 5 V device, it could be harmed (e.g., Arduino Uno and Mega). It is advised to use a logic level converter to safeguard the HC-05. If the "EN" pin is pushed to logic 0, the HC-05 will lose power. The weight of this JC-05 is 4 g with a size of 36.5 * 16 mm. It has safety features of authentication and encryption. Transmit power can operate under – 20 to 55 °C temperature. The operating frequency of Bluetooth module hc-05 is 2.4 GHz ISM frequency band. The pins of the Bluetooth module are the positive pole of the power source, GND, TXD, the transmission terminal, and RXD, the receiving terminal [16]. Exactly a little bit from claiming stranded center wire with a decent robust pin con-

nection looking into possible wind. They have that adaptability by claiming stranded wire will fit straightforwardly under breadboards and female pin headers [17]. These jumpers are perfect for connecting several breadboard components together. The wires are strong, resilient, and reusable. They have rigid ends that are simple to press into the holes in the breadboard. Jumper wires that are flexible at both ends have hard pins attached to them [15].

A relay module transfers table meets expectations from a 5 V supply to the transfers. Furthermore between 1 V also 5 V to the exchanging indicator. It employs a transistor and isolator with a switch the transfer with respect to thereabouts could make associated straightforwardly will a microcontroller pin. Switches up to 10 Amps. Rated at dependent upon 250 V. Power and ground are provided to the module via the pins designated VCC and GND, respectively. Low inputs to the IN1, IN2, IN3, and IN4 inputs energize the relays. There are four relays, each of which offers outputs for dry contacts. The common (COM), usually open (NO), and normally closed (NC) terminals are all provided by each relay, respectively [18]. This is basically four different circuits on one board. The channels are separate from one another, other than the fact that they share ground and VCC.

For 4 channel 5V relay module 10A, its current consumption is 75 mA at 5 V for each when it is switched on. There is a transistor, and this is known as an on-board switching transistor with on-board back EMF protection which is very necessary for 4 channel 5V relay module 10A. It is easier because it has an easy 1-wire drive. Next, we are eager to use is Sonar sensor for two main concepts—one is thermal detection, and another one is Sonar Sound Navigation and Ranging detection.

3.5 Sonar Sensor

For human entry passes (HEP), the Sonar sensor uses a vision sensor for detection and tracking. The tracking algorithm has been set only to detect humans or any other species which necessitates an effective strategy for feature extraction and tracking. Conditions have put the range to 25 m from at any distance; it could detect human if a human goes outside the room or not [19]. The Sonar sensor's VCC and trigger are connected with Arduino's PD6. The other Echo Plus is connected with PD7.

4 Hardware Implementation with Result Analysis

We have implemented the proposed model by using the above apparatuses in Fig. 5. At first, we connected the Bluetooth module with the Arduino. We utilize 5 V to power the Bluetooth module because we are using an HC-05 Bluetooth module with a board. Next, we attached the Bluetooth module's RX (pin 0) pin to the Arduino's TX (pin 1) pin and TX (pin 0) pin to the Bluetooth module's TX pin. Furthermore, we used a wire from the Arduino 5v pin to BT's Vcc pin and another pin is GND,

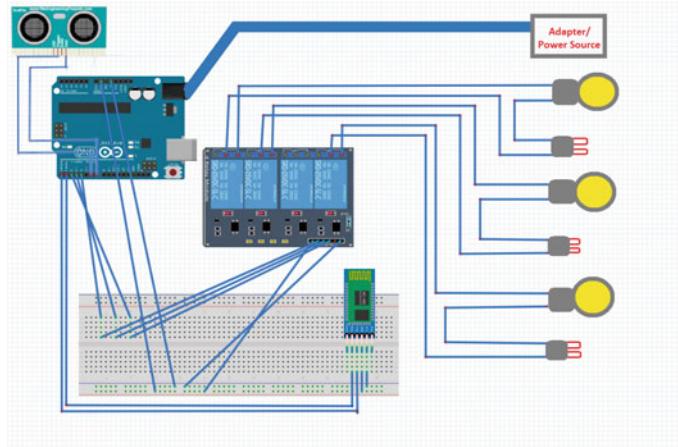
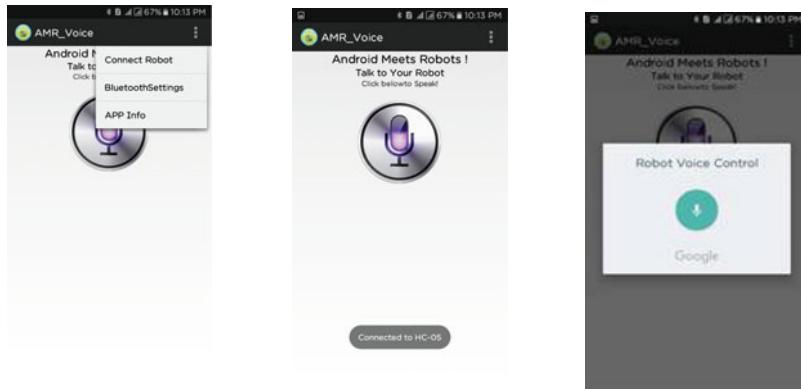


Fig. 5 Circuit diagram



(a) Connecting to Bluetooth Module (b) Connection Successful (c) Waiting for Voice Command

Fig. 6 Connecting to bluetooth module HC-05

which we plugged in a place of the breadboard. Then, we connected the electrical devices to the relay module with wires. The relay module is used to turn the devices on/off, and it works when the Arduino sends the high/low signal.

Further, we needed to download the app ‘BT Voice Control for Arduino’ from Google Play Store. After that, we selected the ‘Connect Robot’ option, shown in Fig. 6a, and choose Bluetooth Module HC-05 which is in Fig. 6b and used in this model. Then, we had to wait for the confirmation of the connection.

Therefore, the app starts working by pressing the mic button where the user can deliver his command, and the process is displayed in Fig. 6c. The application first recognizes the voice command and then displays the word and sends the string to the Arduino via Bluetooth.

Table 1 Result analysis among literatures

Literatures	Accuracy (%)	Range (m)	Distance module/sensors
Voice recognition system for controlling electrical appliances in smart hospital room [6]	75	None	None
Voice control of home appliances using Android [7]	Not needed	20	Via bluetooth
Automatic lighting and control system for classroom [8]	x	x	Relay control
Analysis of security issues and possible solutions in the Internet of Things for home automation system [9]	x	x	PIR sensor and triggering circuit
Voice command automation system (VCAS) for controlling electrical devices using Arduino	80.51%	25	Ultrasonic sensor (sonar sensor)

After receiving the voice command, the instruction is sent to the Arduino via Bluetooth, and then, Arduino decodes the instruction and sent the signal to the relay module. Devices are switched off when a low voltage is applied to the relay module and on when a high voltage is applied. As technology develops day by day, smart home systems are advancing quickly. With the development of smart home automation, numerous research projects have been created to simplify daily living, provide security, and save time and energy. Using less-expensive sensors, a device with a low-cost, adaptable home control monitoring system (HCMS) might be constructed [20].

From Table 1, we are comparing the features with our project. We couldn't get full access of literature. Among the works of literature, we have analyzed a few of them that give either 75% accuracy or worked of distance within a 20 m range or used PIR sensors to detect human presence. But, what we have worked on better accuracy of 80.51%. We got a 25 m range via Bluetooth connection and used a Sonar sensor (Ultrasonic sensor) which is more reliable than IR sensors.

5 Conclusion and Future Works

The project we are working on attempts to address a workable solution to the requirement for very basic automation in our homes and hospitals. With our entire project providing 80.5% accuracy with the detection presence model of a human, the user will be able to have complete control over any appliance without exerting any effort by simply saying the appliance name and the corresponding number assigned to that specific appliance and telling it to switch on or off.

However, we have limitations as well. This project is fully handmade, and the whole application works based on the accuracy of all the elements we have used.

Each part might carry a few errors. We have used not-so-expensive materials. We need more accuracy by using expensive elements. In addition, the range could be another issue as nowadays people try to make use of control system from far distance which we have given possible of 25 m. We are really eager to make it more distance as far as possible. In addition, we have focused here only on disabled people, but we want to also upgrade it for blind people too. Delightfully want to make a device that will serve as a link between blind users and end users as a result of advancements in telecommunications technology. In the future, we plan to interface the Braille pad with desktop computers or mobile devices to connect the system and allow a dual-impaired person to use it independently. Text on a Braille pad could be converted to voice using a translator [21].

Additionally, we want to switch from Arduino to field programmable gate array (FPGA) . FPGA is a method that was created with industrial control system applications in mind. When portable hardware description languages and system-level programming are used more frequently, it is beneficial for development tools and applicable CAD environments. Setting up modeling to assess the surroundings for whole industrial electronics systems is unique [22]. FPGA solves complex operations and real-time applications as well. For the reconfigure ability and boosting of the process, the FPGA could make that possible. Wireless communication is enormous in the industrial security system as it has a lot more complex configuration that the FPGA could handle for its practical applications.

A user can easily understand the functions of android applications in a very certain limited time because of their user-friendly quality. It is saving manual labor, human efforts, and time after electronic appliances. Though this system is primarily aimed to reduce human effort, also, it will be a handful for the old aged and physically handicapped people.

References

1. Ramil T, Dabimel NN, Parimon N, Porle RR (2016) Simple speech controlled home automation system using android devices. *J Sci Res Dev* 3(1):33–38
2. Kumar M, Shimi SL (2015) Voice recognition based home automation system for paralyzed people. *Int J Adv Res Electron Commun Eng (IJARECE)* 4(10)
3. Pinola M (2013) History of voice recognition: from Audrey to Siri | IT Business. ITBusiness.ca | Business Advantage Through Technology. Available at: <https://www.itbusiness.ca/news/history-of-voice-recognition-from-audrey-to-siri/15008>
4. Pinola M (2011) Speech recognition through the decades: how we ended up with Siri (PCWorld). Retrieved from: https://www.pcworld.com/article/477914/speech_recognition_through_the_decades_how_we-ended_up_with_siri.html
5. Kikel C (2022) A brief history of voice recognition technology. Total voice technologies. Available at: <https://www.totalvoicetech.com/a-brief-history-of-voice-recognition-technology/>
6. Agustin EI, Yunardi RT, Firdaus AA (2019) Voice recognition system for controlling electrical appliances in smart hospital room. *TELKOMNIKA (Telecommun Comput Electron Control)* 17(2):965–972

7. bt Aripin N, Othman MB (2014) Voice control of home appliances using Android. In: 2014 electrical power, electronics, communications, control and informatics seminar (EECCIS), pp 142–146. Available at: <https://ieeexplore.ieee.org/abstract/document/7003735>
8. Suresh S, Anusha HNS, Rajath T, Soundarya P, Vudatha SVP (2016) Automatic lighting and control system for classroom. In: International conference on ICT in business industry & government (ICTBIG), pp 1–6. Available at: <https://ieeexplore.ieee.org/abstract/document/7892666>
9. Ramya PS, Nandan D (2020) Analysis of security issues and possible solutions in the internet of things for home automation system. In: Soft computing: theories and applications. Springer, Singapore, pp 825–836. Available at: https://link.springer.com/chapter/10.1007/978-981-15-4032-5_74
10. Burnett R (2020) Ultrasonic vs infrared (IR) sensors—which is better? MaxBotix Inc. Available at: <https://www.maxbotix.com/articles/ultrasonic-or-infrared-sensors.htm>
11. Yan M, Shi H (2013) Smart living using Bluetooth-based Android smartphone. *Int J Wirel Mob Netw* 5(1):65. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.5096&rep=rep1&type=pdf>
12. What is an Arduino? (2021). Available at: <https://learn.sparkfun.com/tutorials/what-is-an-arduino>
13. Arduino void setup and void loop functions [explained]—the robotics back. End (2021). Available at: <https://roboticsbackend.com/arduino-setup-loop-functions-explained/>
14. Tomeczak S (2013) Breadboard basics 1: what is a breadboard? Structure and overview (little scale). Available at: <http://little-scale.blogspot.com/2013/03/breadboard-basics-1-what-is-breadboard.html>
15. Science Buddies. How to use a breadboard. Available at: <https://www.sciencebuddies.org/science-fair-projects/references/how-to-use-a-breadboard>
16. BlueTooth-HC05-Modules-How-To (2016) Available at: <https://arduino-info.wikispaces.com/BlueTooth-HC05-HC06-Modules-How-To>
17. Breadboard jumper wire set (140 PCs pack). Available at: [https://www.seedstudio.com/Breadboard-Jumper-Wire-Set-\(140-PCs-Pack\)-p-1562.html](https://www.seedstudio.com/Breadboard-Jumper-Wire-Set-(140-PCs-Pack)-p-1562.html)
18. Components101. 5V four-channel relay module. Available at: <https://components101.com/switches/5v-four-channel-relay-module-pinout-features-applications-working-datasheet>
19. Kim S, Oh SY, Kang J, Ryu Y, Kim K, Park SC, Park K (2005) Front and rear vehicle detection and tracking in the day and night times using vision and sonar sensor fusion. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 2173–2178. Available at: <https://ieeexplore.ieee.org/abstract/document/1545321>
20. Singh AK, Agrawal S, Agarwal S, Goyal D (2020) Low-cost and energy-efficient smart home security and automation. Computational network application tools for performance management. Springer, Singapore, pp 95–108. Available at: https://link.springer.com/chapter/10.1007/978-981-32-9585-8_10
21. Korde A, Gaikar O, Nikam S, Rukhande S (2020) Wireless emanation of braille to text/voice and vice versa. Soft computing: theories and applications. Springer, Singapore, pp 403–411. Available at: https://link.springer.com/chapter/10.1007/978-981-15-4032-5_37
22. Monmasson E, Cirstea MN (2007) FPGA design methodology for industrial control systems—a review. *IEEE Trans Ind Electron* 54(4):1824–1842. Available at: <https://ieeexplore.ieee.org/abstract/document/4267891>

Turing Machines Behind the Horizon: Modeling Black Hole Interiors as Transfinite Limited Turing Machines



Ajay Agarwal 

Abstract The information encoded in the Hawking radiation emitted from an evaporating black hole has puzzled astrophysicists and computer scientists alike for decades. Recent developments in quantum complexity theory and post-quantum cryptography shed light on a family of radical approaches to treat black hole interiors as systems beyond thermal states with high energy and density. In this paper, this study attempts to model the interior of a black hole using weaker variants of Infinite Time Turing Machines. This study curates two separate models—first based on the entanglement entropy of the evaporating black hole following the Page curve and the other based on the EBH quantum states of the same. This approach also highlights how the choice of the limit state significantly impacts the treatment of black hole interior encoding. Finally, the study shows how our model follows the Page curve and the entanglement wedge casually as a consequence of the correct choice of transition function. The contribution of our paper is to highlight the utility of classical computational models to generalize over high-energy thermal states like black holes.

Keywords Infinite time turing machines · Black hole information paradox · Page curve · Post-quantum cryptography · Classical computational complexity

1 Black Hole Information Paradox

Information loss from black holes has always dictated the attention of computer scientists and astrophysicists alike [9]. Approaches to understanding the nature of the encoding that escapes the black hole as Hawking radiation while robust and exhaustive swing rapidly from describing one paradox or another—in terms of either describing an AMPS firewall interior that prohibits Susskind's four postulates on a black hole or to the end of describing errors in Hawking's radiation calculation leading to a possibility of a Python's lunch that attempts to model black hole interiors using quantum extremal surfaces in Anti-de Sitter space (AdS)

A. Agarwal (✉)
DIT University, Dehradun 248009, India
e-mail: verslinfiniaudela@gmail.com

[6]. Notions of quantum mechanics, computational complexity, and semi-classical physics demanding revision are a common ask from all these approaches [13]. However, casually speaking, there is one inference that is beyond disputable—developing a model of the black hole and its consequent information flow dynamics will at all costs require an interdisciplinary approach from both classical computational complexity and quantum mechanics [7].

This paper builds upon the prior radical works in the field of describing black holes as quantum pseudorandom encoders [10]. Recent works of Kim and Tang [25] have demonstrated, not only, the possibility to interpret Hawking radiation as a pseudorandom stream of information, but also allowed the introduction of post-quantum cryptography as a possible method to approach the same problem [5]. This study introduces the possibility of modeling a transfinite (quantum) limited Turing Machine as the black hole interior which attempts to approach the Black Hole Information Paradox problem by considering two potential input models. The contribution of our paper is:

1. Understanding the utility of defining Hawking radiation as a pseudorandom flow of information [17].
2. Modeling black hole interior as weak Infinite Time Turing Machine (wITTM) with halting condition adjusted with the timeline of an evaporating black and input symbol space as entangled quantum states supporting black hole complementarity [15].
3. Modeling black hole interior as wITTM with halting condition adjusted with Page time of an evaporating black hole and input symbol space as maximum entanglement entropy values [14].

It must be noted that each of the chosen approaches is limited by their interpretation of the Black Hole Information Paradox and the choice of geometric space in consideration [4]. However, the discussion in this paper extends beyond scope of the paper [3, 8].

2 Definition of wITTM

ITTMs, also known as Infinite Time Turing Machines, can be interpreted as standard Turing Machines with major three modifications:

1. ITTMs run for infinitely many steps.
2. The head of ITTMs returns to the leftmost end of the tape, once ITTM reaches the *limit stage*.
3. As it returns to the leftmost cell in the tape, ITTM replaces the content value of each cell with *lim sup* of the previous values it read.

While the above modifications can be utilized to briefly describe infinitary computation, it is likely to appear that ITTMs hold equivalent, if not, lesser computational power than Turing Machines. However, such is not the case. Avoiding unnecessary

extension of our work as a primer on the same, a succinct statement on the computational power of the ITTMs is that—*Halting problem is infinite time decidable for ITTMs as either the ITTMs halts in finitely many steps as the solution to the problem in question, or it reaches the halting state after ω steps*. Consequently, any ITTM can decide which sets are either decidable or undecidable on ordinary Turing Machines given their infinite time computation. Since the introduction of infinitary computability through ITTMs by Hamkins, the notions of infinitary computability have been extended by curating variants of the ITTMs that possess varied advantages against the same for varied sets of reals and their corresponding arithmetic function. For this work, this study reiterates the works of Bianchetti on the weaker variants of ITTMs [23, 25].

As stated earlier, ITTMs extend the model of computability of the Turing Machine to infinite ordinal time by modifying the content of each cell at the limit step from its original value to the *lim sup* of the content [18]. Hence, a question of a computationally strictly weaker version of ITTMs would require the machine to effectively halt before the machine reaches the limit stage, or to say ω steps. Bianchetti [24], in 2019, introduced the concept of wITTM, weaker variants of the Infinite Time Turing Machine, unlike an ITTM that replaces the content value with *lim sup* value, with an “eventually constant” rule [24]. This rule, effectively, means that the value of each cell’s content shall be defined only in the case the cell’s content value has been stabilized before the machine reaches the limit step. In such a case, the value of the cell is, hence, equal to the stabilized content value. Consequently, wITTM are computationally more powerful than ordinary Turing Machine and weaker than ITTMs. To clearly understand the “eventually constant” rule, this borrows its mathematical definition from Bianchetti’s original work on the same [19, 20].

For a wITTM with a limit step at α , then content in some cell c that fails to stabilize before α denoted by $c[\alpha]$ is:

$$c[\alpha] = \begin{cases} 0 & \text{if } \exists \beta < \alpha \forall \gamma \in (\beta, \alpha) c[\gamma] = 0 \\ 1 & \text{if } \exists \beta < \alpha \forall \gamma \in (\beta, \alpha) c[\gamma] = 1 \\ \uparrow & \text{Otherwise} \end{cases} \quad (1)$$

Here, β and α are steps on the wITTM. Notice, the computationally weaker definition of the wITTM against the ITTM arrives from the ability of the weaker variant failing to converge at any output when the contents in some cell c are unable to stabilize before the limit step α . That is to say, the wITTM essentially hangs at α , or in more objective words, the machine never reaches the limit step α . This forms the basis of our work, as to how in the black hole interior (by the first approach), the entanglement entropy at any given time during its evaporation never reaches its *max* limit as it progresses through Page time, hence following Page curve [12]. Also, as per the second approach of treating a black hole as a robust quantum cloner for pseudorandom Hawking radiation, this wITTM takes an input of early Hawking radiation denoted by $|E|$ and the size of the remaining black hole denoted by $|H|$ to

be the input and the scratch tape, respectively, and the transition function effectively performing quantum entanglement with a low Krauss rank [2]

3 Black Hole as Weak Infinite Time Turing Machine Using Entanglement Entropy and Page Time

This begins by defining a wITTM denoted by M that has the following properties:

1. M is a three-tape model of wITTM, two input tapes X and R , and one output tape S_R .
2. The input tape takes into consideration the microscopic states (both classical and quantum) of the evaporating black hole.
3. The transition function is defined by the Faulkner–Lewkowycz–Maldacena formula that utilizes conformal field theory to calculate the entanglement entropy of a classical extremal surface anchored to the holomorphic derivative of that region, or that of a quantum extremal surface.
4. The *limit step* α is defined as the time it takes for the evaporating black hole to achieve the entanglement wedge, that is to effectively state the time when the straddle shock occurs and the entanglement entropy begins to decline/fall.
5. The limit stage of the wITTM, hence, becomes equal to the Page time which lies under the results obtained from the Hayden–Preskill protocol.
6. The *lim sup* is replaced by the *dilaton-induced extremum* operation, which shall comprise the following step:
 - a. Extremize the generalized entropy using the dilaton $(\phi - \phi_0 = \frac{2\phi_r^-}{x^+})$ in the ∂^+ direction.

This defines a wITTM for holographic entanglement entropy for an evaporating black hole in Anti-de-Sitter space. Derivations of the quantum extremal surfaces in the AdS-CFT field theory lie outside the scope of this work, and hence, the reader is suggested to refer to the following for the same [1].

In summary, an evaporating black hole's interior is modeled as a weaker variant of the Infinite Time Turing Machine in the above conceptualization. Mathematically, we extend our above reasoning to provide the following definitions of transition function δ using the Faulkner–Lewkowycz–Maldacena formula:

$$\delta(\Gamma) = \delta(X, R) = S_{\text{gen}}(X_R) = S_{\text{gen}}(\mathcal{X}_R) + \mathcal{O}(\hbar) \quad (2)$$

where Γ is spacelike slice in the spacetime \mathcal{M} that acts as an AdS-Cauchy surface with the property:

$$\partial X = \partial R \quad (3)$$

essentially implying that X is homologous to R , where R is a religion in the AdS/CFT slice, and X is codimension 2, a spacelike extremal surface. This property is also necessary as it stands as the precursor to Hubeny, Rangamani, and Takayanagi (HRT) and Faulkner, Lewkowycz, and Maldacena (FLM) calculation of entanglement entropy for the quantum extremal surface in question, hence, the formulation of our transition function [11]. In the case of the formula of our transition function, X_R represents the classical extremal surface anchored to the quantum extremal surface \mathcal{X}_R .

Here, $\mathcal{O}(\hbar)$ essentially represents the order in the difference of magnitude of the entanglement entropy of the classical extremal surface and the quantum extremal surface. Consequently, \hbar represents Planck's length. Correspondingly, the microscopic configurations of this quantum extremal state act as the alphabet for wITTM replacing the Cantor space used as the alphabet for classical wITTMs [16–18].

As stated earlier, similar to how a wITTM reaches a limit position after ω steps, in our case, an evaporating black hole's interior encode wITTM denoted as M reaches a halting position after it reaches Page time, which is the point of the black hole's evaporation thermodynamic irreversibility, often referred to as the "halfway point." However, it must be noted that Page time has nothing to do with the evaporation time and rate of change of horizon area concerning entropy. Page time, hence, is the halting time of M when the entanglement entropy is maximum and wITTM halts to move back to the leftmost cell of the tape replacing the content of the previous cells with the *dilaton-induced extremum* operation. Correspondingly, the following lemma could be stated:

Lemma 1 *For wITTM denoted by M encoding evaporating black hole interior using dilaton-induced extremum operation, the content of the cell α after ω steps, where ω is the halting step of M , as at cell α , the content of the cell is the maximum entanglement entropy of the evaporating black hole, which occurs at Page time which is the halting time for wITTM M .*

$$c[\alpha] = \max(c[\beta]) \text{ if } \beta < \alpha \quad (4)$$

This completes our formulation of the wITTM for black hole interior using Page curve and entanglement entropy (or average von Neumann entropy). For succinct visualization of our conceptualization, refer to Fig. 1.

4 Black Hole as Weak Infinite Time Turing Machine Using EBH States

In this approach, our main focus is on the treatment of black holes as quantum cloners of Hawking radiation. The states in consideration for our approach are: 1. early Hawking radiation (denoted by E), 2. remaining black hole (denoted by H), and 3. recently emitted Hawking radiation (B) [22]. Our approach shall develop

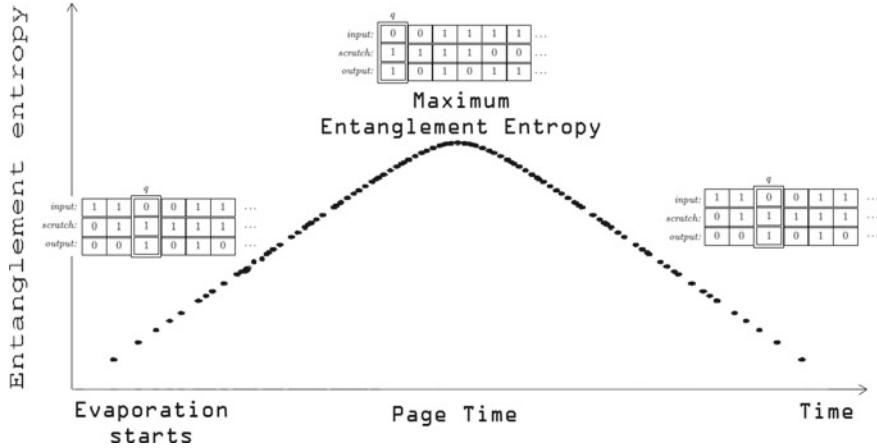


Fig. 1 Notice the figure. The original Page curve from his seminal paper has been superimposed with the state of weak Infinite Time Turing Machine at each step. The peak of entanglement entropy occurs at Page time

using existing works of entanglement wedge reconstruction using toy models of evaporating black holes in AdS-CFT space [16, 21].

Consider a wITTM denoted by M defined by the following characteristics:

1. M is a four-tape model of wITTM with three input tapes and one output tape.
2. Each input tape represents one of the three states— E (early Hawking radiation), H (remaining black hole), and B (recent Hawking radiation).
3. The transition function depends on the treatment of the three states E , H , and B . Here, the unitary operator U_{bh} is the transition function, which is the quantum gate for black hole.
4. Here, the *limit stage* ξ is defined as when any two-outcome measurement S with a polynomial of quantum complexity with size equal to $|H|$ fulfills the given condition, where ρ_{EB} is the marginal mixed state of EB, and σ_{EB} is the maximally mixed state of EB obtained from the output tape content $c[\xi]$.

$$|\Pr(\mathcal{M}(\rho_{EB}) = 1) - \Pr(\mathcal{M}(\sigma_{EB}) = 1)| = 2^{-\alpha|H|} \quad (5)$$

where , $c[\xi] = \psi_{EBH}$

5. The *lim sup* is replaced by S which is a *two-outcome measure* that involves the difference between the maximally mixed quantum state of Hawking radiation emitted till then and the current quantum state of the Hawking radiation emitted till then.

This completes the formulation of the wITTM for black hole interior using EBH quantum states. Now, we proceed to mathematically infer certain definitions of our wITTM.

Notice the role of output tape in this wITTM model. It acts as the mathematical space of all possible quantum entangled states for our three basis states E, B, and H. Based on the works of Kim et al. [25], we can now define this output tape set as:

$$|\psi\rangle_{EBH} = \frac{1}{2^{|H|}} \sum_i |\psi\rangle_i_{EB} \otimes |i\rangle_H \quad (6)$$

The limit step ξ essentially acts as the local maximum of the difference between the marginally mixed state and maximally mixed state, i.e.,

$$|\Pr(\mathcal{M}(\rho_{EB}) = 1) - \Pr(\mathcal{M}(\sigma_{EB}) = 1)| \leq 2^{-\alpha|H|} \quad (7)$$

The above condition follows directly as a consequence of works by Kim et al. [25]. Finally, we draw our attention to a lemma that follows directly as a result of our formulation of the wITTM.

Lemma 2 *For wITTM denoted by M encoding evaporating black hole interior using EBH states, and given limit stage ξ after ω steps, where ω is the halting step of M , as the maximum difference between two-outcome measurement of the marginal and maximally mixed states,*

EB is quantum pseudorandom

where EB is the total Hawking radiation mode.

The corollary of the lemma was proven by Kim et al. [25]. Consequently, the same can be formulated for wITTM M .

As we proceed toward the completion of wITTM for EBH states, we succinctly define what the entanglement wedge shall appear in this case. Given the output tape stores a quantum mixed state for E , B , and H and the focus of the halting condition is to be defined by the range of possible errors in the quantum measurement of the marginally and maximally mixed state, the wedge diagram shall be geometrically similar to the Page curve as described in Fig. 1, with the entanglement entropy replaced with $|\Pr(\mathcal{M}(\rho_{EB}) = 1) - \Pr(\mathcal{M}(\sigma_{EB}) = 1)|$. Also, note, **unlike the Page curve, here, the peak of the value shall not occur after Page time.**

5 Conclusion

The possible future scope of our work includes exploring various transition functions for the 4-tape wITTM developed in the second approach. Some approaches that might be worth mentioning are: mapping from black hole microstates to similarity between encoded and decoded information obtained from Hawking radiation (hence, obtaining a rather inverted wedge), mapping from black hole microstates to Kolmogorov's complexity of the obtained string (encoded information), or more

generally, Kolmogorov's complexity of the encoded information for every possible pair of black hole microstate configurations. Further questions to explore might include whether the extremum of the difference in the measurement of the marginally and maximally mixed state of EBH is obtained at Page time, and if not, then at what scrambling time delay one must wait for the same.

6 Future Work

Recent developments in the field of semi-classical physics and quantum mechanics have given rise to a varied number of approaches to solving the Black Hole Information Paradox. Ranging from the treatment of such as either an infinite thermal barrier at a Plank's length behind the event horizon known as the "firewall" to a quantum error correction code that prevents the external observer ever from knowing the true state of a black hole, the development of approaches is varied and has causally resulted in curating a robust understanding of AdS-CFT correspondence and quantum complexity theory. In our current work, we extend the two sides of the dialog and accommodate both in the form of a weaker variant of the Infinite Time Turing Machine, effectively reviving the utility of classical complexity theory in face of more robust approaches. This highlights not only the mathematical utility of ITTMs for high-density particle physics but also affirms the utilization of Turing Machines at the center of our universe. While the question of whether we would ever be able to decode "almost correctly" the Hawking radiation or ever be able to measure the correct state of a black hole remains unsolved, the question Hawking asked surely has a new amendment—"What if there was just a Turing machine at the beginning of Time itself?"

References

1. Raju S (2022) Lessons from the information paradox. *Phys Rep* 943:1–80
2. Page DN (1993) Information in black hole radiation. *Phys Rev Lett* 71(23):3743
3. Page DN (2005) Hawking radiation and black hole thermodynamics. *New J Phys* 7(1):203
4. Almheiri A, Engelhardt N, Marolf D, Maxfield H (2019) The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole. *J High Energy Phys* 2019(12):1–47
5. Engelhardt N, Wall AC (2015) Quantum extremal surfaces: holographic entanglement entropy beyond the classical regime. *J High Energy Phys* 2015(1):1–27
6. Hubeny VE, Rangamani M (2012) Causal holographic information. *J High Energy Phys* 2012(6):1–35
7. Akers C, Engelhardt N, Penington G, Usatyuk M (2020) Quantum maximin surfaces. *J High Energy Phys* 2020(8):1–43
8. Casini H, Huerta M, Myers RC (2011) Towards a derivation of holographic entanglement entropy. *J High Energy Phys* 2011(5):1–41
9. Lashkari N, McDermott MB, Van Raamsdonk M (2014) Gravitational dynamics from entanglement "thermodynamics." *J High Energy Phys* 2014(4):1–16

10. Faulkner T, Lewkowycz A, Maldacena J (2013) Quantum corrections to holographic entanglement entropy. *J High Energy Phys* 2013(11):1–18
11. Headrick M, Hubeny VE, Lawrence A, Rangamani M (2014) Causality & holographic entanglement entropy. *J High Energy Phys* 2014(12):1–36
12. Espindola R, Güijosa A, Pedraza JF (2018) Entanglement wedge reconstruction and entanglement of purification. *Eur Phys J C* 78(8):1–20
13. Hamkins JD, Lewis A (2000) Infinite time Turing machines. *J Symbolic Logic* 65(2):567–604
14. Hamkins JD, Seabold DE (2001) Infinite time Turing machines with only one tape. *Math Logic Q Math Logic Q* 47(2):271–287
15. Deolalikar V, Hamkins JD, Schindler R (2005) $P \neq NP \cap co\text{-}NP$ for infinite time Turing machines. *J Log Comput* 15(5):577–592
16. Hamkins JD (2007) A survey of infinite time Turing machines. In: International conference on machines, computations, and universality, pp 62–71. Springer, Berlin, Heidelberg
17. Žák S (1983) A Turing machine time hierarchy. *Theoret Comput Sci* 26(3):327–333
18. Carl M (2020) Space and time complexity for infinite time Turing machines. *J Log Comput* 30(6):1239–1255
19. Copeland BJ (1998) Super Turing-machines. *Complexity* 4(1):30–32
20. Parikh MK, Wilczek F (2000) Hawking radiation as tunneling. *Phys Rev Lett* 85(24):5042
21. Medved AJM, Vagenas EC (2005) On Hawking radiation as tunneling with back-reaction. *Mod Phys Lett A* 20(32):2449–2453
22. Hajicek P (1987) Origin of Hawking radiation. *Phy Rev D* 36(4):1065
23. Visser M (2003) Essential and inessential features of Hawking radiation. *Int J Mod Phys D* 12(04):649–661
24. Bianchetti M (2020) Weaker variants of infinite time Turing machines. *Arch Math Logic* 59(3):335–365
25. Kim I, Tang E, Preskill J (2020) The ghost in the radiation: robust encodings of the black hole interior. *J High Energy Phys* 2020(6):1–65

A Review on Deep Learning-Enabled Healthcare Prediction Technique: An Emerging Digital Governance Approach



D. Anand, Venkateswarlu Tata, Jitendra Kumar Samriya, and Mohit Kumar

Abstract Heart disease is one among the critical human diseases as well as health-care issue, that affects human very severely. It occurs when heart is unable to supply appropriate amount of blood to the parts of body. It is the most fatal issue which cannot be seen with naked eye. At the exact time, it requires accurate diagnosis. For preventing heart failure, it is significant to give treatment to the heart disease accurately on time. In many aspects, diagnosing the heart disease is a traditional one which is not reliable. Machine learning and deep learning methods are significant as it is used to find persons ailing from heart disease. Nowadays, research on heart disease prediction is increasing which completely summarizes the research on it. This paper discusses the recent research work related to the prediction of heart disease with comparative table in terms of state of the art of various clinical support systems carried out by different researchers using deep learning and data mining methods along with its issues.

Keywords Heart disease · Health care · Diagnosis · Prediction · Machine learning · Deep learning techniques

1 Introduction

Nowadays, human life is varied by information technology. By using technology, health sector is emphasized rapidly throughout the world. This gives more data about patients worked by technologies in machine learning which is useful for knowledge and data. These data help to construct an expert system which discovers the disease

D. Anand
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andra Pradesh, India
e-mail: ananddama92@gmail.com

V. Tata
Guntur Engineering College, JNTUK, Kakinada, India

J. K. Samriya · M. Kumar (✉)
Dr B. R. Ambedkar NIT Jalandhar, Jalandhar, India
e-mail: kumarmohit@nitj.ac.in4

by providing low cost with less processing time and enhanced diagnose. Every day, modern medicine gives huge data to solve challenges successfully [1].

By using the medical history of patients, healthcare systems enhance human life by increasing medications' use and healthcare services. Some critical challenges such as data threat, minimal medical data, avoidable errors, diagnosis inconsistencies and medical data transmission delay are faced by health support systems. Even for typical disease, available data are so massive and challenging one to make steady and accurate decision, so that clinical decision support system (CDSS) and electronic health recorder (EHR) are introduced to support disease diagnosis process. From the investigation, the aim is to design Disease Prediction Scheme (DPS) and several research communities consider this area for research. Researchers have greater interest on computerized medical data systems. It is easy to predict a person with illness using the medical parameters of patients and non-patients [2]. In modern world, heart disease has highest growth rate. Based on the survey, about more than 17.7 million deaths have occurred worldwide due to heart disease [3], in which 7.4 million and 6.7 million occurred because of coronary heart disease and stroke [4], respectively.

Heart attack is the one which occurs at time without any request, and it is not predictable by doctors. Designing powerful cardiovascular disease prediction method is significant because of lack of specialist and increase in wrong diagnoses. This has gained the thirst of researchers to develop various data mining and machine learning methods related to medical field. To reveal data secret, multiple hidden layers are used in deep learning method. Also, for other diseases, this method is used for prediction [5]. By using deep neural network, heart disease is predicted at first time. This method is fully based on deep neural network because it is fully automated data processing method. For evaluation, multiple datasets are taken which achieve accuracy when compared to other methods. By predictions about future infectious disease risks we can avoid, general standards for 'good practice' in the development and usage of predictive models would be a good place to start.

Paper organization is as follows: Section 2 describes literature survey, Sect. 3 presents the summarization of literature survey in tabular form and Sect. 4 gives the conclusion.

2 Literature Survey

Heart disease is the one which affects at any time, and many risk factors exist for diagnosing it. Risk factors include weight, sex and age. Smoking, diabetes, obesity and high blood pressure are some of the other risk factors. To diagnose and assess disease type is hard for physicians. For predicting disease, based on the literature, many researchers use deep learning, data mining and machine learning techniques in health care. Nowadays many studies are done on predicting heart diseases. Still today, numerous problems are faced by heart disease. Survey about predicting heart disease with the use of deep learning method is discussed below.

Sharma et al. [6] presented heart disease prediction model. Quality of heart disease classification was improved by deep neural network (DNN). By using support vector machine (SVM), k-nearest neighbor (KNN), random forest and Naive Bayes method, classification was performed. Heart UCI dataset was utilized to determine Talos hyperparameter optimization which was efficient when compared to others [6].

Wang et al. [7] presented Weight-based Multiple Empirical Kernel Learning with Neighbor Discriminant Constraint (WMEKL-NDC) technique for predicting deaths due to heart failure. To find crucial clinical features, this method used feature selection by evaluating F-value. Based on centered kernel alignment method, various weights are given to each empirical kernel space. Finally, neighbor discriminate constraint was combined with multiple empirical kernel methods to discriminate sample data. During March 2009 and April 2016, from Shanghai Shuguang Hospital, extensive experiments are done on clinical data which contain 198 patient records. From the experimental results, it was observed that it attained better prediction performance in hospitals and was more accurate against state-of-the-art approaches. Top ten significant clinical features were together identified to assist clinicians for diagnosing heart failure [7].

Latha et al. [8] detected features related to heart disease and reduced the dimensionality by using feature selection approach. UCI Machine Learning Repository Data were used for analysis. By six ML classifiers, dataset containing 74 features was validated. Chi-square and principal component analysis (CHI-PCA) along with random forests (RFs) have provided an accuracy of 98.7%, 99% and 99.4% for Cleveland, Hungarian and Cleveland–Hungarian (CH) datasets, respectively. From this analysis, features like highest heart rates, cholesterol, chest pain and heart vessels and depression were derived by ChiSqSelector. Experimental results reveal that Chi-square with PCA provided good efficiency. PCA on raw data provided low results, and thus to improve the results, PCA required higher dimensionality [8].

Latha et al. [9] presented ensemble classification for enhancing accuracy of weak techniques by integrating multiple classifiers. From the comparative study, it was discovered that this method enhanced the accuracy in the prediction of heart disease. This work focused in predicting disease at the initial stage. Result of methods such as boosting and bagging increased the accuracy of weak classifiers and performed better in determining heart disease. With help of this method, accuracy of weak classifiers was increased by 7%. Further, its efficiency is increased by feature selection method and shows that it has high prediction accuracy [9].

Ashraf et al. [10] designed DNN method for determining automated prediction of an automated system. It was estimated by using multiple dataset and produced better accuracy. Method offered promised automated method in preprocessing dataset. Under consideration, this method achieved minimum accuracy of 87.64% on any dataset [10].

Mohan et al. [11] suggested machine learning method to determine important features which results in improving accuracy of predicting heart disease. This is achieved with various combinations of features. With the help of hybrid random forest with linear model (HRFLM), proposed method achieved 88.7% of accuracy [11].

Zhongzhi [12] suggested deep learning method which used recurrent neural network with multitask framework to capture high-order and temporal interactions which was avoided by the conventional regression approaches. To calculate prediction accuracy, C-static and mean absolute error were utilized. To determine onset depressive issues in elderly people, machine learning methods were employed. For elderly people, this method was an decision support system to take decision and inform intervention by clinicians [12].

Dinesh [13] implemented data preprocessing methods such as removal of mission and noisy data for decision and prediction at various levels. Using classification, performance measures like sensitivity, specificity and accuracy were evaluated. This method was used to accurately predict heart disease to give diagnosis or awareness on that. By using methods such as SVM, gradient boosting, Naive Bayes classifier and logistic regression, accuracy of methods was compared [13].

Kishore et al., [14] suggested deep learning method, especially RNN to predict heart disease of the patients. RNN is efficient method which used deep learning method in artificial neural network. Based on theory, this paper proposes major system models. To give accurate results with less error, this method combines data mining and deep learning methods. For development of heart attack prediction, this research gives direction and precedent [14].

Ul Haq et al. [15] proposed machine learning method for predicting heart attack using disease dataset. Cross-validation, three feature selections, seven machine learning methods and seven performance evaluation metrics like specificity, Matthews correlation coefficient, accuracy, execution time, sensitivity were used. This method easily classified people with heart disease from normal people. Further, for each classifier, receiver optimistic curves and area of curves were estimated. Its performance was validated for full and reduced set of features. By considering execution time and accuracy of classifiers, feature reduction has greater impact on classifier performance. This method was significant to assist doctors for diagnosing heart patients [15].

Huang [16] implemented stacked denoising autoencoder (SDAE) method to identify the risks of ACS patients from Electronic Health Records (EHRs). To get efficient prediction results, two constraints were added with SDAE, for determining patient's characteristics and preserve data at various levels. This method n was updated by real clinical data consisting of 3463 samples of ACS patient. Performance of this method was 0.868 and 0.73 of AUC and accuracy, respectively. When compared with state-of-the-art approaches, this method achieved competitive performance in clinical heart prediction. Suggestive hypotheses were validated by examining medical domain, but it was not consistent with existing methods [16].

Karthikeyan et al [17] suggested Deep Belief Network (DBN) to predict heart-related diseases. This research describes comparison of DBN with Convolutional Neural Network (CNN) methods. CNN is an unsupervised technique which provided the accuracy of 82% and DNN 90% in prediction of heart disease. Accuracy of predicting heart disease was improved by using this technique [17].

AQSA RAHIM et al. [18] defined that a Machine Learning-Based Cardiovascular Disease Diagnosis (MaLCaDD) framework is proposed for the effective prediction of

cardiovascular diseases with high precision. Here, first deals with the missing values (using mean replacement technique) and data imbalance (using Synthetic Minority Oversampling Technique—SMOTE). Subsequently, feature importance technique is utilized for feature selection. Finally, an ensemble of logistic regression and k-nearest neighbor (KNN) classifiers is proposed for prediction with higher accuracies of 99.1, 98.0 and 95.5 % which are achieved, respectively. Yuepeng [20] constructed a heart disease prediction model based on random forest and LSTM to screen out the main features that may lead to heart disease. Then, LSTM, KNN and DNN algorithms are used to test whether the prediction accuracy is improved or not (Table 1).

3 Summarization Table

Summarization table for literature survey is presented below:

Based on the above literature survey, the following contributions are under deep learning techniques:

In this paper, we deploy a deep learning neural networks (DNNs) using Talos optimization. Talos optimization is newly optimization techniques in DNN. Talos provide better accuracy (90.76%) to other optimizations [6].

Deep Neural Network (DNN) is deep learning mechanism which is used for getting high accuracy for the purpose of predicting heart attack in the patient [10].

The proposed multitask LSTM model can successfully capture high-order and temporal patterns that traditional methods ignore [12].

Deep learning techniques, specifically recurrent neural network to predict the likely possibilities of heart-related diseases of the patient. Recurrent neural network is a very powerful classification algorithm that makes use of deep learning approach in artificial neural network [14].

We adopt the long short-term memory (LSTM) recurrent neural network, a popular deep learning framework for modeling event sequences [16].

This proposed work contains comparison of Convolutional Neural Network [CNN] and Deep Belief Network classification [DBN] algorithms. Convolutional Neural Network algorithm is one of the unsupervised algorithms. It provides 82% of accuracy in the prediction of heart diseases. But, the proposed Deep Belief Network algorithm provides 90% accuracy in heart diseases' prediction which enhances the prediction accuracy of heart disease prediction system [17].

Deep learning is remarkably powerful for solving classification problems, but all problems cannot be represented in classification format. Some of the limitations of common deep learning algorithms are as follows:

- It requires huge amount of data in order to perform better than other techniques.
- It is very expensive to train due to complex data models.
- There is no standard theory to guide you in selecting right DL tools as it requires knowledge of topology, training method and other parameters.

Table 1 Methodologies comparison

Ref. No.	Methods used	Advantage	Disadvantage	Dataset	Accuracy
[6]	Deep neural Network	Attained better accuracy and made the system more efficient	Need to enhance overall quality of heart disease classification	Heart disease UCI dataset, 303 columns and 14 attributes	90.76%
[7]	WMEKL-NDC technique	On mortality prediction, it was more accurate	Based on time duration when heart failure increased, datasets also increased which caused patient death due to other diseases	10,198 inpatients' records were collected from Shanghai Shuguang Hospital in March 2009 and April 2016, 10 crucial clinical features	Top ten F-values
[8]	CHI-PCA with RF	Improved raw data results for heart disease prediction	Due to small sample size, it was difficult to find heart disease	UCI dataset, 74 feature	CHI-PCA with RF had the maximum performance, with 98.7% accuracy for Cleveland, 99.0% accuracy for Hungarian, and 99.4% accuracy for CH
[9]	Ensemble classification method	Prediction of disease at an early stage	Need to improve accuracy by enhancing the feature selection technique	Cleveland heart dataset from the UCI machine learning repository, 14 attributes and 303 instances	The highest accuracy was obtained with majority voting with the feature set FS2
[10]	Deep neural network method for automated system	Improved accuracy while evaluating heart attack prediction	To extend other deep learning algorithms for achieving better results regarding accuracy	UCI repository, 14 attributes and 303 instances	Proposed method is 87.64%
[11]	HRFLM approach	Efficient prediction of heart disease	To predict heart disease efficiency, new feature selection methods are produced to get broader perception	UCI repository, 14 attributes and 303 instances	With an accuracy level of 88.7%

(continued)

Table 1 (continued)

Ref. No.	Methods used	Advantage	Disadvantage	Dataset	Accuracy
[12]	Multitaskdeep learning-based prediction models	For elderly people, it was used to predict onset depressive disorder	It was generic and was applied to other prediction tasks or risk assessment or clinical outcomes	The health and retirement study (HRS) is a longitudinal household survey dataset	The C-statistic (for task 1) and MAE (for task 2) are 0.869 and 1.086, respectively
[13]	Hybrid classification algorithm	Based on present attributes, it was used to predict heart disease uncertainty levels	Its performance was improved by using more parameters	UCI dataset, 74 features	Logistic regression 0.8651685
[14]	Recurrent neural network	Accurate results with minimum errors	Its accuracy was improved but did not provide better results for silent heart attacks	303 records with 75 medical attributes (factors) from the UCI machine learning data repository. After preprocessing, use of 270 records with 13 medical attributes	RNN 92%
[15]	Hybrid intelligent system framework	Helped in reducing execution time and assist doctors to diagnose heart patients	Other optimization and feature selection methods to be used to improve efficiency	303 records with 75 medical attributes (factors) from the UCI Machine Learning Data Repository	Logistic regression, before feature selection 84, after 89 SVM (RBF) 86, 88
[16]	Regularized stacked denoising auto-encoder (SDAE)	When compared with state-of-the-art methods, this method achieved competitive performance	It has missing data problem which makes uncertainty in ACS patients	We validate our approach on a real clinical dataset consisting of 3,464 ACS patient samples	The performance of our approach for predicting ACS risk remains robust and reaches 0.868 and 0.73 in terms of both AUC and accuracy

Table 2 Comparison of various approaches for heart disease prediction techniques

Authors	Year	Approach	Accuracy
Sharma and Parmar [6]	2020	Deep neural network	90.76%
Wang et al. [7]	2020	WMEKL-NDC technique	Top ten <i>F</i> -values
Garate-Escamilla et al. [8]	2020	CHI-PCA with RF	CHI-PCA with RF had the maximum performance, with 98.7% accuracy for Cleveland, 99.0% accuracy for Hungarian and 99.4% accuracy for CH
Latha and Jeeva [9]	2019	Ensemble classification method	The highest accuracy was obtained with majority voting with the feature set FS2
Ashraf et al. [10]	2019	Deep neural network method for automated system	Proposed method is 87.64%
Mohan et al. [11]	2019	HRFLM approach	With an accuracy level of 88.7%
Zhongzhi et al. [12]	2019	Multitaskdeep learning-based prediction models	The C-statistic (for Task 1) and MAE (for Task 2) are 0.869 and 1.086, respectively
Dinesh et al. [13]	2018	Hybrid classification algorithm	Logistic regression 0.8651685
Kishore et al. [14]	2018	Recurrent neural network	RNN 92%
Amin Ul Haq et al. [15]	2018	Hybrid intelligent system framework	Logistic regression, before feature selection 84, after 89 and SVM (RBF) 86,88
Huang et al. [16]	2017	Regularized stacked denoising autoencoder (SDAE)	The performance of our approach for predicting ACS risk remains robust and reaches 0.868 and 0.73 in terms of both AUC and Accuracy
Karthikeyan and Kanimozhi [17]	2017	Deep belief network classification	Deep Belief Network algorithm provides 90% accuracy
Aqsa Rahim et al. [18]	2021	Machine learning based cardiovascular disease diagnosis (MaLCaDD) framework	MaLCaDD predictions are more accurate (with reduced set of features)
Yuepeng et al. [19]	2020	Random forest and LSTM algorithm	87%

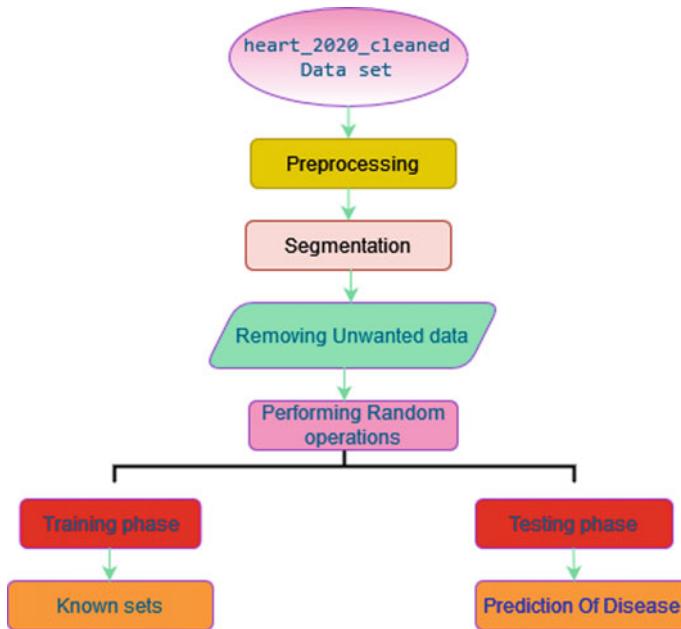


Fig. 1 Framework of heart disease

4 Problems Identified

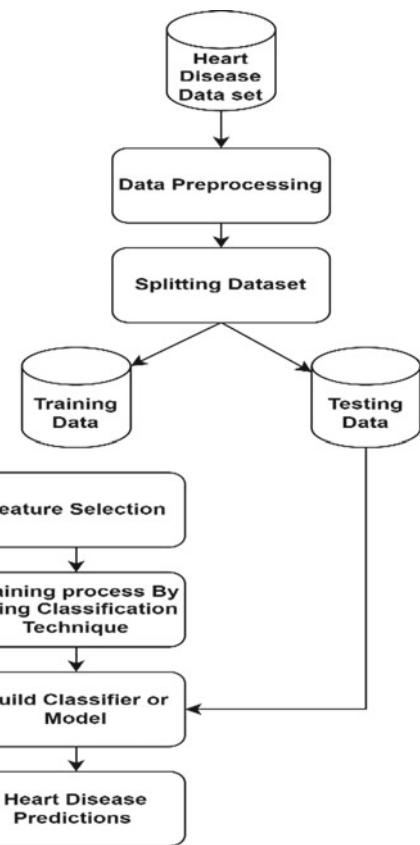
From the several investigations made, it is observed that cardiovascular risk prediction has to be done at the initial stage. Moreover, statistic adoptive approach could increase the clinical risk prediction rate. Further, machine learning techniques can be used to improve the predictive accuracy of all models examined. The major issues were there exists no exact diagnosis of the disease at the early stages where only a single classifier with categorical data was utilized for diagnosis and prediction of the disease. Also, time taken for diagnosis the disease has to be reduced while improving the prediction accuracy and confidence level for the exact prediction.

For this problem, following Fig. 1 shows process of heart disease predictions (Fig. 2).

5 Conclusion

Data mining can be of very knowledge form such suitable dataset. This paper provides survey of heart disease prediction by using data mining method. Different research reviews on predicting heart disease adopting various machine learning, deep learning and data mining methods were studied. In some studies, for single dataset with limited

Fig. 2 Model of heart disease prediction



heart disease features, other data sources are used. Different classification accuracies obtained from heart disease could not be generalized. Finally, this paper provides insight and summary of existing works. To enhance accuracy and scalability of prediction model, there are numerous possible improvements determined.

In future, we will propose methodology for early prediction of heart disease by using hybrid ML and DL techniques with high accuracy and minimum cost and complexity.

References

1. Liu X, Faes L, Kale A (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health* 1(6):271–297
2. Ravindhar NV, Anand HS, Ragavendran GW (2019) Intelligent diagnosis of cardiac disease prediction using machine learning. *Int J Innov Technol Exploring Eng* vol 8, no 11

3. <http://www.who.int/mediacentre/factsheets/fs317/en/>
4. Dangare C, Apte S (2012) A data mining approach for prediction of heart disease using neural networks. *Int J Comput Eng Technol*
5. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
6. Sharma S, Parmar M (2020) Heart Diseases Prediction using Deep Learning Neural Network Model. *Int J Innov Technol Exploring Eng (IJITEE)*, vol 9, no 3
7. Wang Z, Wang B, Zhou Y, Li D, Yin Y (2020) Weight-based multiple empirical kernel learning with neighbor discriminant constraint for heart failure mortality prediction. *Elsevier J Biomed Inf* vol 101, no 103340
8. Gárate-Escamila AK, El Hassani AH, Andrès E (2020) Classification models for heart disease prediction using feature selection and PCA. *Elsevier Inf Med Unlocked*, no 100330
9. Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inf Med Unlocked* vol 16, no 100203
10. Ashraf M, Rizvi MA, Sharma H (2019) Improved heart disease prediction using deep neural network. *Asian J Comput Sci Technol* vol 8 no 2, pp 49–54
11. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7:81542–81554
12. Zhongzhi X, Zhang Q, Li W, Li M, Yip PSF (2019) Individualized prediction of depressive disorder in the elderly: a multitask deep learning approach, *Int J Med Inf* 103973
13. Dinesh KG, Arumugraj K, Santhosh KD, Mareeswari V (2018) Prediction of cardiovascular disease using machine learning algorithms. In: International conference on current trends towards converging technologies (ICCTCT), pp 1–7
14. Kishore A, Kumar A, Singh K, Punia M, Hambir Y (2018) Heart attack prediction using deep learning. *Int Res J Eng Technol (IRJET)* 5(4)
15. Haq AU, Li JP, Memon MH, Nazir S, Sun R (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithm. *Hindawi J Mobile Inf Syst* Article ID 3860146, pp 1–21
16. Huang Z, Dong W, Duan H, Liu J (2017) A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records. *IEEE Trans Biomed Eng* 65(5):956–968
17. Karthikeyan T, Kanimozhi VA (2017) Deep learning approach for prediction of heart disease using data mining classification algorithm deep belief network. *Int J Adv Res Sci Eng Technol* 4(1)
18. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW (2021) An integrated machine learning framework for effective prediction of cardiovascular diseases. *IEEE Access* vol 9, pp 106575–106588
19. Liu Y, Zhang M, Fan Z, Chen Y. (2020) Heart disease prediction based on random forest and LSTM. In: 2020 2nd international conference on information technology and computer application (ITCA). <https://doi.org/10.1109/ITCA52113.2020.00137>, pp 630–635
20. Anand D, Arulselvi G, Balaji GN (2022) An assessment on bone cancer detection using various techniques in image processing. In: Applications of computational methods in manufacturing and product design. Springer, Singapore, 523–529
21. Avinash P, Venkateswarlu T, Anand D (2008) A detail study on biometrics with Matlab. *Int J Eng Technol (UAE)* 7(2.20 Special Issue 20), pp 243–249
22. Srinivasu SVN, Venkateswarlu T, Avinash P (2018) A valuable role of digital payments in building smart cities using IoT technology. *J Adv Res Dynamical Control Syst* 10(2):1890–1896

Evaluation of Deep Learning Technique on Working Model of Self-driving Car—A Review



Somin Sangwan, Gurpreet Singh, Aashima Bagnia,
and Vishwajeet Shankar Goswami

Abstract Artificial intelligence is revolutionising the way we are living. Autonomous vehicles are going to play a major role in that. According to US Department of Transportation, Ohio University and The UK Economic Opportunity self-driving cars will reduce traffic deaths up to 90%, drop of 60% in harmful emissions and reduction in travel time by 40%, respectively. Its a key technology which will help us out in reducing various problems to a larger extent. This is a self-explanatory paper which provides an overview of the functioning of self-driving cars with the use of deep learning. This paper will first make you understand what basically we mean by when we call a car self-driving along with the meaning of deep learning. This will be followed by providing the details on how deep learning is involved at various steps like how the car is able to perceive, localise, use various deep learning algorithms like Markov decision process (MDP) and neural networks like convolutional neural networks (CNNs), recurrent neural networks (RNNs), etc., to predict various possible actions it can take based on the situation and finally with the help of Bayesian optimisation choosing the best possible outcome. This paper will easily make you understand the functioning of self-driving cars and technology behind them.

Keywords Deep learning · Self-driving · LiDAR · RADAR · Perception · Localisation · Decision-making · Sensor fusion

S. Sangwan (✉)

Data Science, Department of Mathematics, Chandigarh University, Mohali, India
e-mail: sominsangwan97@gmail.com

G. Singh

Department of Computer Science and Engineering, Chandigarh University, Mohali, India

A. Bagnia · V. S. Goswami

Department of Mathematics, Chandigarh University, Mohali, India
e-mail: vishwajeet.e9858@cumail.in

1 Introduction

With the advancement in technology, the performance of graphic cards, microprocessors, CPUs and GPUs have risen to such a great extent that they can process huge amounts of data within microseconds. These processors with tremendous power have revolutionised the way we live, and it is going to remain the same. In today's era, self-driven cars have become much powerful than ever. Basically, self-driven cars are the autonomous vehicles capable of sensing the environment, taking decisions of its own and moving safely without or with very less human intervention [1]. These vehicles are integrated basically with the sensors like camera, RADAR and LiDAR from which various inputs are taken which are then processed through deep neural networks and output is generated, which is basically the decision taken by these AI systems [2]. Deep learning algorithms process data through several layers of neural networks, in which each layer passes simple representation of data to the next layer, where layers at the beginning detect the low-level features and subsequent layers combine features extracted from the early layers to build a more holistic representation [3]. As, these are multi-layered networks, hence called deep and are termed neural networks because they mimic the working of brain cells, i.e. neurons [4, 5].

Here, the system contains nodes as the basic units which act as neurons of the human brain. As shown in Fig. 1, these nodes are connected to each other layer-by-layer where some of them are labelled whereas some of them are not, but all together they are grouped into layers. A process takes place in these nodes when a stimulus

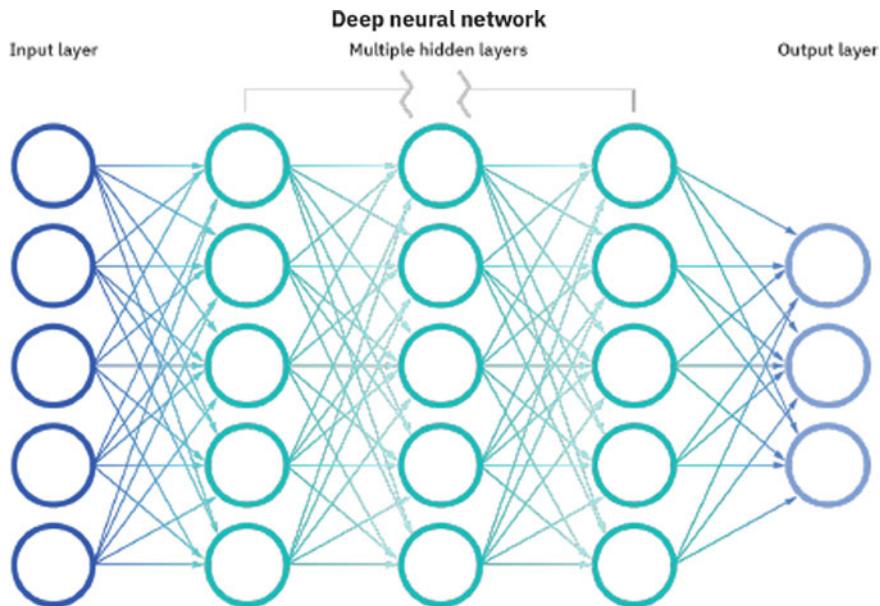


Fig. 1 Deep neural network [4]

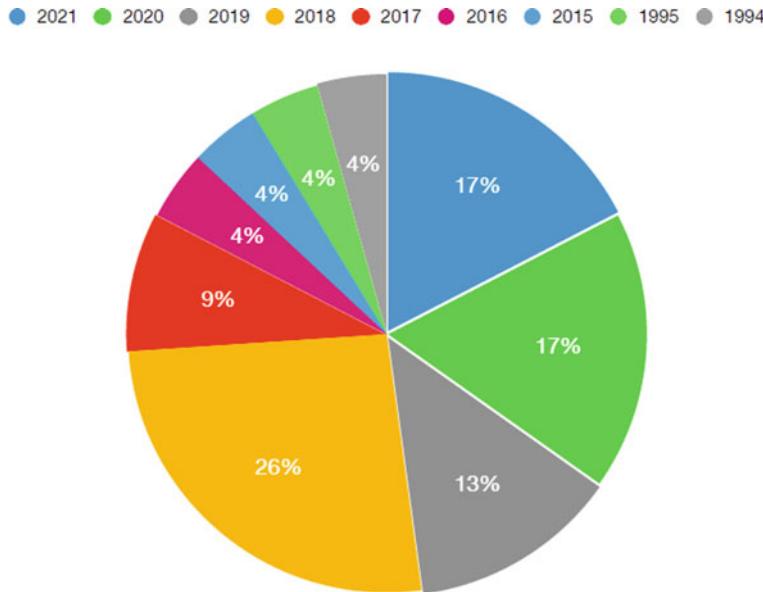


Fig. 2 Year-wise bifurcation of papers

hits them; input data from one end is processed through these layers to get the results [6]. The more the layers the input has to pass through, the deeper the network is considered.

2 Related Work

The literature survey briefly addressed the deep learning techniques are given in Table 1. In relation with the connected work and for taking reference, research papers, books, journals and reports were used from SocTA, IEEE, ACM, Elsevier, Springer and Google Scholar. Year-wise bifurcation of papers is as shown in given Fig. 2.

3 Existing Methodology/Models for Performance Evaluation in Self-driving: Survey

Table 2 describes the various methodology and models used by different researchers related to this research including output of those models, indicating about the success rate of a particular model based on particular algorithms and frameworks used.

Table 1 Survey mechanism of self-driving cars using deep learning

Author(s)	Focus of the paper	Technique used	Parameter analysed	Relevance in the work
Yeong et al. (2021)	Sensor fusion technology in autonomous vehicles	High level fusion, mid level fusion	High quality image resolution, object detection, perception	Fusion of camera and LiDAR to obtain spatial information
Paravarzar et al. (2020)	Autonomous vehicle motion prediction	Long-short term memory, convolutional neural networks, recurrent neural networks	Behaviour prediction longitudinal and lateral trajectory	Prediction about the possible moves of other vehicles
Al-Nima R. et al. (2019)	Road tracking with the use of deep reinforcement learning	Deep reinforcement learning—road tracking (DRL-RT)	Accuracies of various deep learning networks on SYNTHIASEQS-05-SPRING database	Use of reinforcement learning algorithms like MDP and Bayesian optimisation to make decisions
Coskum S. et al. (2018)	Lane changing in a highway	Markov decision process (MDP), fuzzy Markov decision process (FMDP)	Trajectory generation/control accuracy of decision taken by the self driving car	Best decisions are taken out of all possible n-moves with the help of Markov decision process
Zhao J. et al. (2018)	Four keys areas navigation, perception of surroundings planning the path which needs to be followed, controlling the car	Analysed methodology of various research institutions of selfdriving cars	Cars navigation system, location system, electronic map, map matching laser perception etc.	Some of the concept of motion planning is being taken from this survey paper
Häne C. et al. (2017)	Use of multi-camera system for visual localisation 3 d mapping and obstacle detection	Slam-based extrinsic, joint optimisation, sparse mapping, ego motion detection	Calibration of multicamera system, Sparse Map	Use of multicamera's for 3d mapping of surrounding for perception, visualisation and localisation
Paden et al. (2016)	Motion planning and various control techniques for self driving vehicles for urban areas	Motion planning and vehicle control based on prediction and decision making	Accuracy of prediction of steering angles, braking, parking etc.	Decisions on speed of the car, changes of the lanes overtaking decelerating in section under motion planning
Li C. et al. (2015)	Path planning for dynamic environments using model based approach	Model based trajectory generation candidate trajectories generation	Cost of threat probability area, cost of deviation from centreline trajectory	It is used in path planning section of the paper where once the vehicle has localised itself, them sees all the possible routes and plans the best out of them
Anderson, J. A. et al. (1995)	In depth explanation of neural networks and their functioning	Artificial neural networks (ANNs), convolutional neural networks (CNN), recurrent neural networks (RNN)	Results of ML and DL algorithms	Definition of neural networks and deep learning

Table 2 Survey related to performance of various deep learning techniques

Author	Model	Performance
Häne C. et al. (2017)	SLAM-based extrinsic calibration, structure-based calibration	Parameters analysed = 9, mean error = 7.06 cm
Verucchi, M. et al. (2020)	Density-based spatial clustering of applications with noise (DBSCAN)	Miscalibration error (EP%) = 6
Al-Nima, R. et al. (2019)	Deep reinforcement learning-road tracking (DRL-RT)	Driving accuracy = 94.37%
Al-Nuaimi, R. et al. (2021)	Probabilistic model checker (PRISM), probabilistic timed programme (PTP)	Maximum collision probability (i) pedestrian = 0.252, (ii) car = 0.003
Li C et al. (2015)	Hermite interpolation-based connecting trajectory	Car reaches the destination successfully for both straight and curved road
Verucchi, M et al. (2020)	Frameworks: darknet, deep steam, tkDNN (FP32, FP16)	Frames per second (fps): Yolov2—10.02 Yolov2-tiny—67.06 Yolov3—7.44

4 Working of Self-driving Cars

A series of steps are performed by these autonomous vehicles to reach the desired location safely. To understand the working of these cars, we will be basically examining four major parts.

4.1 Perception

Perception is basically making the car see the world around itself [7], and along with that being able to recognise and classify the things it sees [8]. Here, perception is way more than its general meaning, with perception we not only mean recognising and classifying but also being able to evaluate the distance and take decisions like whether to accelerate or slow down depending upon the situation. To perceive the environment and take inputs from its surrounding, the car uses three sensors:

- Camera
- LiDAR
- RADAR.

Camera: Cameras are basically the eyes of the car; they provide vision to the car. The input received from these cameras is used to interpret the environment details



Fig. 3 Cameras of a self-driving car [9]

like identifying sign boards, traffic lights, pedestrian crossing, etc. In order to get a complete 360-degree view, as shown in Fig. 3, cameras are integrated on every side, i.e. front, back and on the sides. These cameras are of very high resolution which not only take visuals but also take audio input from the environment to assess the condition of the environment [9]. After identification of images and videos is done, then deep learning algorithms do classification and segmentation of the data.

Cameras do the perception task very efficiently when the weather is good, the environment is clear, but are unable to provide clear visuals when the conditions are foggy, or there has been heavy rain and also during night time. During these extreme conditions, cameras ability to capture images, enabling lane detection, reading various sign boards, identifying peoples, animals, etc., deteriorates, and all it captures a lot of outside noise and discrepancies, which hampers the quality of input received and henceforth, the decision-making which eventually leads to hamper safety of passengers. To overcome such a situation, we need sensors which can function without light and being able to create a blueprint of the object and measure distances accurately.

LiDAR: LiDAR stands for light detection and ranging, also laser imaging, detection and ranging, is a special 3D-laser scanning system made from combination of 3D-scanning and laser scanning [10]. LiDAR sensor is equipped with the scanner, laser and a GPS receiver. Laser beam is fired at an object, and distance is measured by calculating the time it takes to return back to the LiDAR source. To provide depth to

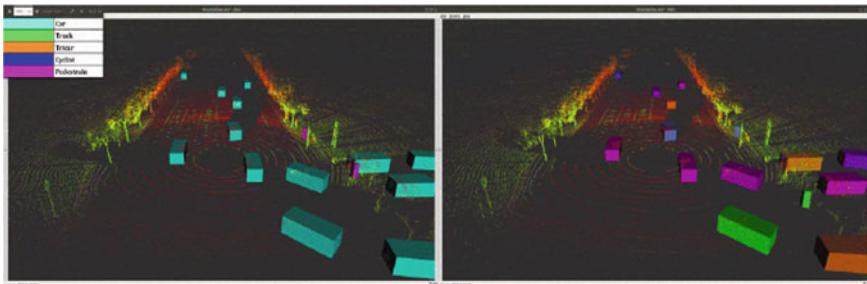


Fig. 4 Object detection using LiDAR [10]

the images captured by the cameras and create a 3D perception (as shown in Fig. 4) of what's going around the car, camera sensor is fused with LiDAR sensor with the use of deep learning and is called sensor fusing [9, 11]. Now, the 2D and 3D information received from camera and LiDAR is turned into spatial information. And then, this data is fed into deep neural networks, and the systems can more precisely predict the actions of nearby or vehicles close to it. But, one cannot completely rely on LiDARS also, as LiDARS have limitations too that can be catastrophic too. Like, LiDARS use laser beams for their functioning which works very well in night and dark environments but may fail to work due to noise from heavy rains or fog [12].

Now, to overcome this, we need one more sensor to perceive more accurately and that is RADAR. As, Elon Musk also openly stated that 'anyone relying on LiDARS are doomed'. It clearly indicates when the leader of self-driving car says such thing, means we need some other instruments too, can't rely on LiDAR alone.

RADAR: We have seen that LiDAR used laser beam to evaluate distance and identify objects, whereas in RADAR as the name suggests radio detection and ranging; therefore, it uses radio waves to detect the location and speed of the object [13]. RADARs have a very high efficiency because unlike lasers, radio waves can work in any type of conditions, whether it be rainy, cloudy, foggy, dark, or very low-visibility conditions and compatibility of LiDAR and radar shown in Fig. 5. This is reason they are being used by militaries all over the world [14].

Along with the great compatibility and high efficiency of RADARs to work in mostly all type of weather conditions, one thing to note is that RADARs are the noisy sensors. Hence, they capture a lot of unnecessary noise which needs to be filtered, and the data needs to be cleaned to make appropriate decisions. The image in Fig. 6 compares the data received from a LiDAR sensor and RADAR sensor of a self-driving car in green colour. As, it clear from the image that radar has captured much more unnecessary noise from its surrounding which needs to be filtered. The data from the radar sensor is cleaned by a process called thresholding, in which weak signals are separated from strong ones and a clean input is fed to the deep learning algorithms. After the input is cleaned and processed through deep learning algorithms, outputs from these LiDAR and RADAR are generated in the form of point-based data as shown in Fig. 7.

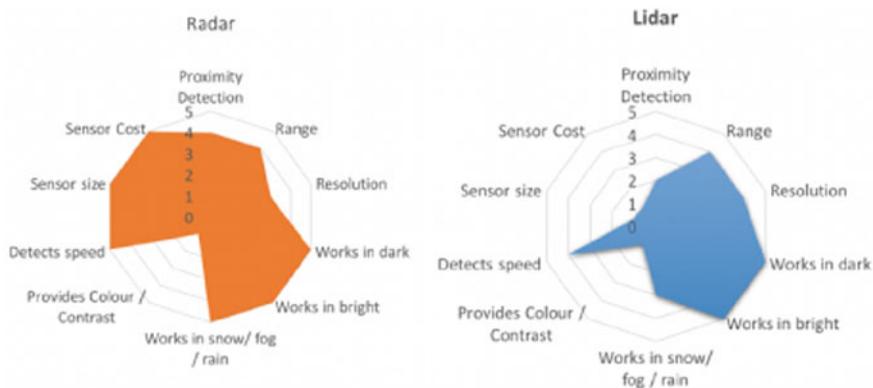


Fig. 5 Compatibility of LiDAR and RADAR in different conditions [13]

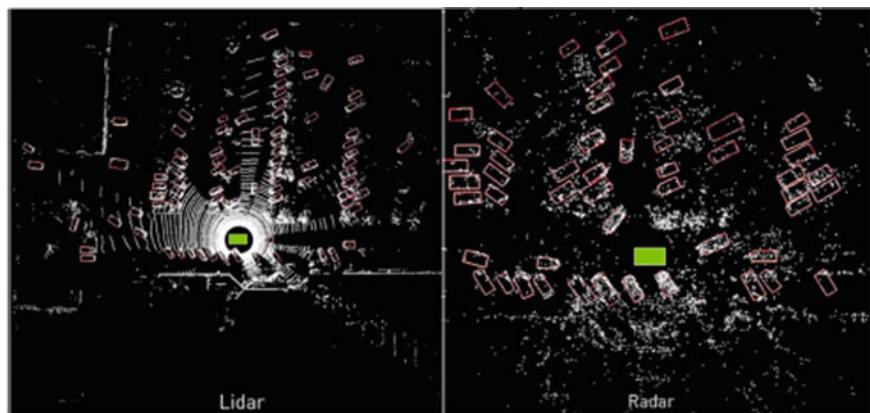


Fig. 6 Comparison of data received

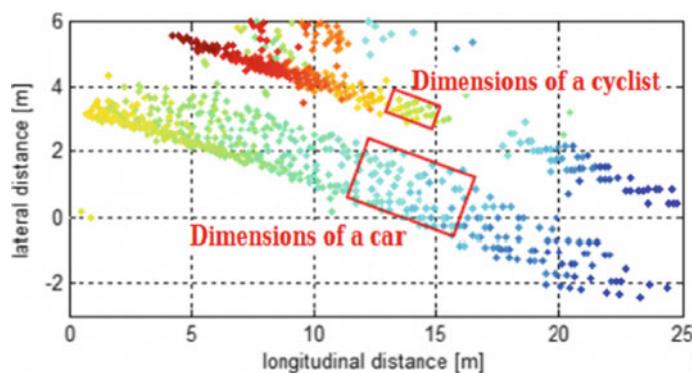


Fig. 7 Point-based output generated from LiDAR and RADAR [13]

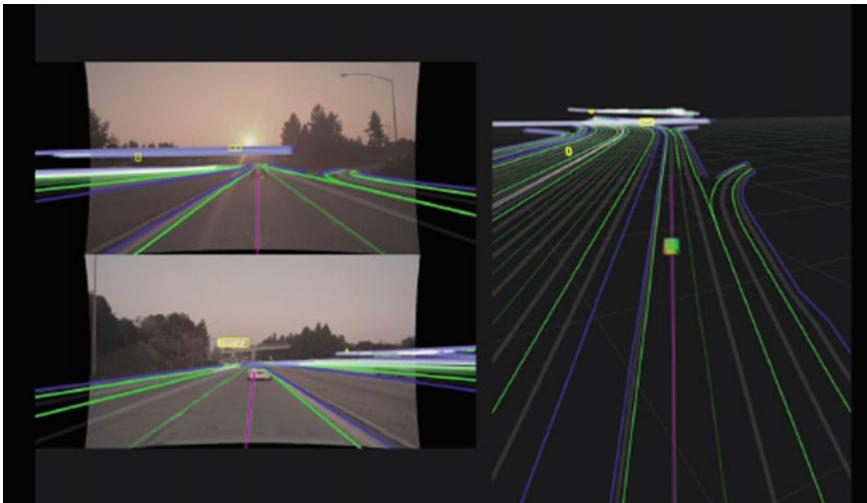


Fig. 8 Localisation in self-driving cars [15]

4.2 *Localisation*

Localisation is basically determining the vehicle's precise position on map, when the vehicle knows where exactly it is and is able to relate that with actual surroundings and with the map in real-time scenario as depicts in Fig. 8. Localisation is important because based on this other components of the autonomous vehicle are stacked to take decisions like, e.g. path planning can be done based on localisation only, once the vehicle knows where it is and where to go, then it decides the best suitable path from that place. Along with planning the path, when we have a constant track of the vehicle, then based on chosen path, the vehicle is aware of the route, turns, sign boards, etc., which are going to come in the chosen path and can be pre-plan to act accordingly. This makes the drive even more smoother and more safer as the car is already aware of what is coming in front.

Now, you might struck with the question that we all have smartphones and cars coming with builtin GPS, so why not use GPS for localisation purpose. The answer to this is that yeah, we can keep a track of where the vehicle is, but not exactly where the vehicle is. Imagine how much is 1–2 m and think of the situation of the vehicle being wrong up to 1–2 m of where it actually is. I think you have gotten your answer by thinking of the possibility of getting a localised position which is 1–2 m wrong only. Major technique used for localisation of self-driving cars is matching, wherein the vehicle matches what it sees through perception with the map it is following.

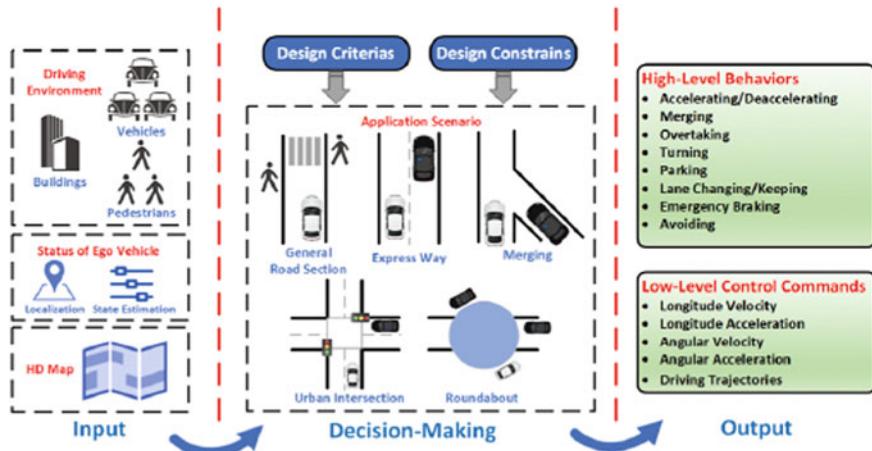


Fig. 9 Prediction in self-driving cars [18]

4.3 Prediction

After perceiving the environment, by the help of sensor fusion and localisation, now its time for the self-driving car to make predictions about the possible moves which other vehicles, pedestrians or nearby objects in its surrounding can make [15, 16]. For a self-driving vehicle to operate safely, it is very important to predict the future behaviour of other road users. Based on the possible moves by other road users, self-driving car will take decisions which will avoid collisions and hence, enhance the road safety. With the help of the camera, LiDAR and RADAR sensors and the fusion of these sensors with each other. Then, this information is fed to to deep learning algorithms like Markov decision process (MDP) [17], and it comes up with all the best possible n no. of moves which the other road users can take. The scenario become very complex when the number of moving vehicles is very high in number, which increases the number of possible moves for the self-driving car itself. Now, as the system has predicted what all moves other road users can take, its time to decide the best possible move for itself, so that it can avoid any collisions and manoeuvre to the desired destination as given in Fig. 9.

4.4 Decision-Making

It is a vital process in self-driving cars. The system needs to be very precise, dynamic and be able to make correct decisions in real-time scenarios, i.e. microseconds. All the data received from the sensors is considered to be correct and the fact that human can make choices which can't be predicted beforehand are taken into consideration

[18]. These uncertain human choices can be measured with good accuracy, and appropriate decisions can be taken accordingly, but can't be predicted with much accuracy. These AI systems use deep reinforcement learning (DRL) [19, 20] to make decisions. Earlier we have read that input data is processed through Markov decision process (MDP), and finite number of moves are predicted by the system. Now, to select best possible move for itself out of these n no. of moves, the deep learning model is optimised with Bayesian optimisation. The Bayesian optimisation [21] helps system to select the best possible solution of the solutions provided by Markov decision process. The decision-making is a hierarchical process, wherein the self-driving vehicle has to make decisions at several steps to complete all the kinematic manoeuvres. They are four general components to complete this process, which are as follows:

Path Planning: Planning the route which the car is going to take is the first of the four decision-making stages [22]. Once the vehicle has localised itself and knows where it is and where to go, then it sees all possible routes and decides the best amongst them.

Behaviour Arbitration: Once the route is decided, now, the car about all the static elements like road, traffic lights, signs, pedestrian crossing, intersections, road congestions, etc., but can't know exactly about the actions which will be taken by other road users through out the route [23].

Motion Planning: Once the route is set and behaviour layer has decided how to navigate through that set route with all the static and dynamic elements in the route, then the motion planning system comes into picture and sets the motion of the car [24]. This layer decides the speed of the car, changing of the lanes, overtaking other vehicles, decelerating in front of other objects or according to sign boards, etc.

Vehicle Control: Now, in the last stage, the vehicle moves on to the path referred by the motion planning system [25]. Basically, here, the vehicle acts according to the commands received from the motion planning system and continues its journey accordingly on the referred route.

5 Conclusion

In this paper, we reviewed and studied about the use of deep learning in perception, localisation, prediction and decision-making in self-driving cars. With the advancement in CPUs, GPUs and microprocessors, the deep learning algorithms like CNN, RNN, MDPs, Bayesian optimisation etc., can process results within real-time scenarios, and instant accurate decisions can be taken. These advancements in processing power of microprocessors have given a huge boost to self-driving cars industry, and the way, deep learning is being explored and studied by researchers nowadays, it will not take much time to improve these neural networks used in the field of self-driving cars, which in turn has made us believe that fully developed SAE level 6 automated

cars are on the verge of becoming reality. Therefore, now, the question of self-driving cars is not, of if and how, but, of when. As the tech giants like Tesla, Google, Nvidia have already given the answer of if and how by innovating such amazing self-driving cars, which shows that we are not far away from developing SAE level 6 automation vehicles.

References

1. Lal AM, Aju D (2021) Deep learning models for object detection in self-driven cars. In: *Integrating deep learning algorithms to overcome challenges in big data analytics*. CRC Press, pp 17–38
2. Batista KBDSL (2018) Self-driven cars, self-driven patients, and company-driven orthodontists? *Angle Orthodont* 88(6):841
3. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. *IEEE Access* 7:53040–53065
4. Anderson JA (1995) *An introduction to neural networks*. MIT Press
5. Giri JP, Giri PJ, Chadge R (2018) Neural network-based prediction of productivity parameters. *Soft computing: theories and applications*. Springer, Singapore, pp 83–95
6. Sarle WS (1994) *Neural networks and statistical models*
7. Sheth S, Ajmera A, Sharma A, Patel S, Kathrecha C (2018) Design and development of intelligent AGV using computer vision and artificial intelligence. *Soft computing: theories and applications*. Springer, Singapore, pp 337–349
8. Häne C, Heng L, Lee GH, Fraundorfer F, Furgale P, Sattler T, Pollefeys M (2017) 3D visual perception for self-driving cars using a multi-camera system: calibration, mapping, localization, and obstacle detection. *Image Vis Comput* 68:14–27
9. Wang Z, Wu Y, Niu Q (2019) Multi-sensor fusion in automated driving: a survey. *IEEE Access* 8:2847–2868
10. Hecht J (2018) Lidar for self-driving cars. *Opt Photonics News* 29(1):26–33
11. Yeong DJ, Velasco-Hernandez G, Barry J, Walsh J (2021) Sensor and sensor fusion technology in autonomous vehicles: a review. *Sensors* 21(6):2140
12. Verucchi M, Bartoli L, Bagni F, Gatti F, Burgio P, Bertogna M (2020) Real-time clustering and LiDAR-camera fusion on embedded platforms for self-driving cars. In: *2020 fourth IEEE international conference on robotic computing (IRC)*. IEEE, pp 398–405
13. Zhaohua L, Bochao G (2020) Radar sensors in automatic driving cars. In: *2020 5th international conference on electromechanical control technology and transportation (ICECTT)*. IEEE, pp 239–242
14. Muskan, Singh G, Singh J, Prabha C (2022) Data visualization and its key fundamentals: a comprehensive survey. In: *2022 7th international conference on communication and electronics systems (ICCES)*, pp 1710–1714. <https://doi.org/10.1109/ICCES54183.2022.9835803>
15. Paravarzar S, Mohammad B (2020) Motion prediction on self-driving cars: a review. *arXiv preprint arXiv:2011.03635*
16. Singh J, Singh G, Bhati BS (2022) The implication of data lake in enterprises: a deeper analytics. In: *2022 8th international conference on advanced computing and communication systems (ICACCS)*, vol 1. IEEE, pp 530–534
17. Coskun S, Langari R (2018) Predictive fuzzy Markov decision strategy for autonomous driving in highways. In: *2018 IEEE conference on control technology and applications (CCTA)*. IEEE, pp 1032–1039
18. Al-Nuaimi M, Wibowo S, Qu H, Aitken J, Veres S (2021) Hybrid verification technique for decision-making of self-driving vehicles. *J Sens Actuator Netw* 10(3):42

19. Al-Nima RRO, Han T, Chen T (2019) Road tracking using deep reinforcement learning for self-driving car applications. In: International conference on computer recognition systems. Springer, Cham, pp 106–116
20. Singh J, Singh G, Verma A (2022) The anatomy of big data: concepts, principles and challenges. In: 2022 8th international conference on advanced computing and communication systems (ICACCS), pp 986–990. Conference series, vol 1000, No 1, p 012101. IOP Publishing. <https://doi.org/10.1109/ICACCS5159.2022.9785082>
21. Deshwal A, Simon CM, Doppa JR (2021) Bayesian optimization of nanoporous materials. *Mol Syst Des Eng* 6(12):1066–1086
22. Li C, Wang J, Wang X, Zhang Y (2015) A model based path planning algorithm for self-driving cars in dynamic environment. In: 2015 Chinese automation congress (CAC). IEEE, pp 1123–1128
23. Grigorescu S, Trasnea B, Cocias T, Macesanu G (2020) A survey of deep learning techniques for autonomous driving. *J Field Robot* 37(3):362–386
24. Paden B, Čáp M, Yong SZ, Yershov D, Frazzoli E (2016) A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Trans Intell Veh* 1(1):33–55
25. Zhao J, Liang B, Chen Q (2018) The key technology toward the self-driving car. *Int J Intell Unmanned Syst*

Analysis of the Distractions in Youth Due to Social Media and the Effects on Their Concentration Abilities



S. Prajwal, N. Aditi, Dharithri B. Sharma, S. Syed Afreeth, K. Ashwini, and Srirupa Guha

Abstract Distraction is a phenomenon that diverts one's attention from a desired task. It includes external factors such as addiction toward social media, web surfing, electronic gadgets, texting, and also, the internal factors like laziness, nervousness, lack of concentration, etc. Sometimes, distraction can be useful to divert from a trauma or misery. However, it creates an adverse effect on the person's mental health by causing anxiety, depression, behavioral changes, loss of interest in humans which affects their career and even leads to death. This study mainly focuses on distractions in youth and its effects on their personal mental health. A quantitative approach was used in collecting and analyzing the data by distributing google forms and verbal response from to the students. The results obtained were analyzed using various machine learning classifiers. Among different classifiers, Logistic Regression obtained the result of 94.68% accuracy and also analyzed on various factors which can cause distraction in youth.

Keywords Pandas · NumPy · Matplotlib · Distraction · Anxiety · Concentration · Gadgets · Technology

1 Introduction

Distraction is a state which diverts one's attention away from doing their present task. Healthy distraction could be useful to deviate from trauma or to elevate the mood. However, it negatively impacts mental health, thereby causing anxiety, depression, fear of missing out. The recent pandemic situation, due to the outbreak of COVID, has affected the youth by switching their mode of learning from offline to online mode of learning. This has increased youth's level of addiction to social media because

S. Prajwal · N. Aditi · D. B. Sharma · S. S. Afreeth · K. Ashwini (✉)

Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

e-mail: dr.ashwinik@gat.ac.in

S. Guha

National Institute of Technology Durgapur, Durgapur, India

of which students get more prone to distractions urging them to view social media more often during classes and also leading to less focus on studies, thus affecting their academic performance. The research model done by Sumich et al. [1] was used to study “Problematic Internet Use (PIU).” According to this, a study was conducted on the behavior of youth which showed the intensity of usage of internet by using tracking application installed on the phone which is also supported by Christakis et al. [2].

Social media is a very important medium for every youth. It is the biggest source which can give instant solution for any queries or subject, but some subjects may be irrelevant to the youth with respect to his academics, law and order, political and economic developments of his surroundings which may take away his precious time in a day, but he invests his time in it due to the fear of missing out on current affairs as states by Elhai et al. [3].

Smart phones are the most common access electronic infotainment device with approximately 31% of the world’s population having access, so the social media more or less has become nearly a basic commodity to the larger part of the population in the world as the number of smartphone users are increasing exponentially day after day according to as per the research done by Gowthami et al. [4].

Most of the gadgets such as smartphones, smart watches, video games, laptop have become the main source for distracting the youth. As they have been made available in compact nature having internet facility, social medias such as Instagram, WhatsApp, Twitter, Facebook, the youth are very much dependent on these all of the time, viewing and reacting to the notifications popping up on their gadgets. This distracts their attention frequently in classrooms, family functions, public gatherings keeping them absent minded in their present physical place according to Ishii et al. [5] and Throuvala et al. [6].

Distraction can be even caused by the surrounding environment, in which they are brought up which includes friends, family background which can lead them to lose focus on their academics affecting their performance and intern making it difficult to achieve goals, as cited by Mayildurai et al. [7].

Meeting friends, relatives and spending quality time in their leisure hours are very much essential to all, but when a youth looks into it as entertainment factor of life and ignores the self-development aspect and as a result, his or her social life may end up in lagging behind in nurturing his potential and in turn affecting his performance in academics.

This study mainly focuses on various factors leading youth to various distractions, i.e., social media, gadgets, environmental conditions, outing. It analyzes the level of distraction the youth undergoes from the above factors.

2 Literature Survey

The literature review covers the analysis and fact-based research references on the factors contributing to the distraction troubles in youth and its outcome on their concentration pattern.

As stated by Bharadwaj [8], education is such an inseparable trait for each individual to live in this quick paced society. It is the basic foundation for a nice livelihood. Nowadays, education has become a criterion for judging the intelligence and behavior of a person. Education is not only what is thought in schools and colleges, it is what we learn newly in our daily lives too.

Learning methods have seen a slow transition from pedagogy to a visual-based learning method. Many students have cited that social media and the digital forum have played a huge part in their academic success and financial success. The digital forum has acted as a guide to many academic professionals, according to Joanne et al. [9].

Social media, which a large chunk of the internet consists of, is also a form of learning and digital learning through different social mediums which have become a stir-causing revolution. In this social network, people can interact with any other random person on his fingertips. But, we know that every boon comes with a bane. In this case, usages of social media in classrooms have become a spot of concern to the teaching faculties of various educational institutions as per Antoine et al. [10].

3 Effect of Social Media in Academic Aspects

According to a study conducted by Kuppuswamy et al. [11] and Tayo [12], more than one-third of their responders have used the internet for unproductive work and more than 75% of the people while away their time on social media/chatting, etc. This has a direct result in the form of a dip in the students' academic performance. The previous sentence is backed up by the survey conducted by Singh et al. [13], Siebers et al. [14] and Yeboah et al. [15], which showed that active usage of social media sites and applications in the classroom affected the holistic academic performance of the individual, which is explained by Agarwal et al. [16].

Most of the times, the usage of social media is overdone without our knowledge itself. This over usage of technology can have adverse effects on the academic persona and mental health of the subject and is prone to problems like depression and fatigue in the long run as stated by Al Meyanes et al. [17] and Abi-Jaoude et al. [18]. It can also cause serious health issues like obesity problems, eye strains, disturbance in the individual's regular sleep cycle, etc.

As mentioned in the above content, most of the adolescents end up being addicted to the social interfaces as they are driven by a common phenomenon known as Fear of Missing Out (FOMO). This puts pressure on the subject to log on to their social media accounts in order to know the trending affairs which are socially popular at that period of time as mentioned by Gannamani et al. [19, 20, 21]. Being too much addicted to social media can also affect the person's social behavior and relationships and can imply the likely qualities of self-obsession, lack of empathy, and carelessness toward their real life on goings and their physical and mental well-being according to James et al. [22–24].

In the upcoming section, the results of the survey done and the methods used for it will be explained in detail and a conclusion will be drawn.

4 Methodology

The roadmap followed for obtaining a final overview and analysis on the problem statement is as follows.

1. A series of questions were collectively discussed and set after.
 - i. Comprehending/examining the aforementioned topic thoroughly.
 - ii. Conducting a thorough research through various sources like informational websites and vastly open amount of available research/survey papers.
2. The questions were later put out and shared through **GOOGLE FORMS** and other mediums [25, 26]. Answers were also acquired through verbal communication.
3. A total of approximately 626 responses were recorded over a certain span of time and were saved in **.csv**, **ipynb**, and **.py** formatted files and were categorized into three different categories.
4. There were implementations of different machine learning models like K-Nearest Neighbor, Logistic Regression, and Naïve Bayes Classification method which were used to get a better understanding of the concentration pattern and distraction frequencies of the various student communities.

Machine learning models were implemented to the above dataset in order to understand the data in a better and conclusive way.

5. New data were entered, and the prediction for that data was registered. This step is repeated every time a new set of data is introduced.
6. A final conclusion was stated based on the analysis, inferencing, and processing of the whole set of the obtained data.

5 Experimental Section

5.1 Logistic Regression Method

The above method, contrary to its name, is used as an alternate for linear regressor models, but when the model requires an output in the form of a classification, i.e., Yes/No, High/Moderate, etc., it is used most of the times in order to classify two categories. If more than two categorizations need to happen, it is termed as multinomial regression. It uses the logit function in order to assign values and predict the values strictly from 0 to 1, unlike linear regression which can sometimes exceed the threshold limit. This model does not require a relation between the input and output variables. Using different scikit models and packages, the accuracy for the prediction of the chances of a new person get addicted to the gadgets and distract themselves was found to be 94.68%.

5.2 K-Nearest Neighbors' Method

The K-Nearest Neighbors or KNN algorithm is a supervised learning classifier, which is the concept of distance or proximity from a cluster of some given set of points. It can be used both as a classification and regression algorithms. It is a lazy learning method, in which the model undergoes computation when the classification and prediction are being made as opposed to predicting a value in advance. In this method, the number of clusters to be made, i.e., “ k ” is defined. Then, with the help of the algorithm, the new data point is joined to any of the clusters, but the number of clusters are made equal to “ k ” that is predefined in nature.

The accuracy of the prediction obtained using this model is 93.08%.

5.3 Naïve Bayes Classifier Method

This classifier is strongly based on the “Bayes Theorem” which is used in calculating probabilities. It assumes independence among the different available predictors (or) no feature is related to the presence of any other feature. It is used particularly for larger datasets. It predicts on the basis of a probability of an event. It is easy to build and faster in computation compared to other machine learning algorithms.

The accuracy obtained using this algorithm for the above problem statement is 87.23%.

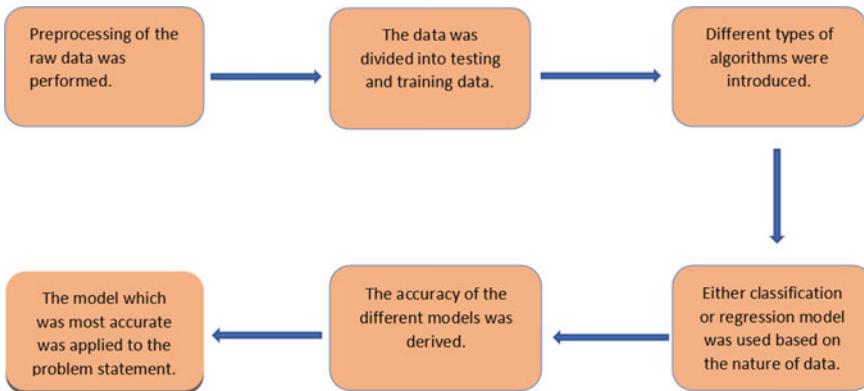


Fig. 1 Flowchart representing the steps followed in finding the accuracy of the machine learning models for predicting addiction probability in a new individual

The following flowchart gives an insight on the process carried out in order to get different inferences from the public for the survey based on the mentioned subject using machine learning. It also shows that the Logistic Regression Model is the most accurate of the above used machine learning models (Fig. 1).

The flowchart representation of the machine learning process is provided so as to provide a peek into the different algorithms which were needed to frame the contents present in the article, for the reader to understand the different machine learning processes easily without any type of confusions or dilemma regarding the main methodological process.

The steps are as mentioned below:

1. The data were preprocessed in order to remove null values or extreme outliers in the dataset or to normalize it.
2. The dataset was divided into training and testing datasets in a fixed ratio in order to train the model for future predictions.
3. Different set of algorithms were used in order to carry out the machine learning process, for example, Naïve Bayes Algorithm.
4. Either classification or regression model was implemented upon it.
5. The different models were tested for accuracy and the final accuracy was calculated.
6. The model which provided the highest accuracy was selected for making future prediction based on the newly provided data attribute values.

The following methods were able to provide and infer to the team, an accurate view of the state of things related to the given statement, and obtain an analytical and technical approach on the latter problem.

6 Results

At present day, nobody can reject the proposition of social media like social political, personal, and official uses. The youth aged between 18 and 21 years is the reason for this survey, for the youth in this level; it is a chance for them to concentrate on their future. Perhaps to have better future, it is necessary to reduce the usage of gadgets in daily life unfortunately, a modern life style is pry into gadgets. It is evident that a person is improper with gadgets around them and one cannot imagine life without gadgets and the applications and games present inside it.

Therefore, a survey was conducted to assess the different dimensions of distraction and the factors promoting them. The questions and the results are stated accordingly, in a clear form in the upcoming sections (Fig. 2).

The below graph Fig. 3 indicates the seriousness/gravity of the addiction to different electronic gadgets which are labeled as mild, moderate, and severe, respectively. The responses were then converted into a bar graph which is shown below.

Several machine learning algorithms were used to obtain an accuracy score for predicting the chances of an individual getting addicted to gadgets in the future, and each of the algorithm was compared with one another. The accuracy of each of the algorithm is depicted in the form of a graph as shown in Fig. 4.

The accuracies of each of the method are as follows:

1. Logistic Regression: 94.68%
2. K-Nearest Neighbor: 93.08%
3. Naïve Bayes Classifier: 87.23%.

We can also infer from the graph the fact that the Logistic Regression method is the most efficient and accurate method for predicting future outcomes.

Fig. 2 Graph representing the count on the gravity of the addiction to electronic gadgets on the masses

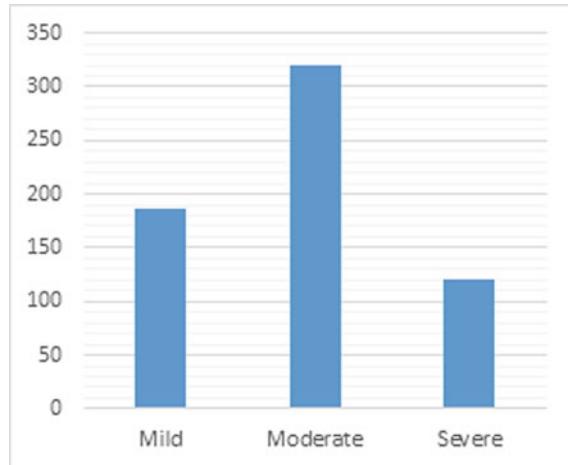


Fig. 3 Chart representing the accuracy obtained through different ML models

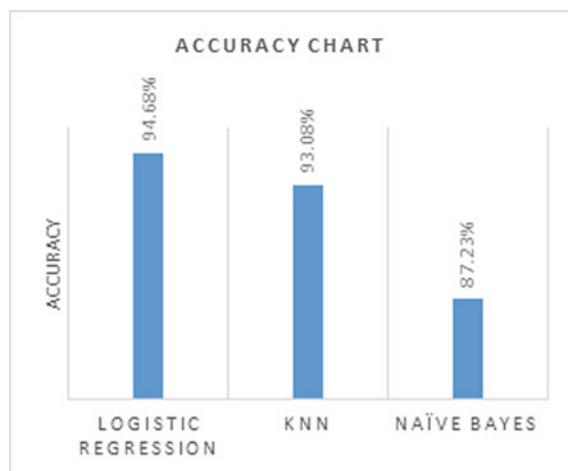
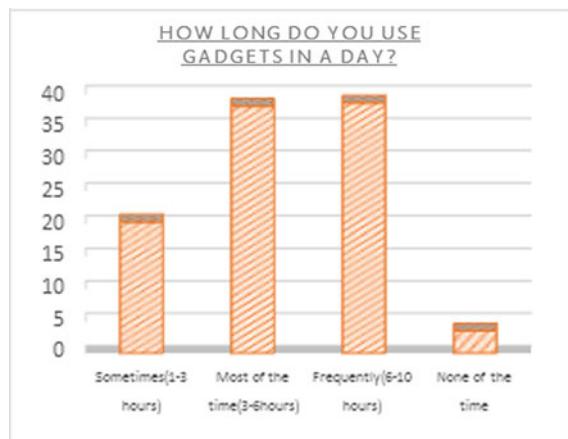


Fig. 4 Graph showing usage of gadgets by youth in a given day

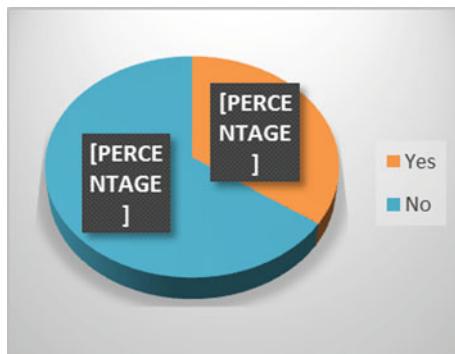


Some relevant questions were then poised to the responders which were used to obtain a clear idea about the pattern in the severity of the addictions and provide a conclusion to it.

6.1 “How Long Do Use Gadgets in a Day?”

From the above Fig. 5, we can infer on the amount of time each person dedicates to using their social media sites or any other forms of entertainment using gadgets. As mentioned in the above graph, 38 percentage of the users who have attended the survey have claimed that they use their devices for most of the time, i.e., they are

Fig. 5 Pie chart showing the amount of time youth can be away from gadgets in a day



on their gadgets for at least 3–6 h. This speaks out the fact that they spend one-third to one-fourth of their day on their gadgets surfing social media sites or other entertainment websites. This means that there is a space for scope in improvement by detaching themselves from the electronic media by cultivating new hobbies, practicing a new art, etc. About 38.5 percentage of the users have admitted to using gadgets for 6–10 h frequently. This means that they spend a major chunk of their day by doing mostly unproductive work on their respective gadgets. This infers that they should find a way to distract themselves from the excessive usage of the gadgets in the form of hobbies or any sport or any random solution. About 20.1 percentage of the responses state that they use the devices for only 1–3 h, i.e., one-eighth of the day and keep themselves preoccupied for the day. This means that they are not overly dependent on the devices surrounded around them. Thus, from the above figure, we can prove that a major part of the people who responded are dependent on gadgets in one or the other way. From this, we can conclude the analysis of the information of how long do people use gadgets in a day.

6.2 “Do You Go Out Regularly?”

From the above Fig. 7, we can infer the amount of time an individual dedicates to using their time for going out frequently to do unwanted things like traveling, wasting time in any unproductive work, etc., as mentioned in the above Fig. 3; 56.6% of the youth who have attended the survey tells us that the youth are going out frequently because of this reason, the youth cannot focus on their respective future and they are not preparing for the upcoming challenges in their career. Technically, the teenagers and adolescents would not be able to take important decisions on their own while solving different real life problem scenarios. They are planning about which place they need to go out and have lots of entertainment. These bunch of youth cannot attain success in the near future as they are not serious about their career. The 43.4% of youth have admitted that they do not go out frequently. This means that they can be focused on their future and they are not wasting time for any unwanted things to

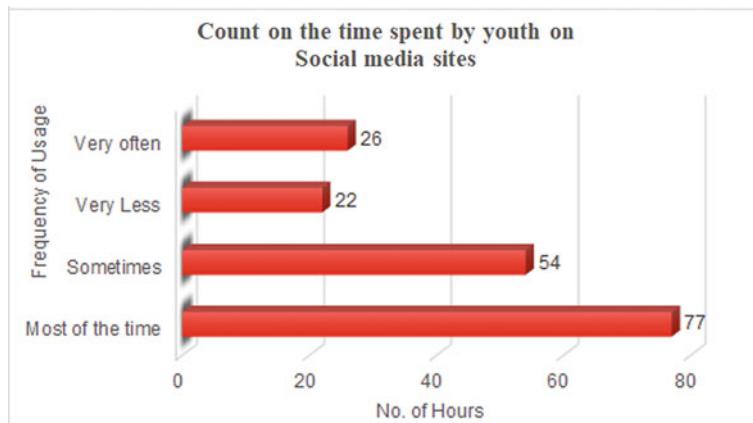


Fig. 6 Graph representing the amount of time youth spend viewing social media sites in a day

do. From the above figure, we can conclude the analysis of the data how youth go out frequently and the academic performance result obtained by posing the above question to them.

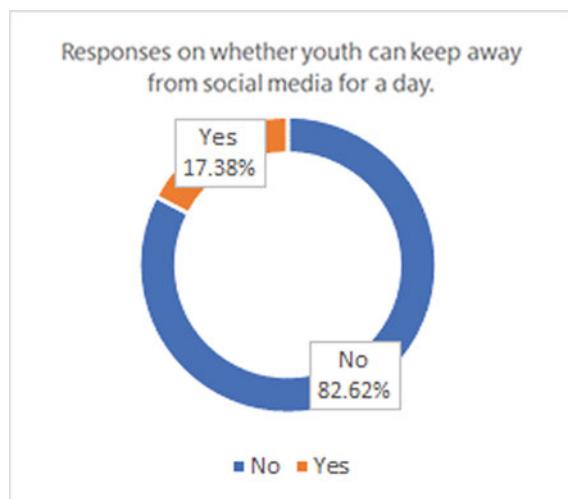
6.3 “How Frequently Do You Go Out?”

The survey question shown above was published, and a response count of 626 was received. It is observed that 33.5% of youth go out every day, 29.1% go out every week, and 37.4% go out only during certain limited occasions. With the above response, it is evident that a high percentage of students tend to move out every day. By practicing this habit regularly, it occupies their precious time dedicated to studies. This may impact the students by piling up of portions together at the end of the academic year which would cause hardship in preparing for the exams. Due to this, many students may not be able to prepare completely for the examination and may end up with backlogs. This is one of the major distractions which happens in youth unknowingly and affects their academic performances/abilities (Fig. 6).

6.4 “How Long Do You View Social Media or Other Sites in a Day?”

Social media is playing a major role in everyone’s life by sharing huge knowledge on all aspects. Even though social media such as Instagram, Facebook, Twitter, TikTok surfing has many advantages, it can also have a negative impact to those who gets addicted by overspending their personal time on these media.

Fig. 7 Pie chart representing the percentage of youth answering whether they can keep away from social media for a day



It can take away precious time from the youth if they do not use it essentially and it can cause huge damage not only with respect to time but also their mental health. The survey was done on this with the questioner “How long do you view social media/other sites in a day?” It was noted that about 43% of the students use social media most of the time in a day. This shows that the youth are more prone to get engaged in the social media and are getting distracted from their regular academics. So, this is one of the major mediums which is distracting the youth (Fig. 7).

7 Conclusion

This study has focused on various disturbances and distractions in younger generations mainly addiction toward social media, web surfing, electronic gadgets, etc. Some of these problems can even have negative impact on mental health causing behavioral changes and depression, affecting the learning ability, and can even make it impossible to reach their goals which can make them feel dejected about themselves.

There are many people who are badly affected by poor mental health conditions and depression due to their stressful lives and many have even lost their lives due to depression, which is a very grave problem of the hour.

As per the result of a survey conducted across India in 2021, it is found that around 43% Indians are suffering from depression and 14% of them are youths of age 15–24 [12]. In conclusion, we can state that everyone has ability to accomplish their goals. But, we must focus on our goals without getting distracted from external factors.

References

1. O'Brien O, Sumich A, Kanjo E, Kuss D (2022) WiFi at university: a better balance between education activity and distraction activity needed. *Comput Educ Open* 3:100071
2. Christakis DA, Moreno MM, Jelenchick L, Myaing MT, Zhou C (2011) Problematic internet usage in US college students: a pilot study. *BMC Med* 9(1):1–6
3. Elhai JD, Yang H, Montag C (2020) Fear of missing out (FOMO): overview, theoretical underpinnings, and literature review on relations with severity of negative affectivity and problematic technology use. *Brazilian J Psychiatry* 43:203–209
4. Gowthami S, Kumar SVK (2016) Impact of smartphone: a pilot study on positive and negative effects. *Int J Sci Eng Appl Sci (IJSEAS)* 2(3):473–478
5. Ishii K, Aoyagi K, Shibata A, Javad Koohsari M, Carver A, Oka K (2020) Joint associations of leisure screen time and physical activity with academic performance in a sample of Japanese children. *Int J Environ Res Public Health* 17(3):757
6. Throuvala MA, Pontes HM, Tsaousis I, Griffiths MD, Rennoldson M, Kuss DJ (2021) Exploring the dimensions of smartphone distraction: development, validation, measurement invariance, and latent mean differences of the smartphone distraction scale (SDS). *Front Psych* 12:199
7. Mayildurai R, Logeshkumar S, Priyanka A, Mythili AS (2019) Destructive effects of distraction on younger generation. *Int J Eng Adv Technol (IJEAT)* 8(65):76–78
8. Bhardwaj A (2016) Importance of education in human life: a holistic approach. *Int J Sci Conscious* 2(2):23–28
9. Gikas J, Grant MM (2013) Mobile computing devices in higher education: student perspectives on learning with cellphones, smartphones & social media
10. Van Den Beemt A, Thurlings M, Willems M (2020) Towards an understanding of social media use in the classroom: a literature review. *Technol Pedagog Educ* 29(1):35–55
11. Kuppuswamy S, Narayan PS (2010) The impact of social networking websites on the education of youth. *Int J Virtual Commun Soc Network (IJVCSN)* 2(1):67–79
12. Tayo SS, Adebola ST, Yahya DO (2019) Social media: usage and influence on undergraduate studies in Nigerian universities. *Int J Educ Dev Using Inf Commun Technol* 15(3):53–62
13. Singh M (2017) A study on the impact of social media on education and academic performance of engineering students in Lucknow city. *Int J Res Econ Soc Sci (IJRESS)* 7(11):171–178
14. Siebers T, Beyens I, Pouwels JL, Valkenburg PM (2022) Social media and distraction: an experience sampling study among adolescents. *Media Psychol*, pp 1–24
15. Yeboah J, Ewur GD (2014) The impact of WhatsApp messenger usage on students performance in tertiary institutions in Ghana. *J Educ Pract* 5(6):157–164
16. Agarwal M, Bishesh B, Bansal S, Kumari D (2021) Role of social media on digital distraction: a study on university students. *J Content Commun Commun* 13:125–136
17. Al-Menayes JJ (2015) Social media use, engagement and addiction as predictors of academic performance. *Int J Psychol Stud* 7(4):86–94
18. Abi-Jaoude E, Naylor KT, Pignatiello A (2020) Smartphones, social media use and youth mental health. *CMAJ* 192(6):E136–E141
19. Bandaru AKR, Fiaidhi J, Gannamani VS (2020) FOMO-social media engagement-smartphone addiction and distraction _IEEPPaper.pdf
20. James C, Davis K, Charmaraman L, Konrath S, Slovak P, Weinstein E, Yarosh L (2017) Digital life and youth well-being, social connectedness, empathy, and narcissism. *Pediatrics* 140(Supplement_2), S71–S75
21. Sanjana S, Shriya VR, Vaishnavi G, Ashwini K (2021) A review on various methodologies used for vehicle classification, helmet detection and number plate recognition. *Evol Intel* 14(2):979–987
22. Kusuma T, Ashwini K (2021) Modular ST-MRF environment for moving target detection and tracking under adverse local conditions. In: International conference on big data analytics (pp 93–105). Springer, Cham
23. Kusuma T, Ashwini K (2018) Real time object tracking in H. 264/AVC using polar vector median and block coding modes. *Int J Comput Inf Eng* 12(11):981–985

24. Kodipalli A, Devi S (2021) Prediction of PCOS and mental health using fuzzy inference and SVM. *Frontiers in public health*
25. Kusuma T, Ashwini K (2022) Analysis of deep learning frameworks for object detection in motion. *Int J Knowl-Based Intell Eng Syst* 2022, ISSN: 1327-2314.<https://doi.org/10.3233/kes-220002>
26. Kusuma T, Ashwini K (2022) Multiple object tracking using STMRF and YOLOv4 deep SORT in surveillance video. *Int J Res Trends Innov* ISSN:2456-3315

The Intervention of Technology in Education Under Isolation: Intuitions from Covid



Stephen Owusu Afriyie, Joseph Akwasi Nkyi, Gertrude Amoakohene, Mohammed Musah, and Peter Yao Lartey

Abstract The novel Coronavirus Disease (COVID-19) is recognized as a global pandemic, affecting more than 530 million individuals worldwide. The COVID contagion presents an exclusive challenge to education. Due to the restrictions recommended by the World Health Organization (WHO), governments, health professionals, and other organizations, the educational sector moved from offline to online pedagogy. The purpose of this study is to assess the impact of technology in teaching and learning process, student engagement, and faculty involvement toward virtual classrooms as triggered by COVID. A cross-sectional study was conducted to gather information from some teachers and students of Ghana. An online questionnaire was developed to collect data on probable academic outcome envisaged through technology in the midst of educational restriction by the COVID pandemic. About 500 teachers and students were engaged to participate in the investigation. The empirical evidence provided in this study suggests that during the lockdown period, schools undergone technological adoption processes, and students are still involved with various online learning modes. COVID awareness among students was massive, but there existed anxiety, fear, stigma, and discomfort. Change in the education sector through COVID has propelled innovation diffusion. This study showed that the change in the education sector by engaging students with various virtual sessions has created intellectual harmony in the minds of students and faculty members toward technological adaptation for academic success.

Keywords Technology · Education · Restrictions · Virtual classrooms · COVID

S. O. Afriyie (✉) · J. A. Nkyi · G. Amoakohene · M. Musah
Ghana Communication Technology University, I.T. Business, PMB 100, Accra-North, Ghana
e-mail: safriyie@gctu.edu.gh

P. Y. Lartey
School of Accounting, Universidade Federal de Uberlândia, Uberlândia, MG, Brazil

1 Introduction

The outbreak of the COVID-19 pandemic has caused a lot of transformation in every aspect of the society today, most particularly in education, social activities, economy and finance, tourism, labor, and employment. COVID-19 was declared a pandemic by the World Health Organization (WHO) on March 12, 2020 [1]. Various governments across the world have implemented policy initiatives on a broad base in response to the impact of the 2019 global pandemic on education. The governments took those decisions by restricting education based on the original data available to them and how the virus was spreading in their countries to fight the deadly coronavirus [2]. Some of such directives include emergency response centers, suspension or restriction of interpersonal academic activities and indefinite closure of schools. The impact of COVID also encouraged the introduction of technological innovation such as launching online teaching platforms for students to learn from homes in order to curb the outbreak of the deadly disease. This study focuses on the effect of technology on the educational sector during the academic restriction actuated by COVID. The study also accentuates on the transmission from face-to-face interaction or classroom interaction between students and teachers to online teaching through the use of technology. Dorgu [3] believes the attitude of teachers and students should be highly considered with the aim of understanding how they discharge their respective duties while adapting to new teaching modalities which are consistent with global preventive and safety measures. Policymakers have estimated severe difficulties and new levels of stress on teaching and learning in educational institutions. According to a recent survey by Nandkar [4], teachers and students suffered emotionally and constant stress, anxiety, and extreme fatigue associated with nationwide lockdown and closure of schools. Apart from these issues, students' academic performance equally suffered a setback due to excessive bureaucratic standards, inflexible guidelines, vague instructions, non-availability of Internet facilities, and lack of technological devices to support online teaching and learning. In the study conducted by Durak and Çankaya [5], the COVID pandemic has impacted negatively on education due to the lack of basic infrastructure, teaching materials, lack of teacher training, poor communication channels, and people living in deplorable home conditions. Poor teaching and learning could also be associated with lack of safety and protective materials, since people fear of their lives been at risk when they come out for school. The human and social insecurity due to the pandemic generated a strident revolution in the teaching fraternity toward adapting to technology and virtual engagement of students.

2 Literature Review

2.1 *Educational Investment*

The impact of COVID on the economic and financial capability of educational institutions, students, and teachers cannot be undermined. The adverse effects have already been widely estimated which include an upward adjustment of tuition fees and academic user facilities as a result of the educational institutions' inability to cope with the high financial and economic uncertainties such as cost and losses. Smith [6] elucidates that as the cost of sustaining academic activities continues to grow due to tragic public health systems to deal with the spread of the COVID pandemic, educational institutions are facing a reduction in enrollment and disruption of academic programs without measure. Cost of education is one of the factors threatening the sustainability of education in the world, especially in developing countries [7]. From the contribution of Buheji et al. [8], the financial challenges imposed by the pandemic have accounted for high unemployment rate resulting from loss of jobs which has already affected financing education. Currently, cost of education with the use of Internet facilities has incontrovertibly added to the financial burden of people. Pokhrel and Chhetri [9] epitomize that tutors and students must rely on costly and stable Internet connectivity for teaching and learning. This conforms to a previous study in India by Mishra et al. [10] which reported that COVID has demanded a mandatory, but expensive stable Internet for pedagogy. Additionally, Government's regular financial budget meant to support education will defiantly fall as the impact of the pandemic spread across the economy has been pathetic. As highlighted by Sanchez-Moyano [11], government projections to restore the micro and macroeconomic conditions will further stiffen the budget allocation for the educational sector. As a result, there will still be insufficient funding or limited investment in education such as scholarship packages to support needy students. This implies that students who could rely on bursary, grant, or studentship may not be guaranteed their sources of funding during the COVID pandemic [12]. Education is a critical societal investment and has long been championed for its vast array of personal and social benefits, and for its empowerment to communities. Even though various governments have offered some amount of support during the pandemic to the populace, parents, students, and teachers have spent a lot of money on teaching gadgets and Internet.

2.2 *Technological Innovation*

Innovation is key to the success of every economic, social, political, health, technical, and educational activities. In the absence of innovation, educational institutions encounter stagnation, especially during the COVID contagion. Mathivanan et al. [13] hinted that countries are facing challenges in sustaining educational activities amid

the pandemic, hence the desire to innovate through technology is on the agenda of every government than ever. Kaplan et al. [14] believe policymakers have consistently emphasized on the need for broad technical investment in education to help the sector survive the deadly impact of the COVID pandemic. The current situation requires that plans regarding educational policies are changed rapidly through creativity and ingenuity in order to realize the opportunities associated with innovation. Therefore, technology adoption is vital to maintain the education of residents during the COVID wave. The global pandemic also presents the best opportunity to assess the use of “creative destruction” that is suitable for the educational sector [15]. From the point of view of Jha et al. [16], for education to survive the impact of the pandemic, stakeholders need to implement human creativity and knowledge management, provide infrastructure, and nurture talent that will support virtual teaching and learning. Educational institutions must be poised to create their own innovative systems to support online teaching and learning. However, students and teachers must develop a habit or make use of these technologically innovative capabilities to bridge the gap created by restrictions; closure of schools or suspension of face-to-face classes due to COVID [17]. The innovative provision ensures continuous teaching and learning without a hitch. Heleta and Bagus [18] recounted that the basic foundation for innovation in education should include ensuring equality in social and economic achievements for all students. Additionally, changes in family social life and work schedules of parents and guardians could support compliance with safety preventive protocols. Rapid technology adoption in teaching, learning, and assessment of student performance also contributes passively to quality education [19]. Replacing the traditional method of teaching with effective virtual tools and platforms may be helpful, as teachers deliver lessons remotely, expediently, and communicate knowledge to students. While it is more convenient for students to develop their knowledge through practical experiments with teachers in a traditional classroom environment, knowledge is presented through online tools and platforms for only those having access to the technology infrastructure. Lee [20] alluded that innovation may also include designating and adopting new teaching and administrative practices that will help students fulfill their academic requirements. The sensitivities toward studies have changed, giving technology usage a priority over conventional lessons. Even though COVID has disrupted the academic calendar and day-to-day life on the school campus, Shenoy et al. [21] believe it has created a revolution in education, as the teaching fraternity has revised its physical training methods and accepted practices toward adaptation to technology and virtual engagement with students for pedagogical purposes.

3 Methodology

The method used for this exploratory study was a survey. A cross-sectional enquiry was conducted to gather data from some teachers and students of Ghana. An online questionnaire (close-ended) was developed to collect data on probable academic

outcomes envisaged through technology in the midst of educational restrictions by the COVID pandemic. A total of 500 teachers and students were engaged to participate in the investigation. As the investigators wanted fair inference, they selected 250 teachers and 250 students who were involved in virtual schooling during the pandemic. The participants were chosen according to their profound experience in the field of online teaching and learning. When participants are regarded as people with crystallized viewpoints for pragmatically exclusive information, they can provide valued information on emerging issues that could be difficult or impossible to receive [22]. One of those rising topics is the essential digitization of teaching and learning amid the COVID pandemic. However, five point-scale measurements, from strongly disagree = 1 to strongly agree = 5 were used to measure the respondents' answers. Respondents were asked to select from these scales related to statements involving the impact of technology on online teaching and learning process as triggered by COVID. After explaining the goal of the research to the contributors, the questionnaires were administered by E-mail, Zoom, Telegram, and Whatsapp. Out of the 500 respondents, 478 appropriately filled the questionnaires and submitted, representing a successful recovery rate of 95.60%. Preliminary tests were also conducted, including reliability using Cronbach's Alpha to check internal consistency, validity, and multicollinearity tests to determine if the data collected were appropriate for the study and could therefore suit the predictable model. Once they were considered reliable and valid, the data gathered were analyzed using PLS-SEM. An inference was then drawn in the study based on the results received (from the variables in Table 1).

4 Findings and Discussions

From Fig. 1, the model consists of two exogenous variables (Learning Design and Student Assessment) and three endogenous variables (COVID effect, technological use, and academic performance). The structure categorized as a reflective model has five constructs and their indicators. In the study, the R^2 values of the endogenous variables such as COVID effect, technological use, and academic performance are 0.591, 0.949, and 0.819, respectively. These R^2 values are used to evaluate the structural model, since they measure the model's prediction accuracy. The R^2 values represent the combined impacts of the exogenous variables on the endogenous variables. It encapsulates the degree of variance in the endogenous constructs explained by the exogenous construct (s) linked to them. The structural model is assessed based on the R^2 ; the larger the R^2 , the higher the proportion of variation explained [23], and the better the model fit of the observations [24]. The R^2 value of COVID effect for this study is 0.591. This indicates that the amount of variance, approximately 60% in the endogenous constructs is explained by the entire exogenous constructs connected to it. Nevertheless, the amount of variance in technological use and academic performance explained by all of the exogenous constructs linked to them astoundingly happened to be 0.949 (94.9%) and (0.819) 81.9%, respectively.

Table 1 Variable labels

Latent/Implicit variables	Observable/Explicit variables	Labels
Learning design	Course content	LEN1
	Objectives and outcome	LEN2
	Study materials	LEN3
Student assessment	Report writing	ASS1
	Quizzes	ASS2
	Term examinations	ASS3
COVID effect	Fear	COV1
	Stigma	COV2
	Lockdown	COV3
Technological use	Zoom	TEC1
	Google meet	TEC2
	Demio	TEC3
	WebinarNinja	TEC4
Academic performance	Satisfactory	ACA1
	Unsatisfactory	ACA2
	Needs improvement	ACA2

This shows that the structural model developed in this investigation has realistically predictive significance and remarkable precision level, due to accurate estimation power of the exploratory variables on the model. Statistically, the greater the percentage of R^2 for accurate predictions, the better the model fit for the study. However, the use of technology in online learning during the pandemic is essential to uphold academic performance. Figure 1 also depicts a structural model assessment to evaluate the relationship among the constructs. The path coefficients generated by the model were incredibly high and positive values. Technological use has a positive relationship with Academic performance. The figures obtained as a result of the aforesaid implicit variables' relationship is 0.905. The study shows that technology has the proclivity to improve academic performance. In view of this output, technology plays a noteworthy role to achieve academic success even during the COVID period. Tam [25] posits that students were satisfied with virtual classes as online platforms enhanced their accomplishments. From the model, cross-loadings between the latent and observable variables displayed high scores. Cross-loadings allow the assessment of correlation among factors. From Fornell and Larcker [26], the measurement of cross-loadings should be more than 0.7. Almost all cross-loadings in the study were above the 0.7 criterion. For instance, the cross-loadings between Learning design and LEN1, LEN2, and LEN3 were 0.934, 0.943, and 0.905, respectively. Also, the loadings between Student assessment and its indicators measured between 0.713 and 0.972. Congruently, the cross-loadings between COVID effect and

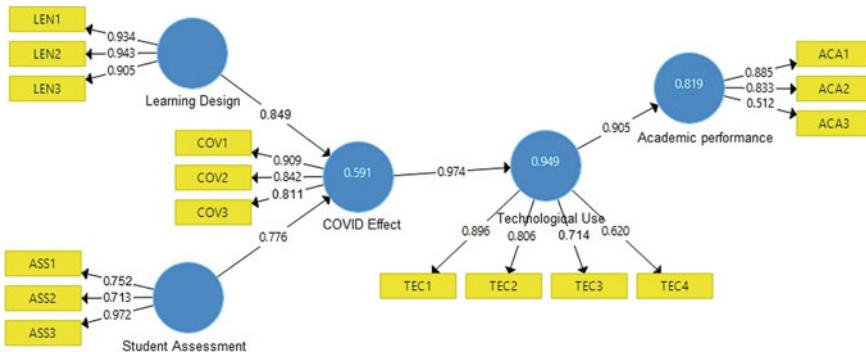


Fig. 1 Structural model displaying path analysis of variables

COV1, COV2, and COV3 were 0.909, 0.842, and 0.811, respectively. This stipulates that the constructs and their indicators have a strong relationship.

5 Conclusion

In this rapidly evolving coronavirus contagion, an acceptable form of educational flexibility will be essential for teachers, students, and learning methods. Both educators and learners are required to avail themselves for the transition by offering alternative resources that permit easy access to studies to circumvent uncertain bottlenecks [27]. To ensure effective implementation of distance learning tools, they must encourage educationalists to deliver effectively, such that learners would have access to high-quality education during these trying times. As the world battles with COVID, education is experiencing unexpected lockdown, but virtual lessons can still continue by adopting the most efficient methods of education [28]. These methods may be used in the future (maybe with an upgrade) to drive innovations in the education sector. This task cannot be executed by only tutors and students, but faculty leadership should also show keen interest to maintain high standards of education. As a result of the situation, most educational institutions have accepted and implemented technology-based or virtual classes, and their involvement has been phenomenal. In some cases, the online engagement of students records higher attendance even better than normal classrooms, since learners conveniently partake schooling from home. The student-centric factors like overall study engagement, attentiveness, and attitude toward online classes contribute in enhancing the quality of teaching and learning [29]. This unique circumstance will change the conservative way of educating students, but will require cooperation and novelty. However, schools, educators, and policy-makers should take these virtual classes as innovation (although coming through a revolution) for other educational institutions to implement the similar technique. This is because the innovation adopted by educational faculties during the COVID

crisis fortified teachers to continue to impart knowledge to students. This is not to say online education should or could even substitute the traditional approach. It is not really that at all. The authors of this study are just suggesting an advancement of the argument around virtual learning and the need for prudential readiness to swiftly adjust delivery methods, depending on the situation or circumstance (like the COVID pandemic). Teaching remotely requires essential capacity and appropriate infrastructural support to efficiently deliver in an online mode [30]. Efficient online teaching demands tutors' support for students, which encompasses regular monitoring of students' learning processes, as with the face-to-face schooling.

References

1. Rahimi F, Abadi ATB (2020) Tackling the COVID-19 pandemic. *Arch Med Res* 51(5):468–470
2. Kautish S (2022) Coronavirus-related disease pandemic: a review on machine learning approaches and treatment trials on diagnosed population for future clinical decision support. *Current Med Imaging* 18(2):104–112
3. Dorgu TE (2015) Different teaching methods: a panacea for effective curriculum implementation in the classroom. *Int J Secondary Educ* 3(6–1):77–87
4. Nandkar RS (2020) To study the effect of lockdown on physical, mental and emotional health of common people. *Int J Innov Sci Res Technol* 6:777–785
5. Durak G, Çankaya S (2020) Undergraduate students' views about emergency distance education during the COVID-19 pandemic. *Online Submission* 5:122–147
6. Smith WC (2021) Consequences of school closure on access to education: lessons from the 2013–2016 Ebola pandemic. *Int Rev Educ* 67:53–78
7. Cobbina PB, Erdiaw-Kwasie MO, Amoateng P (2015) Rethinking sustainable development within the framework of poverty and urbanisation in developing countries. *Environ Dev* 13:18–32
8. Buheji M, da Costa Cunha K, Beka G, Mavric B, De Souza YL, da Costa Silva SS, Hanafi M, Yein TC (2020) The extent of covid-19 pandemic socio-economic impact on global poverty. a global integrative multidisciplinary review. *American J Econ* 10(4):213–224
9. Pokhrel S, Chhetri R (2021) A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher Educ Future* 8:133–141
10. Mishra L, Gupta T, Shree A (2020) Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *Int J Educ Res Open*. 100012
11. Sanchez-Moyano R (2022) Lessons learned from small business lending during COVID-19: a case study of the California rebuilding fund. *Commun Dev Res Brief* 3:1–38
12. Oleksiyenko A, Blanco G, Hayhoe R, Jackson L, Lee J, Metcalfe A, Sivasubramaniam M, Zha Q (2021) Comparative and international higher education in a new key? Thoughts on the post-pandemic prospects of scholarship. *Compare J Comp Int Educ* 51(4):612–628
13. Mathivanan SK, Jayagopal P, Ahmed S, Manivannan SS, Kumar PJ, Raja KT, Dharinya SS, Prasad RG (2021) Adoption of e-learning during lockdown in India. *Int J Syst Assurance Eng Manage*, pp 1–10
14. Kaplan S, Lefler J, Zilberman D (2022) The political economy of COVID-19. *Appl Econ Perspect Policy* 44(1):477–488
15. Strielkowski W, Wang J (2020) An introduction: COVID-19 pandemic and academic leadership. In: 6th international conference on social, economic, and academic leadership (ICSEAL-6–2019) Atlantis Press, pp 1–4
16. Jha A, Jha N, Bhatate KR, Patsariya S, Pahariya JS (2022) COVID-19: a creative destruction in the education sector of India. In: Cases on practical applications for remote, hybrid, and hyflex teaching, pp 26–55

17. Ali W (2020) Online and remote learning in higher education institutes: a necessity in light of COVID-19 pandemic. *High Educ Stud* 10(3):16–25
18. Heleta S, Bagus T (2021) Sustainable development goals and higher education: leaving many behind. *High Educ* 81(1):163–177
19. Almaiah MA, Hajjej F, Lutfi A, Al-Khasawneh A, Shehab R, Al-Otaibi S, Alrawad M (2022) Explaining the factors affecting students' attitudes to using online learning (Madrasati platform) during COVID-19. *Electronics* 11(7):973
20. Lee K (2021) Openness and innovation in online higher education: a historical review of the two discourses. *Open Learn J Open Dist e-Learn* 36(2):112–132
21. Shenoy V, Uchil R, Alexander J, Mahendher S (2021) COVID 19-a metamorphosis in indian higher education institutions with technology infused learning. *Psychol Educ* 58:3208–3217
22. Innes JE, Booher DE (2015) A turning point for planning theory? Overcoming dividing discourses. *Plan Theory* 14(2):195–213
23. Rights JD, Sterba SK (2019) Quantifying explained variance in multilevel models: an integrative framework for defining R-squared measures. *Psychol Methods* 24(3):309
24. Hair JF Jr, Howard MC, Nitzl C (2020) Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *J Bus Res* 109:101–110
25. Tam ACF (2022) Students' perceptions of and learning practices in online timed take-home examinations during Covid-19. *Assess Eval High Educ* 47(3):477–492
26. Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. *J Mark Res* 18:39–50
27. Lei M, Medwell J (2021) Impact of the COVID-19 pandemic on student teachers: how the shift to online collaborative learning affects student teachers' learning and future teaching in a Chinese context. *Asia Pac Educ Rev* 22(2):169–179
28. Jena PK (2020) Online learning during lockdown period for covid-19 in India. *Int J Multi Educ Res (IJMER)* 9
29. Tripathi VM, Mani AP (2022) Online teaching during COVID-19: empirical evidence during indian lockdown. In: soft computing: theories and applications, pp. 251–262. Springer, Singapore
30. Amin FM, Zulfitri Z (2022) Emergency remote teaching during COVID-19 crisis: an analysis of EFL students' engagement in Aceh. *Englisia J Lang Educ Human* 9(2):46–59

Analysis of Bao-Zhou-Chen-Liu's Hybrid Chaotic System



Meenakshi Agarwal, Arvind, and Ram Ratan

Abstract One of the key encryption techniques commonly used nowadays in a variety of multimedia security applications is the hybrid chaos-based approach. The effectiveness of a hybrid chaotic system proposed by Bao-Zhou-Chen-Liu for multimedia encryption is examined in the paper. The effectiveness is evaluated by analysing the profile, trajectory, sensitivity, Lyapunov exponent, bifurcation, key space, entropy, zero–one test, and performing the randomness tests. Analysis reported in the paper reveals that Bao-Zhou-Chen-Liu's hybrid chaotic system (BZCLS) has a number of flaws including low sensitivity, small chaotic range, weaker ergodicity, smaller key space, failure of zero–one test, and several NIST randomness tests. Because of such reported drawbacks and shortcomings, BZCLS is not sufficiently secure and is unsafe to employ for information security applications.

Keywords Chaotic map · Dynamical system · Hybrid system · Cryptography · Information security

1 Introduction

Information security is an issue well important to be taken in design and development of information management systems for secure storage and exchange of vital information. In today's digital world and present era of technology, digital images are commonly used by modern societies in addition to other form of information. Security of such information can be ensured by considering suitable approaches like

M. Agarwal
Department of Mathematics, University of Delhi, Delhi, India
e-mail: ameenakshi68.ma@gmail.com

Arvind
Department of Mathematics, Hansraj College, University of Delhi, Delhi, India
e-mail: drarvind@hrc.du.ac.in

R. Ratan (✉)
Defence Research and Development Organization, Delhi, India
e-mail: ramratan_sag@hotmail.com

spread spectrum, steganography, and cryptography. The encryption process converts a normal data into encrypted data that seems garbled like a random mess. Wide-ranging uses of image encryption include secure handling of visual data in strategic communication, telemedicine, medical imaging, and multimedia systems. Chaotic encryption emerges as a potential encryption solution to address image encryption to meet encryption speed and memory constraints. Ergodicity, randomness, and initial value sensitivity are the characteristics of chaos, making chaos theory an appropriate choice for applications in cryptography. Chaotic maps are used commonly due to their structural simplicity, and we can say that they are the foundation of chaotic cryptography [1].

In chaos theory, the construction and analysis both are equally important in designing of chaotic systems. Hybrid chaotic systems, which are combinations of chaotic maps, offer high chaoticity, nonlinearity, and unpredictability in comparison of one-dimensional (1D) maps. This motivates us to deeply study the properties of the hybrid chaotic systems and the security strength they provide over existing basic maps. In the past two decades, secure communication has received a lot of interest by the realization of chaotic synchronization [1]. Chaotic encryption is being used intensively in various engineering applications. Initially, Mathews proposed the idea of securing vital data using chaotic maps [2]. For their simple structure and effective computing, the 1D chaotic maps: Logistic map [3], Tent map [4], Sine map [5], and Gauss map [6] have been in use previously to produce unpredictable random output sequences. A general method is presented to construct cryptosystems with discrete dynamical chaos using a fast iterative computation [7]. An algorithm is presented to create numerous keys for symmetric key cryptography using a simple chaotic function [8]. Chaotic systems employing two or more 1D chaotic maps are applied [9–12] to offset the dynamics degradation of chaotic map in digital computer. Many other encryption methods were also published for 2D and higher-dimensional chaotic maps. A 1D sine powered chaotic system (1DSP) [13] reported for high sensitivity and randomness. The difference between sequences of 1D chaotic maps by the Logistic, Sine, and Chebyshev maps [1] is used to get larger chaotic ranges and high chaoticity. The improvised versions of Logistic and Sine map are also reported using the sequences generated by 1D chaotic maps [14, 15]. The 1D Joan-S-MuraliP's (JSMP) map is reported by hybridizing Logistic and quadratic chaotic maps for larger complex chaotic band [16]. The 3D cat map, 2D logistic map, and hyper-chaotic maps such as the 2D-SLMM, 2D-LASM, 2D-SIMM, 2D-LICM, and 2D-SLMM are reported [1]. However, it is seen that some of the chaos-based encryption systems are either found to be computationally complicated or easily breakable [17, 18].

In the paper, we analyse a chaotic system given by Bao-Zhou-Chen-Liu, BZCLS, with claimed high security uses combination of 1D chaotic maps [10]. It is based on the criteria of Geffe generator [19]. It uses Logistic, Sine, and Tent map to design hybrid system and claimed that the chaotic system is effective and has more chaotic and complicated behaviour. For cryptographic security point of view, the chaotic

map should possesses high chaoticity and passing of randomness tests. We observe many weaknesses/shortcomings in this chaotic system which may cause threats to the security. We perform detailed analysis of BZCLS in terms of chaotic profile, trajectory, sensitivity, Lyapunov exponent, bifurcation, entropy, key space, 0–1 test, and NIST randomness tests.

The rest of this paper is organized as follows. Section 2 discusses BZCLS briefly. Section 3 analyses the chaoticity and shortcomings of BZCLS. Section 4 reports the results of NIST randomness tests, and Sect. 5 concludes the paper.

2 Brief Description of BZCLS

Three 1D chaotic maps, Logistic, Sine, and Tent maps are given first that are the essential components of BZCLS. Logistic map is a discrete-time analogue of an idealized population model [13]. Logistic map is expressed by (1).

$$x_{i+1} = rx_i(1 - x_i) \quad (1)$$

where parameter r is within the range of $[0, 4]$. Tent map is a 1D discrete map that can stretch and fold. The output is stretched to the range $[0, 1]$ when the input value is less than 0.5. If the input value is higher or equal to 0.5, the input value of a tent map is folded into the range $[0, 0.5]$ before being stretched into the range $[0, 1]$. It is represented by (2).

$$x_{n+1} = f_u(x_n) = \begin{cases} ux_n & \text{for } x_n < 1/2 \\ u(1 - x_n) & \text{for } x_n \geq 1/2 \end{cases} \quad (2)$$

where $f_u = u \min\{x, 1 - x\}$. The u is the system parameter lying in the range $(0, 2]$. The range of the sine map's inputs is $[0, \pi]$ and the range of its outputs is $[0, 1]$. It is expressed by (3)

$$x_{i+1} = a \sin(\pi x_i) \quad (3)$$

where parameter $a \in [0, 1]$.

BZCLS consists of a controlling sequence and two chaotic output sequence generators. The control sequence generator uses the Logistic map which chooses either Tent map or Sine map randomly for generating chaotic sequences. It is expressed by equations (4–8).

$$f_1(Y_n) = L(Y_n) = rY_n(1 - Y_n) \quad (4)$$

$$q_n = \begin{cases} 1 & f_1(Y_n) < 0.5 \\ 0 & f_1(Y_n) \geq 0.5 \end{cases} \quad (5)$$

$$X_{n+1} = F(X_n, q_n) = \begin{cases} f_2(X_n) & q_n = 0 \\ f_3(X_n) & q_n = 1 \end{cases} \quad (6)$$

$$f_2(X_n) = T(X_n) = \begin{cases} uX_n & X_n < 0.5 \\ u(1 - X_n) & X_n \geq 0.5 \end{cases} \quad (7)$$

$$f_3(X_n) = S(X_n) = a \sin(\pi X_n) \quad (8)$$

3 Chaoticity of BZCLS

The analysis of Lyapunov exponent for Logistic, Sine, and Tent maps shows that their control parameters must lie in the ranges from 3 to 4, 0 to 1, and 1 to 2, respectively, for chaotic behaviour of output sequences [4]. It is also seen that these maps show strong chaotic behaviour when the control parameters are chosen from the higher side values and weak chaotic behaviour when the parameters are chosen from the lower side values. To observe the security weaknesses of this system, we check the chaotic characteristics by making variation in the control parameters within their chaotic ranges. Here, we discuss the weaknesses of BZCLS with respect to different chaotic characteristics measures and present the results obtained by taking control parameter values in upper and lower bound ranges of their chaotic interval. The analysis is carried out by implementing experiments on MATLAB platform.

3.1 Weak Profile

Different profile distributions for data samples are obtained by making variations in control parameters. The profiles of output sequences are shown in Fig. 1 where Fig. 1a, b show the profiles for higher side values $P1 (r = 3.95, u = 1.95, a = 0.95)$ and lower side values $P2 (r = 3.1, u = 1.1, a = 0.88)$ set of parameters, respectively. We observe that none of the profiles are covering the available range $[0, 1]$ since a strong chaotic system's profile should be evenly dispersed in the range $[0, 1]$.

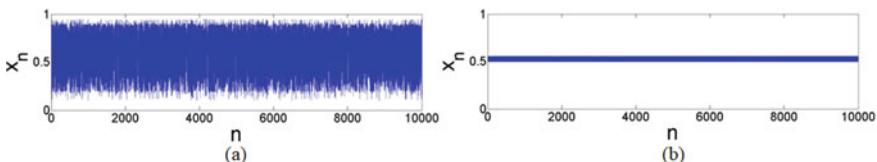


Fig. 1 Profiles of output sequences, **a** for $P1$ and **b** for $P2$ set of parameters

3.2 Non-coverage Trajectory

The 2D and 3D trajectories can be seen in Fig. 2 where Fig. 2a–d show the trajectory for the set of parameters $P1$ and $P2$. Figure 2a, c and b, d show 2D and 3D trajectories, respectively. Figure 2 clearly show that the trajectory coverage is very minimal when the values of parameters are chosen from the lower side. It occupies lesser region of the phase plane even on the higher side values of parameters. This indicates weaker ergodicity for all choices of parameters.

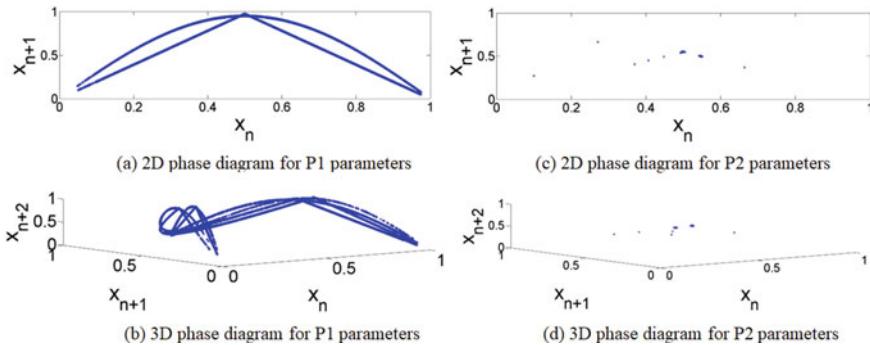


Fig. 2 2D trajectories: **a** for $P1$ and **b** for $P2$ set of parameters, 3D trajectories: **c** for $P1$ and **d** for $P2$ set of parameters

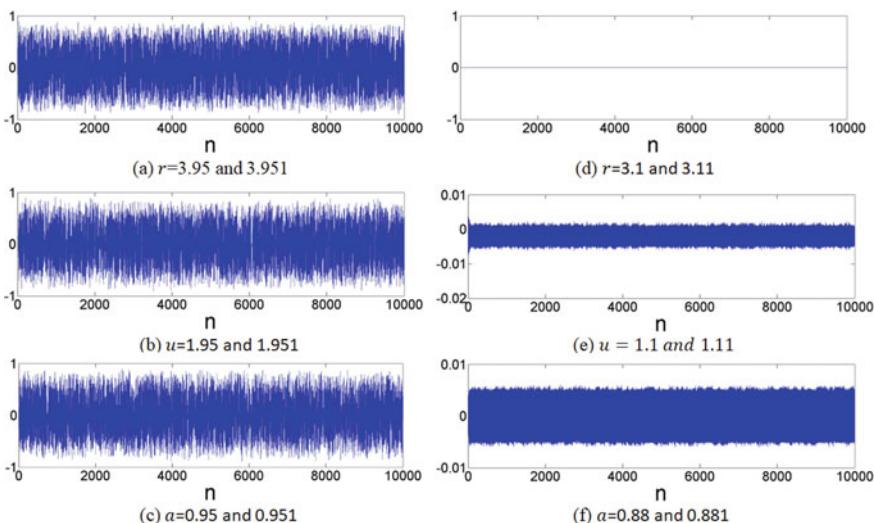


Fig. 3 Difference of sequences for variations in parameters: **a–c** for change in $P1$ and **d–f** for change in $P2$ set of parameters

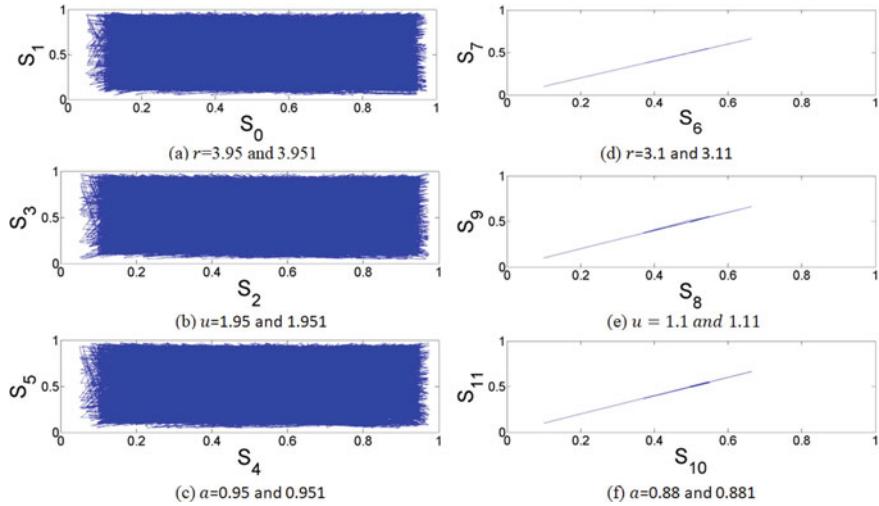


Fig. 4 Relationship between two chaotic sequences by changing one parameter: **a–c** for $P1$ and **d–f** for $P2$ set of parameters

3.3 Low Sensitivity

The difference profiles and random relationship graphs of two sequences are considered to present the sensitivity. These are plotted by making variations in the least significant digits of the values of set of control parameters $P1$ and $P2$. The difference profiles are shown in Fig. 3. Figure 3a–c and d–f represent results for $P1$ and $P2$ set of parameters, respectively.

The random relationship graphs are given in Fig. 4 in a likewise manner. It is seen in Fig. 3a–c and Fig. 4a–c that the difference profiles and random relationship are not covering their complete ranges $[0, 1]$ and $[0, 1] \times [0, 1]$, respectively. It is also clear from Fig. 3d–f and Fig. 4d–f that the system becomes completely insensitive on choosing parameter values from lower side. Hence, we see that the system has weak sensitivity for most of the choices of parameters and diminishing its key space.

3.4 Low Lyapunov Exponent

The LE is observed for each parameter to check the system's chaotic range. Figure 5 represents the LE spectrum of the system with respect to each parameter. Figure 5 (a) and (b) show LE for the values of fixed parameters taken from the higher and lower side, respectively. From Fig. 5 (a), we observe that the system is chaotic for $r \in (2, 4)$, $u \in (0.5, 2)$, and $a \in (0.2, 1)$ and from Fig. 5 (b), we see that the system has

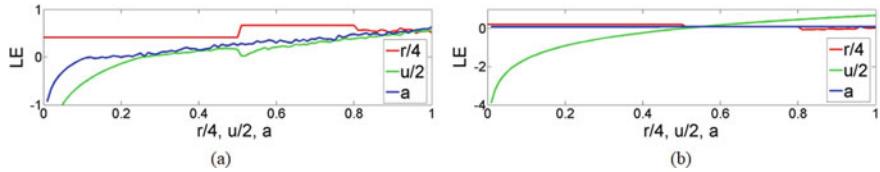


Fig. 5 LE for the BZCLS: **a–c** for $P1$ and **d–f** for $P2$ parameters

chaotic behaviour for $r \in (0, 3.2) \cup (3.9, 4)$, $u \in (1, 2)$, and $a \in (0.1, 1)$. Thus, the behaviour of LE is shown to be depend upon the choice of fixed parameter value.

In practical use, we cannot restrict the choices of fixed parameters, so a chaotic system has to have positive LE irrespective of the values of fixed parameter. Thus, $r \in (2, 4) \cap ((0, 3.2) \cup (3.9, 4)) = (2, 3.2) \cup (3.9, 4)$, $u \in (0.5, 2) \cap (1, 2) = (1, 2)$ and $a \in (0.2, 1) \cap (0.1, 1) = (0.2, 1)$ are the actual ranges of chaotic parameters. From Fig. 5 (a)-(b), we also realize that the value of LE is less positive which indicates low chaoticity and non-randomness.

3.5 Low Bifurcation

For a dynamical system, positive LE implies that the system behaves chaotic, but for good chaotic behaviour, dense bifurcation diagram is an additional property which has to be satisfied by the system. Figure 6 shows the bifurcation diagrams for BZCLS with respect to each control parameter.

Figure 6a–c and d–f show bifurcation diagrams for higher and lower side ranges of control parameters, respectively. From Fig. 6b, f, values of r ranging from $(2, 3.2)$ and $a \in (0.2, 1)$ are not considerable due to low bifurcation. Hence, only two parameters r and u with limited ranges are providing chaoticity to BZCLS. Consequently, the width of chaotic parameters given in Sect. 4.4 is reduced to $r \in (3.9, 4)$ and $u \in (1, 2)$.

3.6 Low Entropy

The entropy values for the output sequences with different parameters are given in Table 1 which are varying from 0.88 to 1. Table 1 indicates that the system does not have adequate entropy values, and hence, randomness on lower side values of control parameters as it is not giving desired entropy for all the values of control parameters.

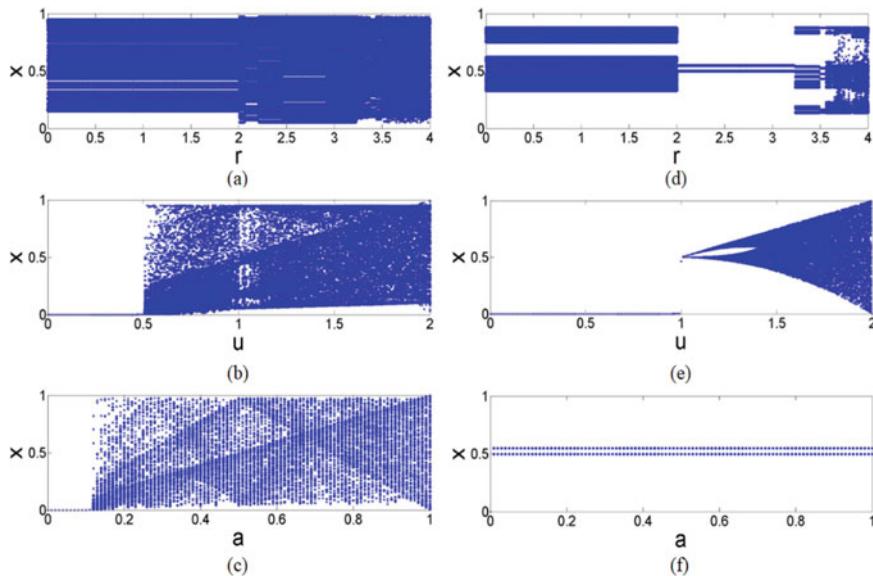


Fig. 6 Bifurcation diagram for BZCLS: **a–c** for $P1$ and **d–f** for $P2$ parameters

Table 1 Entropy values of the system for different parameter values

S. No.	Change in control parameters			Entropy
	Logistic map parameter (r)	Tent map parameter (u)	Sine map parameter (a)	
1	3.1	1.1	0.9	0.8881
2	3.2	1.2	0.91	0.961
3	3.3	1.3	0.92	0.9803
4	3.4	1.4	0.93	0.9906
5	3.5	1.5	0.94	1.0000
6	3.6	1.6	0.95	0.9979
7	3.7	1.7	0.96	0.9879
8	3.8	1.8	0.97	0.9923
9	3.9	1.9	0.98	0.9972
10	3.9999	1.9999	0.9999	0.9952

3.7 Zero–One Testing

The 0–1 test has been explored for checking chaotic system [3]. The 0–1 test evaluates a dynamic system for many rounds of sequences generated. For a real constant c , number of rounds N and sequences $D(n)_{n=1,2,3\dots N}$, the value of K is obtained by (11).

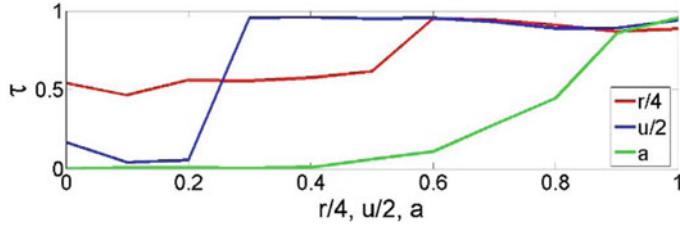


Fig. 7 Result of 0–1 test for BZCLS

$$K = \frac{\log M(n)}{\log n} \quad (11)$$

where

$$M(n) = \lim_{n \rightarrow \infty} \frac{1}{N} \sum_{i=1}^n [p(i+n) - p(i)]^2 + [s(i+n) - s(i)]^2 \quad (12)$$

and

$$p(n) = \sum_{i=1}^n D(i) \cos(i c) \quad (13)$$

$$s(n) = \sum_{i=1}^n D(i) \sin(i c) \quad (14)$$

A given dynamic system is chaotic if $K \approx 1$. We perform 0–1 test for 100 random values of $c \in (\frac{\pi}{5}, \frac{4\pi}{5})$ and for parameter values taken up to 10,000 to show the results.

The control parameters are scaled down to display the result of 0–1 test in $[0, 1]$. Figure 7 shows the results of the 0–1 test for BZCLS with respect to $r/4$, $u/2$, and a . BZCLS passes the 0–1 test for $r \in (2.4, 4)$, $u \in (0.6, 2)$, and $a \in (0.9, 1)$.

3.8 Low Key Space

From the Sects. 3.4, 3.5, and 3.7, the key space comes out to be $r \in (3.9, 4)$ and $u \in (1, 2)$. However, on analysing the profile, trajectory, sensitivity, and entropy values in Sects. 3.1–3.3 and Sect. 3.7, we realize that the system does not have chaotic behaviour for lower values of $u \in (1, 2)$. This further diminishes the system's key space. Consequently, BZCLS exhibits chaoticity in a very small key space which is not sufficient to use it for security applications.

4 Randomness Analysis

We use NIST statistical test suite to check that the considered scheme can generate randomized sequences [20]. We run 14 tests to test randomness of binary sequences. One can apply some other tests which may observe patterns not traceable by NIST tests [21, 22] for more detailed analysis of randomness. To pass a test, the p _value computed must be greater than the significance level α and if the sequence passes all tests, it is considered random; otherwise, considered a non-random sequence. We take 100 binary sequences by transformation of BZCLS output each of length 10^6 to perform randomness tests.

The test results for a binary sequence generated for seed value $x_0 = 0.1234567$ are listed in Table 2. Test results for proportion of passing tests are also shown in Table 2. We see that all the p _values computed are not higher than $\alpha = 0.01$ and the sequences do not pass all the tests. For tested sequences, 5 of the 14 tests are badly failed, and even do not pass any test. Hence, BZCLS does not have the capability to generate sequences of meeting desired randomness criteria.

5 Conclusion

A hybrid chaotic system, BZCLS given by Bao-Zhou-Chen-Liu for multimedia encryption with high information security claimed, has been evaluated in the paper. The profile, trajectory, Lyapunov exponent spectrum, bifurcation, entropy, zero-one test, sensitively, and key space factors for chaoticity analysis, and NIST statistical tests for randomness analysis have been considered to evaluate the effectiveness of BZCLS. It has been found in our analysis that BZCLS possesses poor sensitivity, limited chaotic region, weaker ergodicity, lesser key space, and failure of zero-one test. It has also been seen that the randomness tests badly failed as block frequency, longest run test, rank test, overlapping template matching test, and approximate entropy test fails for almost all the sequences and frequency test, non-overlapping test and random excursions test not qualify the proportionality of passing. Overall, BZCLS has been found insecure and cannot be considered for multimedia security applications. As a future work, the aim is to have a methodology that identifies strong chaotic systems for their optimum and safe utilization in security applications. The work is in progress to develop an efficient algorithm based on fuzzy soft computing approach to find the best one out of several chaotic systems for its usages in security applications.

Table 2 NIST statistical randomness test results for sequences from BZCLS

S. No.	Test name	<i>p</i> _value	Result	Proportion of passing
1	Frequency (monobit) test	0.6821	Pass	50/100
2	Block frequency test	0	Fail	0/100
3	Runs test	0	Fail	0/100
4	Longest runs of ones	0	Fail	0/100
5	Rank test	1	Pass	100/100
6	Non-overlapping template matching test	0.5830	Pass	86/100
7	Overlapping template matching test	1.1139×10^{-6}	Fail	0/100
8	Universal test	0.9192	Pass	100/100
9	Linear complexity test	1	Pass	100/100
10	Serial test	0.391 0.572	Pass Pass	100/100 100/100
11	Approximate entropy test	0	Fail	0/100
12	Cumulative sums test		Pass 1 1	100/100 Pass Pass
13	Random excursions test		Pass 0.1322 0.044 0.0322 0.0277 0.4678 0.9655 0.9895 1	76/100 Pass Pass Pass Pass Pass Pass Pass Pass
14	Random excursions variant test		Pass 1 0.6831 0.7518 0.7893 0.8137 0.8312 0.8445 0.8551 0.8638 0.8711	66/100 Pass Pass Pass Pass Pass Pass Pass Pass Pass Pass

(continued)

Table 2 (continued)

S. No.	Test name	p_value	Result	Proportion of passing
		0.9385		
		1		
		0.8875		
		0.8383		
		0.7928		
		0.7995		
		0.8055		
		0.8111		

References

1. Peng YX, Sun KH, He SB (2020) Dynamics analysis of chaotic maps: from perspective on parameter estimation by meta-heuristic algorithm. *Chin Phys B* 29(3):030502
2. Matthews R (1984) On the derivation of a chaotic encryption algorithm. *Cryptologia* 8:29–41
3. Arif J, Khan MA, Ghaleb B, Ahmad J, Munir A, Rashid U, Al-Dubai AY (2022) A novel chaotic permutation-substitution image encryption scheme based on logistic map and random substitution. *IEEE Access* 10:12966–12982
4. Aouissaoui I, Bakir T, Sakly A, Femmam S (2022) Improved one-dimensional piecewise chaotic maps for information security. *J Commun* 17(1):11–16
5. Lu Q, Linlan Y, Congxu Z (2022) Symmetric image encryption algorithm based on a new product trigonometric chaotic map. *Symmetry* 14(2):373
6. Wu Y, Noonan JP, Agaian S (2011) A wheel-switch chaotic system for image encryption. In: Proceedings on international conference on system science and engineering 2011. IEEE, pp 23–27
7. Kotulski Z, Janusz S (1997) Discrete chaotic cryptography. *Ann Phys* 509(5):381–394
8. Bose R, Banerjee A (1999) Implementing symmetric cryptography using chaos functions. In: Proceedings of the 7th international conference on advanced computing and communications
9. Agarwal M, Yadav A, Ratan R (2022) On the characteristics of 1D-HD chaotic maps for cryptographic applications. *NeuroQuantology* 20(9):6073–6081
10. Bao L, Zhou Y, Chen CP, Liu H (2012) A new chaotic system for image encryption. In: 2012 International conference on system science and engineering (ICSSE). IEEE, pp 69–73
11. Pareek NK, Patidar V, Sud KK (2006) Image encryption using chaotic logistic map. *Image Vis Comput* 24:926–934
12. Wang H, Xiao D, Chen X, Huang H (2018) Cryptanalysis and enhancements of image encryption using combination of the 1D chaotic map. *Signal Process* 144:444–452
13. Mansouri A, Wang X (2020) A novel one-dimensional sine powered chaotic map and its application in a new image encryption scheme. *Inf Sci* 520:46–62
14. Dou Y, Li M (2021) An image encryption algorithm based on a novel 1D chaotic map and compressive sensing. *Multimedia Tools Appl* 80(16):24437–24454
15. Alzaidi AA, Ahmad M, Ahmed HS, Solami EA (2018) Sine-cosine optimization-based bijective substitution-boxes construction using enhanced dynamics of chaotic map. *Complexity*
16. Muthu JS, Murali P (2021) A new chaotic map with large chaotic band for a secured image cryptosystem. *Optik* 242:167300
17. Ratan R (2009) Key independent retrieval of chaotic encrypted images. In: International conference on pattern recognition and machine intelligence. Springer, Berlin, pp 483–488

18. Ratan R, Yadav A (2021) Key independent image deciphering using neighborhood similarity characteristics and divide-and-conquer attack. *Recent Pat Eng* 15(4):34–44
19. Din M, Bhatia AK, Ratan R (2014) Cryptanalysis of Geffe generator using genetic algorithm. In: *Proceedings of the third international conference on soft computing for problem solving*. Springer, New Delhi, pp 509–515
20. NIST Special Publication 800-22 Rev.1a (2010) A statistical test suite for random and pseudorandom number generators for cryptographic applications
21. Ratan R, Jangid BL, Arvind (2020) Bit-plane specific randomness testing for statistical analysis of ciphers. In: *Proceedings of the international conference on soft computing for problem solving 2019, Advances in intelligent systems and computing*, vol. 1138. Springer, Singapore
22. Jangid BL, Ratan R (2020) A new bit plane specific longest repeating pattern test for statistical analysis of bit sequences. In: *Proceedings of the international conference on soft computing: theories and applications, Advances in intelligent systems and computing*, vol. 1154. Springer, Singapore

Identification of Skin Lesion with Adaptive Tasmanian Devil Optimization-Based Transfer Learning



Vineet Kumar Dubey and Vandana Dixit Kaushik

Abstract Skin cancer is a life-threatening and a hazardous disease that produces the death of human in the globe. An advanced development of skin lesion identification techniques is used to detect and identify the skin cancer. However, an early prediction of skin cancer is challenging because of the minimum contrast among the skin portion, melanoma moles, and the maximum color similarity among melanoma-affected and non-affected regions. Hence, this paper introduced the effective skin lesion identification technique for segmenting and identifying the affected region. Here, the skin lesion is identified by the transfer learning (TL) model, namely Adaptive Tasmanian Devil Optimization (ATDO) with convolutional neural network-based transfer learning (CNN-TL). Besides, the TL is modeled by fetching the hyperparameters of AlexNet by training the CNN. The pre-processing is placed by the Gaussian filter and Region of Interest (ROI) extraction process. Moreover, the segmentation process is carried out by the SegNet model. In addition, the ATDO algorithm is modeled by including the adaptive concept in Tasmanian Devil Optimization (TDO) algorithm. Besides, the devised scheme provides the superior performance in the metrics of testing accuracy, sensitivity, and specificity of 0.924, 0.886, and 0.865.

Keywords AlexNet · Gaussian filter · Region of interest · SegNet · Tasmanian devil optimization

1 Introduction

Skin cancer is the general malignancy in Western nations that explicitly accounts for the major kind of skin cancer-based deaths in worldwide. Recently, skin cancer is the collective type of cancer in the entire globe. The general kinds of cancer disease are classified into melanoma, squamous cell carcinoma, and basal cell carcinoma. Among the three kinds of cancer disease, melanoma is the most hazardous and tremendously cancerous. The premature detection of cancer disease can surely

V. K. Dubey (✉) · V. D. Kaushik
Harcourt Butler Technological University, Kanpur 208002, India
e-mail: mrvineetkumardubey@gmail.com

preserve 95% of cases. However, the treatment given through the dermoscopy can preserve the cancer disease between 75 and 84% [1]. The manual detection of skin lesion is human-labor intensive that requires the illuminating and magnified skin images for enhancing the precision of spots. The commonly used algorithms for boosting the dermoscopy is Asymmetry, Border, Color variation, and Diameter (ABCD-rule), 7-point checklist, Menzies method, and 3-point checklist. These algorithms are utilized to observe the malignant melanoma in very initial stage. The manual dermoscopy imaging process is more susceptible to make fault since it requires years of knowledge over tough situations, similarities, huge quantities of visual exploration, and the dissimilarities between various skin lesions [2].

The recently used techniques for detecting the skin lesion is handcrafted-based approaches and deep learning-based methods, which are caused from two probable issues, such as overfitting issues and high computational cost. Generally, an automatic detection of skin cancer commonly involves segmentation and feature mining. However, the procedure of segmentation and feature mining is more complex and challenging, due to the similar shapes of lesions, various skin conditions, and irregularity of lesion. Moreover, the extraction of irrelevant features can lessen the accuracy of segmentation process. Recently, various skin lesion segmentation approaches have been introduced, like thresholding based on type 2 fuzzy approach, stochastic merging segmentation, mean shift algorithm-based segmentation, snaked model-based segmentation, color features, extreme learning machine-based recognition, optimized Histogram of Gradients (HOG) features, and so on. Moreover, the Deep CNN has been introduced to attain the improved performance in classification and recognition. In medical imaging, the small deviations among the lesions, such as melanoma and benign, have reduced the classification accuracy so that the optimal tuning of deep learning is necessary to improve the accuracy [3].

The main invention of this paper is the design of optimization-based TL-based skin lesion identification technique, namely optimized CNN-TL. Here, the TL is modeled by adapting the hyperparameters of AlexNet by training the CNN, and then, the gathered hyperparameters are applied to the layer of CNN. The pre-processing is operated by ROI extraction and Gaussian filter for enhancing the quality of images. The skin lesion identification is done by the CNN-TL wherein the weight of TL is trained by the ATDO algorithm. The main invention of this research is given.

Proposed ATDO + CNN-TL for skin lesion identification: The skin lesion identification is performed by CNN-TL, wherein the hyperparameters of CNN-TL are adjusted by the developed ATDO algorithm. The ATDO is modeled by making the variable in TDO as adaptive.

The association of this paper is explained as below. Section 2 describes the motivation of skin lesion segmentation, Sect. 3 debates the developed model, Sect. 4 shows the result and discussion of planned method, and Sect. 5 provides the conclusion of this paper.

2 Motivation

In medical imaging, the skin cancer diagnosis involves various challenges, like low-contrast lesions, similarity in intraclass, thick hairs, healthy region based on their color and texture, and so on. These challenges make the skin lesion identification process complex. This inspires the researchers for considering the skin lesion identification as research topic.

2.1 Literature Survey

The literature survey of numerous skin lesion identification methods is given underneath. Ratul et al. [2] introduced the Dilated convolution approach for classifying the skin lesions. Here, the TL method is comprised of four dilated networks that provide the improved classification performance. But, the time intake of developed model was high. In order to reduce the processing time, Khan et al. [3] introduced the optimized color feature (OCF) and DCNN model for performing the skin segmentation and classification. However, the developed method provided the degraded performance, due to the selection of irrelevant features. For attaining the better result with the irrelevant features, Nawaz et al. [4] developed the faster region-based CNN (RCNN) and fuzzy k -means clustering (FKM) for performing the skin cancer detection. This method provided the optimal outcome even with low-contrast images, but the training time of this scheme was in maximum. In order to reduce the training period, Jinnai et al. [5] modeled the faster, region-based CNN (FRCNN) to classify the skin cancer. Here, the devised scheme acquired the improved value of positive predictive value, but this method had the difficulty in recognizing the objects from poor-resolution images.

2.2 Challenges

The major challenges of this research paper are explained as below.

- In [2], dilated convolution approach was invented for classifying the skin cancer disease. However, this scheme attained the maximum finishing time and maximum computational complexity.
- In order to lessen the execution period, the devised method in [3] provided the degraded performance, due to the similarity among malignant as well as benign lesions, comparable contrast of benign and malignant lesions as well as the selection of inappropriate features.
- For improving the performance with the small variation in benign and malignant lesions, the FRCNN provided the superior performance. The devised scheme was failed to offer the better result with the real-time applications.

- An automatic skin lesion recognition method using dermoscopic images is a challenging process owing to the issues in lesions, such as irregularity, artifacts, irrelevant feature extraction, and the shape of lesion.

3 Proposed ATDO-Based CNN-TL Algorithm for Skin Lesion Identification

The main aim of the investigation is the progression of ATDO algorithm-based CNN with TL (AlexNet) for skin lesion identification. Initially, the input skin images are fed to the pre-processing to remove the calamities in the images using Gaussian filter [6] and Region of Interest (ROI) extraction. After the pre-processing, the skin lesion segmentation is carried out to segment the affected region using SegNet [7]. The next phase is the feature extraction phase, where statistical features, like mean, variance, standard deviation, kurtosis, skewness, homogeneity, entropy, and CNN features, are extracted. Once the feature extraction is done, the extracted features are dispatched to the skin lesion identification phase, wherein the skin lesion is identified using AlexNet with CNN-TL model [8] in which the CNN is applied with the hyperparameters from AlexNet. Moreover, the weight optimization of CNN is done by the projected optimization algorithm, named ATDO algorithm. In addition, the ATDO is formed by including the adaptive concept with TDO [9] algorithm. The schematic outlook of invented ATDO-based CNN-TL for skin lesion identification is depicted in Fig. 1.

3.1 Data Acquisition

The considered dataset G containing k number of images that are formulated as,

$$G = \{G_1, G_2, \dots, G_d, \dots, G_k\} \quad (1)$$

where, G_d denotes the d th image count and k be the total number of images. Moreover, the image G_d is used for an input of pre-processing.

Pre-processing The pre-processing is done to eliminate the noise in the input images that improves the consequences of skin lesion identification process. The pre-processing is done in two methods, such as Gaussian filtering and ROI extraction that are explained as below.

Gaussian Filter Generally, the Gaussian filter [6] is used to remove the noise by blurring the images. Here, the Gaussian filtering is applied to the input image G_d such that the noise in the image is removed. Moreover, the expression becomes,

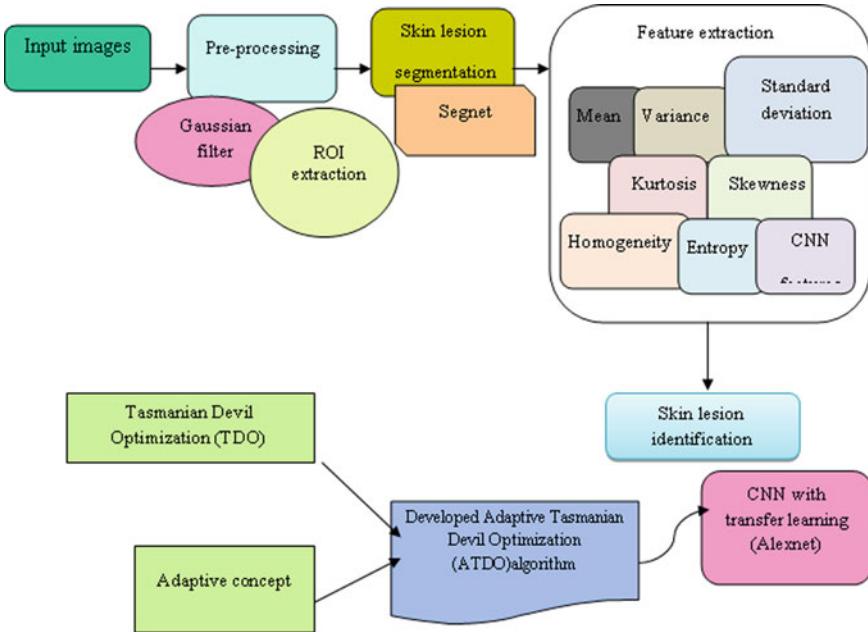


Fig. 1 Block diagram of projected ATDO-based CNN-TL for skin lesion identification

$$M(a) = \frac{1}{\sqrt{2\pi}\varepsilon^2} e^{-\frac{(a-\eta)^2}{2\varepsilon^2}} \quad (2)$$

where ε denotes the standard deviation, a be the gray level image, and η denotes the mean. Thus, the outcome of Gaussian filtering is expressed as d , which is again pre-processed by ROI extraction.

ROI Extraction The blurred and noise-free image d acquired from the Gaussian filtering is given to the input of ROI extraction in order to further smoothen the images. The ROI extraction [10] process considers the noise-free image d , and then, the image smoothening process is done by performing the contrast analysis and local texture feature extraction that improves the efficiency of skin lesion segmentation. Thus, the ROI extracted image is designated as W and is fed to the skin lesion segmentation process.

3.2 Skin Lesion Segmentation Using SegNet

The pre-processed image W is subjected to the SegNet model [7] in order to separate the affected part from the normal images. The advantage of SegNet model is that it is easy to train and it requires lesser computational resources. Moreover, the parameters

of SegNet model are tuned, and then, the skin lesion area is segmented in the testing process. The SegNet model reused the pooling indices, and then, it requires less memory. The SegNet model is comprised of encoding and the decoding path followed by the classification layer. In the encoder, the number of conv layer is equivalent to the number of conv layer in VGG16 network. Moreover, each encoder contains the corresponding decoder layer. The outcome of decoder layer is dispatched to the classification layer to make the final segmented outcome, which is indicated as P .

3.3 Feature Extraction

The segmented region P is sent to the feature extraction process for excavating the significant features from the segmented area. In this research, the statistical features [11], such as mean, variance, standard deviation, kurtosis, skewness, homogeneity, entropy, and CNN features, are extracted from the segmented area. Each of these features is extracted and then fused into a single feature vector in order to get the final feature vector.

Mean The mean value of image is calculated by summing every pixel of an image to the overall number of pixels, that is indicated as,

$$m_1 = \frac{\sum_{a=1}^h b_a}{p} \quad (3)$$

where b_a denotes the pixel value of image and p be the overall pixel number. Thus, the mean feature is notified as m_1 .

Variance The variance is a metric that is used to measure the squared variations among mean and pixel value of image. It is formulated as,

$$m_2 = \frac{\sum_{a=1}^h (b_a - \mathfrak{M})^2}{p - 1} \quad (4)$$

where \mathfrak{M} signifies the pixel value and the variance feature is represented as m_2 .

Standard Deviation

It is stated as the square root of variance and is described as,

$$m_3 = \sqrt{\frac{\sum_{a=1}^h (b_a - \mathfrak{M})^2}{p - 1}} \quad (5)$$

Kurtosis The kurtosis is used to measure the flatness and is derived based on the mean and standard deviation, which is indicated as,

$$m_4 = \frac{\mathfrak{M}_4}{\omega^4} \quad (6)$$

Here, \mathfrak{M}_4 denotes the fourth central moment and ω_4 denotes the standard deviation. Thus, the kurtosis feature is indicated as m_4 .

Skewness The symmetry of an image is described by the term called as skewness, which is denoted as m_5 and the corresponding equation becomes,

$$m_5 = E[(B - \mathfrak{M}\omega)^3] \quad (7)$$

Homogeneity It is a metric, which is used to measure the local information that represents the linearity of computed areas and is indicated as m_6 .

Entropy The entropy feature is used to elaborate the randomness, and it is used to expose the texture of an image. Moreover, the entropy feature is expressed as m_7 .

The final statistical feature is got by combining all the statistical features, which is exposed by,

CNN Features CNN feature is used to extract the size, texture, and boundary of a required region and is indicated as m_c . The CNN feature is extracted with the assistance of CNN model, which contains the convolutional layer, pooling layer, and the fully connected layer.

In addition, the final feature vectors are perceived by merging the CNN features and statistical features, which is given by,

$$m = \{m_s, m_c\} \quad (8)$$

where m_s denotes the statistical features and m_c denotes the CNN features.

3.4 Skin Lesion Identification Using CNN-TL Model

After extracting the overall feature vector m , then the skin lesion identification is carried out using the TL network, namely CNN-TL model [8, 12]. The concept of TL relies on applying the hyperparameters of AlexNet with the CNN model. Figure 2 shows the AlexNet with CNN-TL model. Here, the extracted feature m is allowed to the AlexNet for fetching the hyperparameters by training it. The fetched hyperparameters are utilized for replacing the constraints of CNN. In addition, the weight and bias of CNN are trained by the ATDO algorithm, which is formed by applying the adaptive concept with the TDO algorithm [9].

Pre-trained CNN Scheme CNN [8] is a network to map the input image into the output. Generally, CNN is composed of conv layer, pool layer, activation layer, and softmax layer. Initially, the conv layer extracts the edges of an image, after that the



Fig. 2 CNN-TL model

pooling layer computes the mean of every extracted feature map. In addition, each function of CNN is considered as layer that considers the outcome of preceding layer as an input. In the case of TL, CNN acts as a feature extractor or classifier after the training process. TL is a method to enhance the behavior of machine learning by adapting the intelligence of any other model. In this research, the pre-trained CNN with AlexNet is adapted.

AlexNet Model AlexNet [12] is the CNN architecture, which is designed by Alex Krizhevsky with Ilya Sutskever and Geoffrey Hinton. AlexNet is a classifier model, which is used to detect various classes using deep layers. The AlexNet is comprised of five conv layer, 3 pooling layer, 2 FC layer, and one softmax layer. In this research, the parameters of AlexNet are acquired by training the AlexNet model. Moreover, the fetched parameters are used to train the CNN. In the testing process, the feature vector is directly presented to the CNN so that the skin lesion is identified correctly. Moreover, the skin lesion identification outcome is indicated as z_k .

Training of CNN Using Proposed Adaptive Tasmanian Devil Optimization Algorithm This section explains the ATDO algorithm, which is used to tune the parameters of CNN-TL model. Here, the ATDO is modeled by integrating the adaptive concept with the TDO algorithm. TDO algorithm [9] is a bio-inspired metaheuristic approach, which is modeled by copying the behavior of Tasmanian devil. In the TDO, the location of Tasmanian is updated in two strategies, such as feeding and prey attacking. The developed ATDO is designed by making the random variable adaptive. Then, the algorithmic processes are given below.

Initialization In the initialization step, the members of TDO are modeled based on the location and are formulated as,

$$N = \begin{bmatrix} N_1 \\ \vdots \\ N_x \\ \vdots \\ N_w \end{bmatrix}_{w \times v} = \begin{bmatrix} n_{1,1} & \cdots & n_{1,u} & \cdots & n_{1,v} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{x,1} & \cdots & n_{x,u} & \cdots & n_{x,v} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{w,1} & \cdots & n_{w,u} & \cdots & n_{w,v} \end{bmatrix}_{w \times v} \quad (9)$$

where N signifies the population of Tasmanian devils, N_x signifies the candidate solution of x th value, w denotes the count of Tasmanian devils, and v be the count of defect variable.

Fitness Measure The fitness measure is used to calculate the best solution so that the mean squared value (MSE) is deliberated as a fitness function in this case. Thus, the expression for fitness function is given by,

$$W = \frac{1}{f} \sum_{k=1}^f [z_k^* - z_k]^2 \quad (10)$$

Here, z_k signifies the projected outcome of TL model, f be the entire sample count and z_k^* specifies the expected outcome.

Exploration Phase The feeding of eating carrion is done in the exploration phase. Sometimes, the Tasmanian devils like to eat carrion without participating in hunting process. The other animals that live nearer to the Tasmanian devils may hunt huge animals so that they are unable to completely eat the entire animal. In this case, the Tasmanian devils prefers to eat the remaining flesh and that behaviors are modeled in the exploration phase. Moreover, the location of members in the population is considered as the location of carriions. Thus, the expression becomes,

$$B_x = C_t, \quad x = 1, 2, \dots, w, \quad t \in \{1, 2, \dots, w\} | t \neq x \quad (11)$$

where B_x specifies the carriion chosen by the Tasmanian devils. By considering the chosen carriion, the next location of Tasmanian devil is computed. If the fitness function is better than the new location, then the preceding position is considered as the best solution and the corresponding equation becomes,

$$n_{x,u}^{\text{new, P1}} = \begin{cases} n_{x,u} + q \cdot (b_{x,u} - R \cdot n_{x,u}), & W_{B_x} < W_x; \\ n_{x,u} + q \cdot (n_{x,u} - b_{x,u}) & \text{otherwise,} \end{cases} \quad (12)$$

where, $n_{x,u}^{\text{new, P1}}$ be the new location of Tasmanian devils in first strategy, W_{B_x} be the objective function of nominated carriion, R be the random number between $[1, 2]$ and q is the random variable lies between $[0, 1]$, which is made adaptive.

$$r = 0.02 \left(1 - \frac{t}{T}\right) \times \sigma \quad (13)$$

where T specifies the maximum iteration count, t denotes the present iteration and σ specifies the neighborhood radius.

Exploitation Process The second strategy of TDO is to hunt and eat prey, which involves two stages. The 1st stage is to choose and chase the prey and the 2nd stage is to stop the chasing and eat the prey, and the selection of prey is done using,

$$Y_x = C_t, \quad x = 1, 2, \dots, w, \quad t \in \{1, 2, \dots, w | t \neq x\} \quad (14)$$

where Y_x specifies the chosen prey. After selecting the prey, then the location of Tasmanian devil is computed.

Re-evaluation The re-evaluation process is done in terms of the fitness function. After the exploration and exploitation phase, the location of Tasmanian devils is computed separately. Furthermore, the best location is updated based on the objective function.

Termination All the above mentioned steps are continual till the optimal solution is attained. Moreover, the algorithmic processes of ATDO algorithm are specified in algorithm 1.

Algorithm 1. Pseudo code of ATDO

Initialize the algorithmic parameters N , v , and n_x

Set the constraints T and w

Compute the objective function by Eq. (10)

If the probability $P < 0.5$

Choice B_x value and compute $n_{x,u}^{\text{new}, P1}$ using Eqs. (11) and (12)

Update the best value using Eq. (10)

else

Select Y_x , calculate $n_{x,u}^{\text{new}, P1}$ using Eqs. (12) and (14) and update the best value using Eq. (10)

end

Calculate the feasibility

Terminate

The devised ATDO algorithm is modeled by integrating the adaptive concept in TDO algorithm such that the best solution is attained.

4 Results and Discussion

The results and discussion of devised CNN-TL + ATDO model are explained in this section.

4.1 Experimental Arrangement

The devised CNN-TL + ATDO model is implemented using Python tool using personal computer with Windows 10 OS and Intel i3 core processor.

4.2 *Explanation of Utilized Dataset*

The dataset used for the CNN-TL + ATDO skin lesion identification technique is SIIM-ISIC Melanoma classification dataset [13]. The SIIM-ISIC Melanoma classification dataset contains both metadata and image that is stored in TR and JPEG format having the dimension of 1024×1024 . In addition, the metadata is stored in the format of DICOM.

4.3 *Evaluation Metrics*

The metrics used for examining the effectiveness of CNN-TL + ATDO are testing accuracy, sensitivity, and specificity.

Testing Accuracy Testing accuracy is used to measure the accurateness of skin lesion detection that is exposed by,

$$k_1 = \frac{a + g}{a + g + m + n} \quad (15)$$

where a denotes the true positive, true negative is indicated as g , m depicts the false positive and n indicates the false negative.

Sensitivity It is the measure of effectively identifying diseased person, which is depicted as,

$$K_2 = \frac{a}{a + n} \quad (16)$$

Specificity It is defined as the measure of effectively identifying non-diseased person that is formulated as,

$$K_3 = \frac{g}{g + m} \quad (17)$$

4.4 *Experimental Images*

This part explains the experimentation outcome image of CNN-TL + ATDO model for skin lesion identification. Figure 3a indicates the input image, Fig. 3b explores the pre-processed image after applying the Gaussian filtering, Fig. 3c elucidates the pre-processed image after applying the ROI extraction, and Fig. 3d illustrates the segmented outcome of developed model.

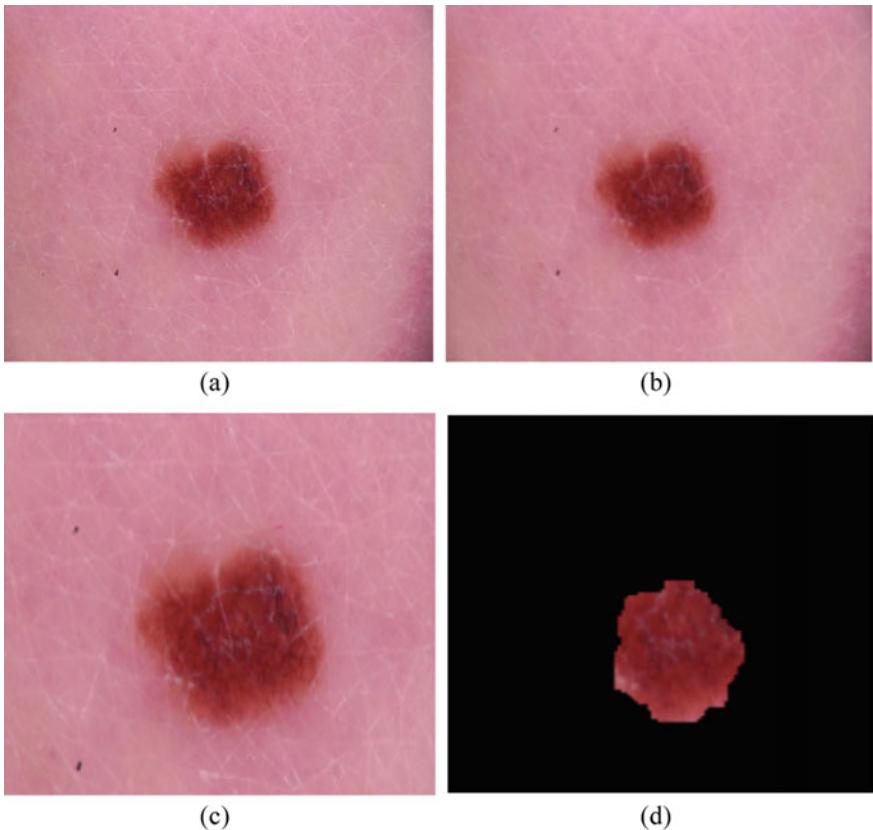


Fig. 3 Experimentation image outcome of invented model for, **a** input image, **b** pre-processing image based on Gaussian filter, **c** pre-processing image based on ROI extraction, **d** segmented image outcome

4.5 Algorithmic Assessment Methods

The methods used for assessing the devised ATDO + CNN_TL algorithm are Particle Swarm Optimization (PSO) + CNN_TL [8, 14], Aquila Optimizer (AO) + CNN_TL [8, 15], Social Optimizer (SO) algorithm + CNN_TL [8, 16], Rider + CNN_TL [8, 17], and TDO + CNN_TL [8, 9].

Population-Based Assessment The analysis of devised ATDO + CNN_TL is done by altering the population size. The testing accuracy of proposed ATDO + CNN_TL is shown in Fig. 4a. The testing accuracy of PSO + CNN_TL is 0.726, AO + CNN_TL is 0.745, SO + CNN_TL is 0.775, Rider + CNN_TL is 0.795, and TDO + CNN_TL is 0.836, whereas the testing accuracy of ATDO + CNN_TL is 0.896 for the population size 50. The performance improvements of ATDO + CNN_TL with existing algorithms are 18.87, 16.77, 13.42, 11.18, and 6.65%. The sensitivity

of ATDO + CNN_ TL is given in Fig. 4b. As selecting the population size as 50, then the sensitivity of PSO + CNN_ TL, AO + CNN_ TL, SO + CNN_ TL, Rider + CNN_ TL, TDO + CNN_ TL, and devised ATDO + CNN_ TL is 0.676, 0.706, 0.723, 0.738, 0.781, and 0.832. The performance upgrading of ATDO + CNN_ TL with existing algorithms is 18.81, 15.14, 13.17, 11.32, and 6.14%. Figure 4c exposed the specificity of ATDO + CNN_ TL. If the size of population is 50, then the specificity values are 0.689, 0.709, 0.736, 0.756, 0.762, and 0.809. The performance upgrading of ATDO + CNN_ TL with existing algorithms is 14.96, 12.38, 9.06, 6.59, and 5.76%.

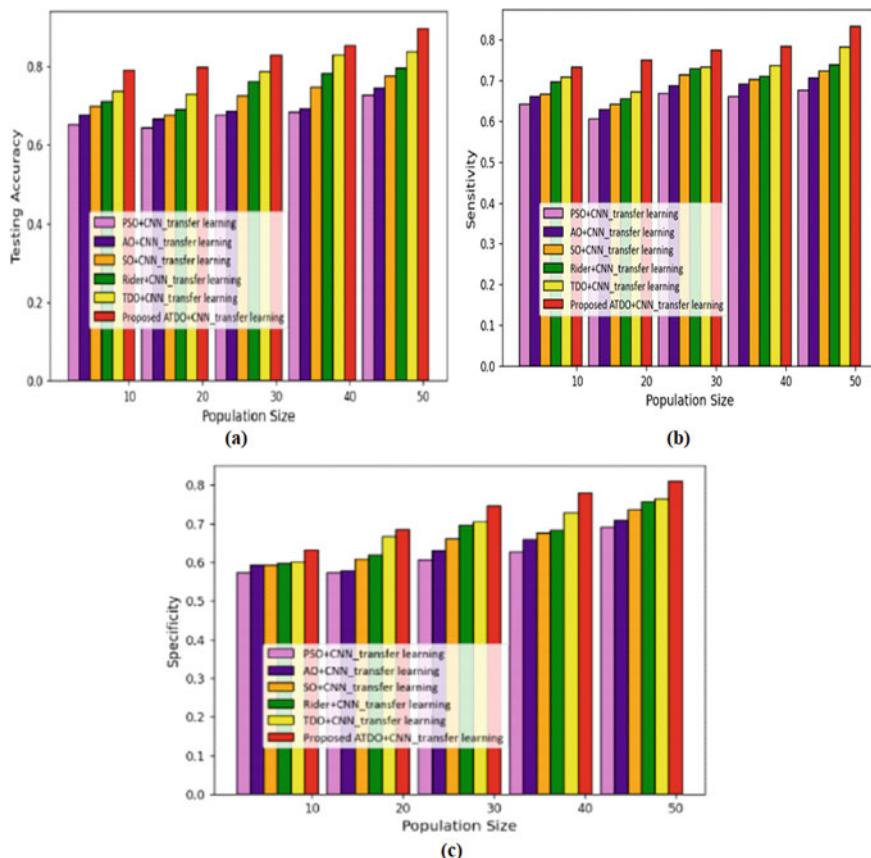


Fig. 4 Comparison of CNN-TL + ATDO model based on, **a** testing accuracy, **b** sensitivity, **c** specificity

4.6 Comparative Methods

The comparison of invented CNN-TL + ATDO model is performed by comparing the new model with the state-of-the-art methods, namely Dilated convolution [6], DCNN [7], RCNN_FKM [8], FRCNN [9], and CNN_TL + TDO.

4.7 Comparative Investigation

The comparative investigation of proposed CNN-TL + ATDO model is done by adjusting the training epoch.

Based on Training Epoch The testing accuracy of newly invented CNN-TL + ATDO model is exposed in Fig. 5a. The testing accuracies, such as 0.743, 0.766, 0.792, 0.807, 0.858, and 0.924, are achieved by the comparison methods and invented CNN-TL + ATDO model, while the train epoch is 90. Thus, the enhanced performance of devised model is 19.55, 17.06, 14.34, 12.68, and 7.14%. The sensitivity graph of CNN-TL + ATDO model is given in Fig. 5b. When the percentage of train epoch is 90, then the CNN-TL + ATDO model got the sensitivity of 0.886, and the previously invented schemes got the sensitivity of 0.716, 0.737, 0.750, 0.782, and 0.805 as the train epoch is 90. Thus, the improved performance of devised model is 19.10, 16.75, 15.32, 11.77, and 9.09%. The graphical depiction of specificity for CNN-TL + ATDO model is visualized in Fig. 5c. While selecting the train epoch is 90, then the specificity of CNN-TL + ATDO model is 0.865 and the existing strategies are 0.698, 0.727, 0.74, 0.790, and 0.845. Moreover, the upgraded performance of devised model is 19.31, 15.92, 14.31, 8.67, and 2.28%.

4.8 Comparative Discussion

Table 1 displays the comparison table of CNN-TL + ATDO scheme with state-of-the-art approaches based on the performance metrics. Here, the testing accuracy value of CNN-TL + ATDO is 0.924, sensitivity is 0.886, and specificity is 0.865. Similarly, the testing accuracies of comparison methods are 0.743, 0.766, 0.792, 0.807, and 0.858, sensitivity is 0.716, 0.737, 0.750, 0.782, and 0.805, and then the specificity is 0.698, 0.727, 0.741, 0.790, and 0.845.

From the analysis, the developed CNN-TL + ATDO scheme accomplished the greater performance due to the better performance of adapted concept. Here, the optimal performance is attained by presenting the adaptive concept in TDO algorithm so that the performance is improved in terms of assessment metrics.

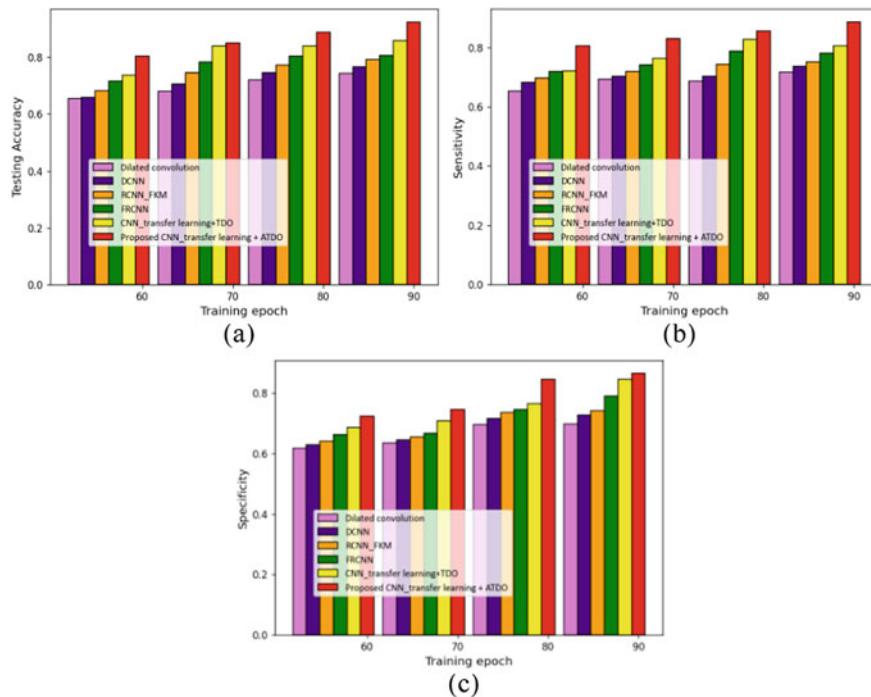


Fig. 5 Algorithmic evaluation of invented CNN-TL + ATDO model based on, **a** testing accuracy, **b** sensitivity, and **c** specificity

Table 1 Comparative discussion

Metrics	Dilated convolution	DCNN	RCNN-FKM	FRCNN	CNN_TL + TDO	Proposed CNN_TL + ATDO
Testing accuracy	0.743	0.766	0.792	0.807	0.858	0.924
Sensitivity	0.716	0.737	0.750	0.782	0.805	0.886
Specificity	0.698	0.727	0.741	0.790	0.845	0.865

5 Conclusion

This paper exposes the devised CNN-TL + ATDO scheme for skin lesion segmentation. Here, the series of steps followed in devised model are pre-processing, skin lesion segmentation, feature extraction, and skin lesion identification. Here, the pre-processing method removes the noise exists in the image using ROI extraction and Gaussian filtering method. The affected region from the image is segmented by SegNet model that helps to progress the efficiency of detection process. In addition,

the skin lesion detection is carried out using CNN_TL in which the hyperparameters of TL are trained by the developed ATDO algorithm. The experimental outcome shows that the devised model attained the superior performance based on the testing accuracy of 0.924, sensitivity of 0.886, and specificity of 0.865, correspondingly. In future, the performance of CNN-TL + ATDO scheme can be improved by applying some other effective concepts in optimization approaches.

References

1. Carli P, Quercioli E, Sestini S, Stante M, Ricci L, Brunasso G, De Giorgi V (2003) Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. *Br J Dermatol* 148(5):981–984
2. Ratul MAR, Mozaffari MH, Lee WS, Parimbelli E (2020) Skin lesions classification using deep learning based on dilated convolution. *BioRxiv* 860700
3. Khan MA, Sharif MI, Raza M, Anjum A, Saba T, Shad SA (2019) Skin lesion segmentation and classification: a unified framework of deep neural network features fusion and selection. *Expert Syst* e12497
4. Nawaz M, Mehmood Z, Nazir T, Naqvi RA, Rehman A, Iqbal M, Saba T (2022) Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. *Microsc Res Tech* 85(1):339–351
5. Jinna S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R (2020) The development of a skin cancer classification system for pigmented skin lesions using deep learning. *Biomolecules* 10(8):1123
6. Kumar A, Sodhi SS (2020) Comparative analysis of gaussian filter, median filter and denoise autoencoder. In: Proceedings of 2020 7th international conference on computing for sustainable global development (INDIACoM), pp 45–51
7. Alqazzaz S, Sun X, Yang X, Nokes L (2019) Automated brain tumor segmentation on multi-modal MR image using SegNet. *Comput Vis Media* 5(2):209–219
8. Shi Z, Hao H, Zhao M, Feng Y, He L, Wang Y, Suzuki K (2019) A deep CNN based transfer learning method for false positive reduction. *Multimedia Tools Appl* 78(1):1017–1033
9. Dehghani M, Hubálovský S, Trojovský P (2022) Tasmanian devil optimization: a new bio-inspired optimization algorithm for solving optimization algorithm. *IEEE Access* 10:19599–19620
10. Kim B, Kehtarnavaz N, LeBoulluec P, Liu H, Peng Y, Euhus D (2013) Automation of ROI extraction in hyperspectral breast images. In: Proceedings of 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 3658–3661
11. Meraj T, Rauf HT, Zahoor S, Hassan A, Lali MI, Ali L, Bukhari SAC, Shoaib U (2021) Lung nodules detection using semantic segmentation and classification with optimal features. *Neural Comput Appl* 33(17):10737–10750
12. Rahman T, Chowdhury ME, Khandakar A, Islam KR, Islam KF, Mahbub ZB, Kadir MA, Kashem S (2020) Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Appl Sci* 10(9):3233
13. SIIM-ISIC Melanoma classification datasets will be acquired from, <https://www.kaggle.com/c/siim-isic-melanoma-classification/data.accessed>
14. Poli R, Kennedy J, Blackwell T (2007) Particle swarm optimization. *Swarm Intell* 1(1):33–57

15. Abualigah L, Yousri D, AbdElaziz M, Ewees AA, Al-Qaness MA, Gandomi AH (2021) Aquila optimizer: a novel meta-heuristic optimization algorithm. *Comput Ind Eng* 157:107250
16. Karimi N, Khandani K (2020) Social optimization algorithm with application to economic dispatch problem. *Int Trans Electr Energy Syst* 11:e12593
17. Binu D, Kariyappa BS (2018) RideNN: a new rider optimization algorithm-based neural network for fault diagnosis in analog circuits. *IEEE Trans Instrum Meas* 68(1):2–26

Synchronization of MLS Chaotic System Using Sliding Mode Control Technique



Pallav and Himesh Handa

Abstract A novel 4-D chaotic system recognized as the modified Lorenz-Stenflo (MLS) chaotic system is the subject of this research study, in which a synchronization tactic on the basis of integral SMC has been developed for the system. One way to construct a sliding mode controller is to pick the integrated switching surface in a suitable manner. This controller ensures that a sliding motion will occur and accomplishes the task of synchronization of MLS chaotic systems in a drive-response arrangement. The synchronization of master and slave system has achieved in about 7 s from the time control signal is activated. The method offers a different approach to constructing a controller for this system from the one that was suggested in (Handa and Sharma in TENCON 2011–2011 IEEE Region 10 Conference, 2011, [1]). In the end, numerical simulations are carried out in order to test how successful the new controller is.

Keywords Chaos · Synchronization · Integral sliding mode controller · Drive system · Response system · MLS chaotic system

Pallav (✉) · H. Handa

National Institute of Technology, Hamirpur, Himachal Pradesh 177005, India
e-mail: pallav.sahay@nith.ac.in

H. Handa
e-mail: himeshhanda@nith.ac.in

1 Introduction

Over the course of the last two decades, the research community has placed a significant amount of emphasis on the topic of chaotic system synchronization [2, 3]. Dynamical systems that are very susceptible to changes in their initial conditions are referred to be chaotic systems. Pecora and Carroll were the first people to propose the concept of the synchronization phenomena of chaos in the year 1990. They did it for two identical systems with distinct beginning circumstances [4]. Since then, the phenomena of chaotic systems synchronizing with each other have received a lot of interest in a variety of scientific and technological disciplines. A number of control strategies, such as active control technique [5], nonlinear active control technique [6], output feedback control methodology [7], adaptive synchronization [8], feedback control technique [9], backstepping control approach [10], and sliding mode control scheme [11], among others, have been anticipated in order to accomplish synchronization of chaotic systems that are evolving from a variety of starting conditions.

Applications for chaos synchronization include secure communication, cryptography, robotics, laser physics, ecology, chemical systems, biological systems, cardiology, and a great deal more besides [12, 13]. It is not that chaos is beneficial in all scenarios; rather sometimes, it has the opposite effect. For example, it is not suitable for mechanical or power systems [14]. Likewise, it is not always appropriate for systems.

Usually, in most of the cases, there are two chaotic systems involved in the chaos synchronization process. The first one is recognized as the drive system, while the second one is seen as the response system. When developing the control methodology inside the drive-response synchronization arrangement, it is done so in a way that the states of the reaction system go along with the states of the drive system.

On the basis of the original Lorenz system, researcher Lennart Stenflo developed a new system identified as the Lorenz-Stenflo (LS) system [15] in the year 1996. After substituting one of the quadratic-nonlinear elements of the LS system with a PWL function, a novel 4-D chaotic system was developed. This system was given the name Modified Lorenz-Stenflo (MLS) chaotic system. The piecewise linear function, often known as the PWL function, is almost entirely made up of the standard sign function [16].

The task of constructing a control method to synchronize MLS chaotic systems utilizing two under-actuated control signals is the focus of the work that will be presented in this article as part of the suggested solution. In order to accomplish this objective, the SMC methodology and the Lyapunov stability approach are used

in the design of an integral sliding surface. This helps to assure the stability of the closed-loop error system while it is in sliding motion. The control scheme that has been presented ensures the existence of sliding motion and furthermore accomplishes synchronization of MLS chaotic systems in a drive-response arrangement. To explain the viability of the intended technique, numerical simulations have been provided toward the conclusion of the article.

The structure of the leftover segment of the document is as follows: In Sect. 2, an overview of MLS chaotic system is offered. A proposal has been made in Sect. 3 about the strategy of sliding surface and sliding mode controllers with the intention of synchronization of the MLS chaotic system. In Sect. 4, the findings of the numerical simulation are offered in order to demonstrate how effective the suggested controller is. The last section, which is a summary of the whole thing, can be found in Sect. 5.

2 Introduction of MLS Chaotic System

Lennart Stenflo conducted research on a number of different equations that regulate atmospheric waves in 1996. As a result of his efforts, he was competent to successfully develop a collection of equations that regulate atmospheric waves by making use of low-frequency and short-wavelength estimates. He created the Lorenz-Stenflo (LS) system of the fourth order by adhering to the same procedure. This LS system is comparable to the well-known Lorenz system; however, it differs in that it has a new control parameter denoted by the letter d and a new state variable denoted by the letter x_4 . The differential equations explaining the Lorenz-Stenflo (LS) system may be expressed in the following manner:

$$\begin{cases} \dot{x}_1 = a(x_2 - x_1) + dx_4 \\ \dot{x}_2 = x_1(c - x_3) - x_2 \\ \dot{x}_3 = x_1x_2 - bx_3 \\ \dot{x}_4 = -x_1 - ax_4 \end{cases} . \quad (1)$$

The LS system displays chaotic behavior when its parameters are a , b , c , and d , respectively, set to 1, 0.7, 26, and 1.5, as seen in Fig. 1.

When a piecewise linear (PWL) function is substituted for the quadratic-nonlinear component x_1x_2 in the 3rd equation of (1), a distinct chaotic system of 4th order is produced. This system has got the name MLS chaotic system because of its mathematical properties.

$$\begin{cases} \dot{x}_1 = a(x_2 - x_1) + dx_4 \\ \dot{x}_2 = x_1(c - x_3) - x_2 \\ \dot{x}_3 = \text{sign}(x_2)x_1 - bx_3 \\ \dot{x}_4 = -x_1 - ax_4 \end{cases} \quad (2)$$

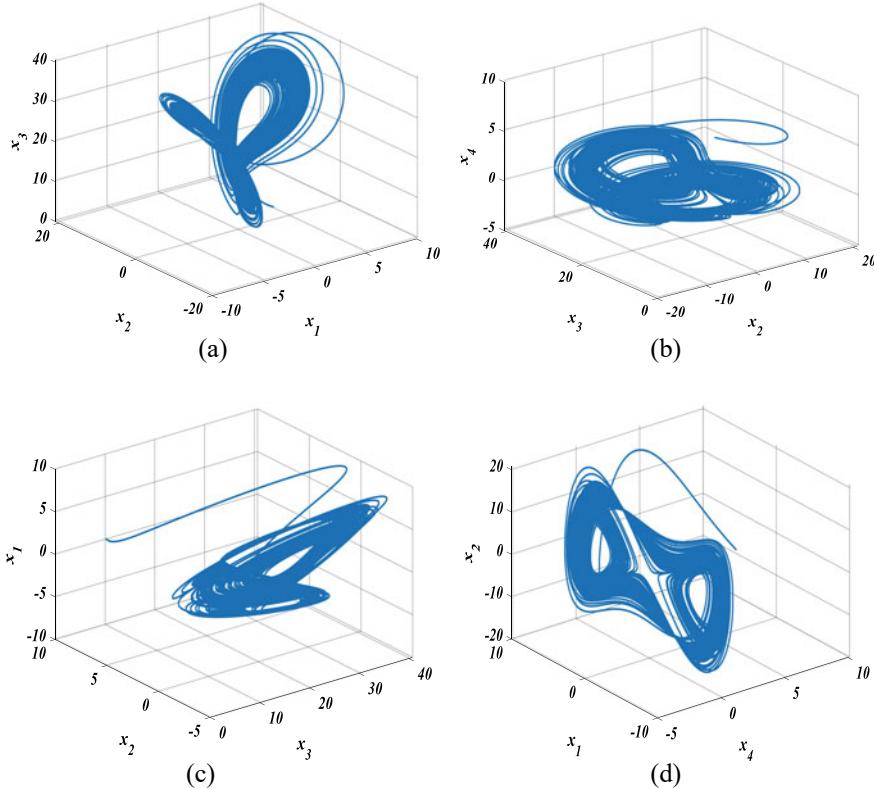


Fig. 1 a-d 3-D chaotic pattern of Lorenz-Stenflo (LS) system

It should be obvious from the differential equations shown earlier that the traditional sign function is being used in the third equation of dynamics presented in (2). The new system, i.e., MLS chaotic system displays a chaotic behavior, as seen in Fig. 2, when the parameters a , b , c , and d correspondingly are set to 1, 0.7, 26, and 1.5.

3 Synchronization of Two MLS Chaotic Systems

In this segment of the paper, the challenge of synchronizing two MLS chaotic systems that are identical by employing an integral sliding mode controller is designed. SMC is one of the nonlinear control methodologies which promises control strategy. Maintaining the system nodes on the sliding surface is the means through which stability may be accomplished in this manner.

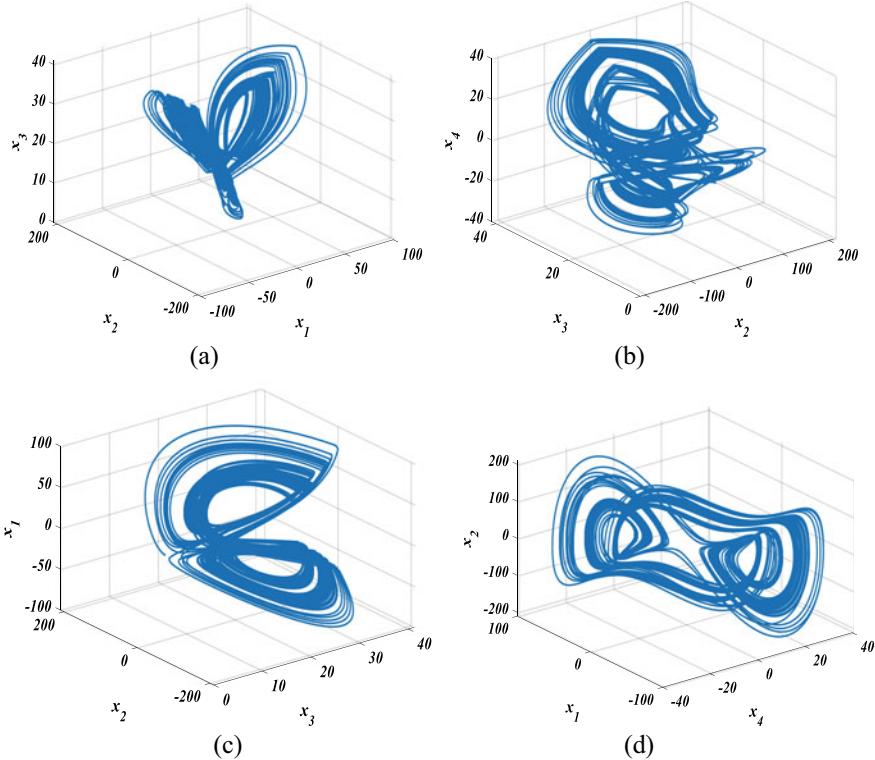


Fig. 2 a-d 3-D chaotic behavior of MLS chaotic system

If system (2) is considered to be the drive or master system, then the response or slave system can likewise be defined as follows:

$$\begin{cases} \dot{y}_1 = a(y_2 - y_1) + dy_4 \\ \dot{y}_2 = y_1(c - y_3) - y_2 + u_1 \\ \dot{y}_3 = \text{sign}(y_2)y_1 - by_3 \\ \dot{y}_4 = -y_1 - ay_4 + u_2 \end{cases} . \quad (3)$$

where u_1 and u_2 refer to the control inputs that are being considered for the response system.

Let the error dynamics between (3) and (2) be

$$e_i = y_i - x_i \text{ for } i = 1 - 4;$$

Error dynamics may be computed as the following:

$$\begin{cases} \dot{e}_1 = a(e_2 - e_1) + de_4 \\ \dot{e}_2 = (c - y_3)e_1 - e_2 - y_1e_3 + e_1e_3 + u_1 \\ \dot{e}_3 = \text{sign}(y_2)y_1 - \text{sign}(x_2)x_1 - be_3 \\ \dot{e}_4 = -e_1 - ae_4 + u_2 \end{cases}. \quad (4)$$

Based on the principle of sliding mode [17, 18], consider the following integral sliding surface

$$\begin{aligned} \dot{\mathcal{S}}_1 &= a(e_2 - e_1) + de_4 + k_1e_1 + (c - y_3)e_1 - e_2 - y_1e_3 + e_1e_3 + k_2e_2 + u_1 \\ \dot{\mathcal{S}}_2 &= \text{sign}(y_2)y_1 - \text{sign}(x_2)x_1 - be_3 + k_3e_3 - e_1 - ae_4 + k_4e_4 + u_2. \end{aligned} \quad (5)$$

Assuming $\dot{\mathcal{S}} = 0$ (Itkis and Utkin condition),

$$\begin{cases} u_1 = -a(e_2 - e_1) - de_4 - k_1e_1 - (c - y_3)e_1 + e_2 + y_1e_3 \\ \quad - e_1e_3 - k_2e_2 \\ u_2 = -\text{sign}(y_2)y_1 + \text{sign}(x_2)x_1 + be_3 - k_3e_3 + e_1 + ae_4 \\ \quad - k_4e_4 \end{cases} \quad (6)$$

Now, based on the exponential reaching law [17], the equivalent control signals are as follows:

$$\begin{cases} u_{1\text{eq}} = -a(e_2 - e_1) - de_4 - k_1e_1 - (c - y_3)e_1 + e_2 + y_1e_3 \\ \quad - e_1e_3 - k_2e_2 - r\mathcal{S}_1 - \rho \text{sgn} \mathcal{S}_1 \\ u_{2\text{eq}} = -\text{sign}(y_2)y_1 + \text{sign}(x_2)x_1 + be_3 - k_3e_3 + e_1 + ae_4 \\ \quad - k_4e_4 - r\mathcal{S}_2 - \rho \text{sgn} \mathcal{S}_2 \end{cases} \quad (7)$$

where r and ρ are positive real constants.

Theorem: *If the integral SMC law (7) is used for the drive-response MLS chaotic systems represented by (2) and (3), they are globally and asymptotically synchronized for every initial condition $x(0)$ and $y(0)$.*

Proof: The result can be justified by referring to the Lyapunov stability theory [19].

Let us assume that the Lyapunov function is represented by (8), which is a positive definite function. In light of the Lyapunov stability theorem, in order to demonstrate that the error dynamics (4) are stable, it is necessary to demonstrate that the derivative of the Lyapunov function represented by (8) has a value which is less than zero.

The following is the suggested expression for the Lyapunov function:

$$V = \frac{1}{2}(\mathcal{S}_1^2 + \mathcal{S}_2^2). \quad (8)$$

And the time derivative of (8) is as follows:

$$\dot{V} = \mathcal{S}_1 \dot{\mathcal{S}}_1 + \mathcal{S}_2 \dot{\mathcal{S}}_2.$$

Now, using Eq. (5),

$$\begin{aligned} \dot{V} = & \mathcal{S}_1(a(e_2 - e_1) + de_4 + k_1 e_1 + (c - y_3)e_1 - e_2 - y_1 e_3 + e_1 e_3 \\ & + k_2 e_2 + u_1) + \mathcal{S}_2(\text{sign}(y_2)y_1 - \text{sign}(x_2)x_1 - be_3 + k_3 e_3 \\ & - e_1 - ae_4 + k_4 e_4 + u_2). \end{aligned}$$

Using relations (6) and (7),

$$\dot{V} \leq \mathcal{S}_1(-r\mathcal{S}_1 - \rho \text{sgn}\mathcal{S}_1) + \mathcal{S}_2(-r\mathcal{S}_2 - \rho \text{sgn}\mathcal{S}_2)$$

By selecting r and ρ values that are greater than zero, \dot{V} will become negative, ensuring that Lyapunov's stability requirement is met.

4 Numerical Simulation and Discussion

In this part of the article, with the aim of verifying and illustrating the usefulness of the intended controller, the MLS chaotic system (2) is investigated after the control law has been applied. The mathematical simulations are performed in MATLAB using ODE 45, and the step size taken is 0.01. The simulation is permitted to run for 15 s. In order to carry out simulations, the initial conditions for both the drive or master system and the response or slave system are taken as (2.0, 4.0, 6.0, 8.0) and (9.0, 7.0, 5.0, 3.0), respectively. Control gains $(k_1, k_2, k_3, k_4) = (15, 15, 15, 15)$. Figure 3a-d illustrates the corresponding states of the master and slave MLS chaotic systems when control signal is not triggered. The time response of synchronization error between drive and response systems before applying controller is illustrated in Fig. 4. It can be seen from the figures that when control signal is not triggered, the states of master and slave are not synchronized, and the error dynamics do not converge to zero. The variation of the corresponding states of master and slave MLS chaotic systems when the control signal is initiated can be seen in Fig. 5a-d, which clearly reveals both the states mimic each other. The variation of synchronization error dynamics has been demonstrated in Fig. 6. The graph illustrates how synchronization errors quickly approach zero.

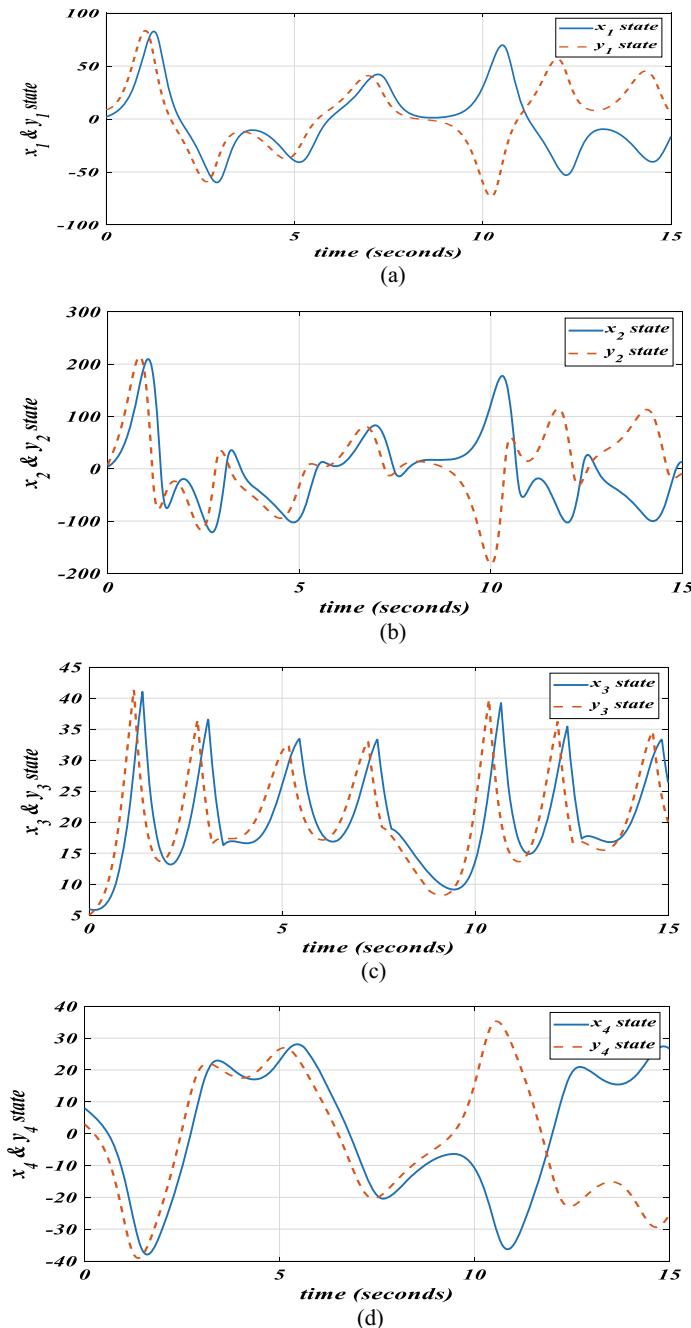


Fig. 3 a-d State response of master (blue) and slave (brown) systems before control signal being triggered

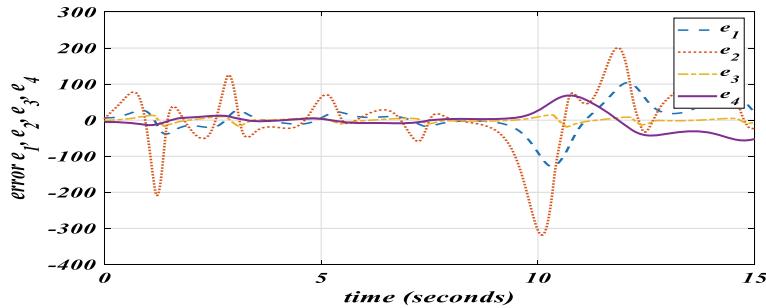


Fig. 4 Variation of synchronization errors before applying controller

5 Conclusion

In this research, an integral sliding surface-based SMC methodology is provided in order to examine the synchronization issue that is associated with the MLS chaotic systems. The SMC approach that has been suggested ensures that the sliding motion will take place. Additionally, it has been demonstrated that the synchronization of the drive or master system and the response or slave systems may be realized by careful selection of the control parameters. It demonstrates that the specified control technique provides the MLS chaotic systems with outstanding transient performance. Synchronization error dynamics take a little more than five seconds, i.e., around 7 s to converge to zero. With the intention of demonstrating the viability of the intended control technique, both a theoretical analysis and numerical simulations have been presented. However, if some other chaotic system is considered, then, the time taken for error dynamics to converge to zero will change, and thus, the time taken to synchronize the states will also change.

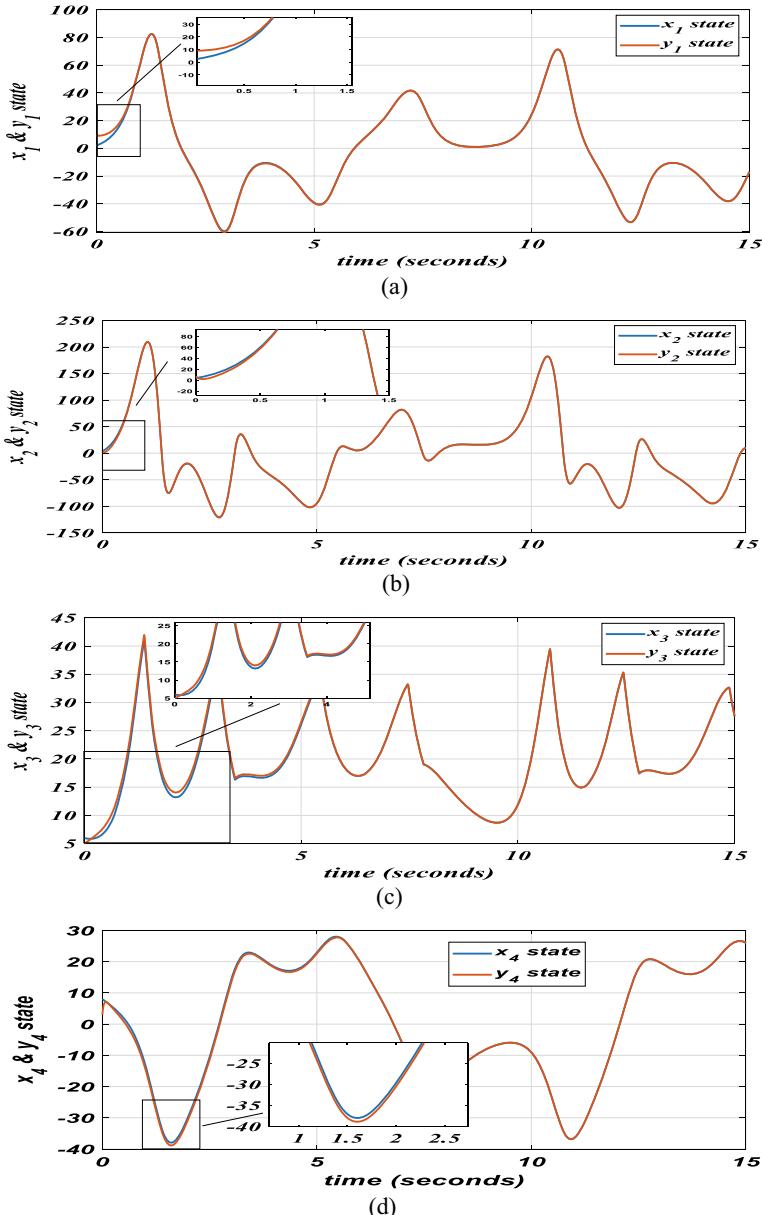


Fig. 5 a-d Corresponding states of master (blue) and slave (brown) systems after control signal has been triggered

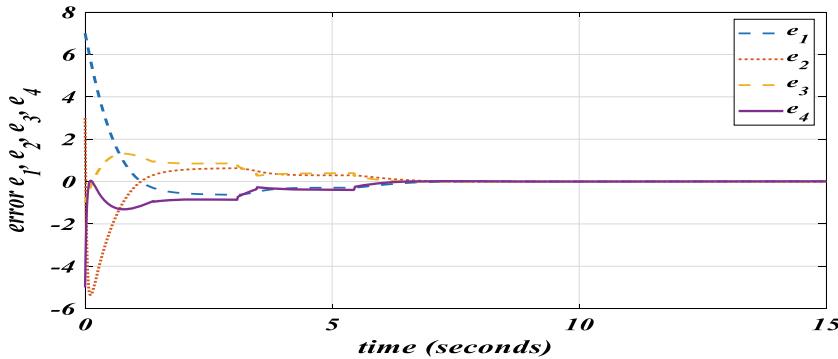


Fig. 6 Synchronization error response between when controller has been applied

References

1. Handa H, Sharma BB (2011) Stabilization and synchronization of MLS chaotic system using PI based Sliding mode control. In: TENCON 2011—2011 IEEE region 10 conference
2. Carroll TL, Pecora LM (1991) Synchronizing chaotic circuits. *IEEE Trans Circuits Syst* 38(4):453–456
3. Chen G, Dong X (1993) On feedback control of chaotic continuous-time systems. *IEEE Trans Circ Syst I Fundam Theory Appl* 40(9):591–601
4. Pecora LM, Carroll TL (1990) Synchronization in chaotic systems. *Phys Rev Lett* 64(8):821–824
5. Pallav and Handa H (2021) Active control synchronization of similar and dissimilar chaotic systems. In: 2021 Innovations in power and advanced computing technologies (i-PACT), pp 1–6. <https://doi.org/10.1109/i-PACT52855.2021.9696832>
6. RM Bora, BB Sharma (2021) Reduced order synchronization of two different chaotic systems using nonlinear active control with or without time delay. In: 2021 International conference on control, automation, power and signal processing (CAPS), pp 1–6. <https://doi.org/10.1109/CAPS52117.2021.9730665>
7. Ranjan RK, Sharma BB, Chauhan Y (2021) Stabilization of a class of chaotic systems with uncertainty using output feedback control methodology. In: 2021 IEEE 6th international conference on computing, communication and automation (ICCCA), pp 533–538. <https://doi.org/10.1109/ICCCA52192.2021.9666409>
8. Sharma BB, Kar IN (2009) Contraction theory based adaptive synchronization of chaotic systems. *Chaos Solitons Fractals* 41(5):2437–2447
9. Pallav and Handa H (2022) Simple synchronization scheme for a class of nonlinear chaotic systems using a single input control. *IETE J Res* 1–14
10. Anand P, Sharma BB (2022) Generalized finite-time synchronization scheme for a class of nonlinear systems using backstepping like control strategy. *Int J Dyn Contr*
11. Singh S, Han S, Lee SM (2021) Adaptive single input sliding mode control for hybrid-synchronization of uncertain hyperchaotic Lu systems. *J Franklin Inst* 358(15):7468–7484
12. Mishra N, Sharma TK, Sharma V, Vimal V (2018) Secure framework for data security in cloud computing. In: Advances in intelligent systems and computing. Springer Singapore, pp 61–71
13. Giri JP, Giri PJ, Chadge R (2018) Neural network-based prediction of productivity parameters. In: Advances in intelligent systems and computing. Springer Singapore, pp 83–95
14. Wang HO, Abed EH (1993) Control of nonlinear phenomena at the inception of voltage collapse. In: 1993 American control conference
15. Stenflo L (1996) Generalized Lorenz equations for acoustic gravity waves in the atmosphere. *Phys Scr* 53:83–84

16. Shan L, Liu Z, Wang Z (2010) A new MLS chaotic system and its backstepping sliding mode synchronization control. *J Comput* 5(3)
17. Slotine J, Li W (1991) Applied nonlinear control. Prentice-Hall, Englewood Cliffs
18. Utkin VI (1993) Sliding mode control design principles and applications to electric drives. *IEEE Trans Ind Electron* 40(1):23–36
19. Vaidyanathan SS, Mamat M, Mohamed MA (2020) Investigation of chaos behavior in a new two-scroll chaotic system with four unstable equilibrium points, its synchronization via four control methods and circuit simulation. *IAENG Int J Appl Math* 50(1):12–21

Performance Analysis of User Behavior Pattern Mining Using Web Log Database for User Identification



Gokulapriya R. and Ganesh Kumar R.

Abstract User behavior analytics is a progressive research domain. Understanding the user's behavior patterns and identifying their behavior patterns will provide solutions to many issues like identity theft and user authentication. So many research works are done in analyzing the frequent access patterns of the users by pre-processing access logs and applying various algorithms to understand the frequent access behavior of the user. From the literature, it finds that the frequent user access pattern identification needs improvement on prediction accuracy and the minimal false positives. To accomplish these, three different approaches were proposed to overcome the existing issues and intended to reduce false positives and improve the frequent pattern mining accuracy based on web access logs. Proposed methods were found to be good while compared with the existing works.

Keywords Mutual pre-processing · Frequent pattern identification · User access pattern · Web access mining

1 Introduction

Web usage mining is the practice of analyzing navigational patterns within web access logs. Access log pattern includes accessing the behaviors of all web users. The web log files include information about each user's web access habits. As discussed in [1], performing analytical study in a big data environment is progressive in nature. The web accessing behaviors provide the number of web pages and the number of times a user visits a website at various intervals. By analyzing behavioral navigation

R. Gokulapriya (✉) · R. Ganesh Kumar

Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST(Deemed to be University), Bangalore, Karnataka, India
e-mail: r.gokulapriya@res.christuniversity.in

R. Ganesh Kumar

e-mail: Ganesh.kumar@christuniversity.in

patterns, the frequently accessed patterns are extracted from the weblog database, and the web user's future access is determined in a significant way. Mining frequent patterns is a difficult task during the analysis of web user behaviors. The lack of identifying frequent behavioral patterns of web users leads to degrading web user identification performance.

The web user requests a search for a specific phrase. The web extracts the requested page after receiving a user request. The user spent a limited time on the retrieved page's homepage. The user also visits a few additional websites, spending varying amounts of time on each. The user logged off after getting search results. Sequentially, the weblog database stores each user's viewed web pages. From stored user data, web user behaviors are analyzed.

The analysis of web user behaviors is conducted by gathering and watching the behaviors of every web user. The web user is identified in a substantial way through the efficient investigation of common web user behaviors. Due to the vast development of web and web users, detection of frequent behavioral patterns of web users is a challenging task. The frequent behavioral patterns of users on the web are efficiently analyzed by using machine learning techniques such as clustering [2] and classification [3] in web usage mining. Various research works [4–6] are designed to study weblog databases to detect frequent behavioral patterns of users. The methods that were developed in the past are ineffective when it comes to recognizing the consistent and recurring behaviors of website visitors.

The architecture of user behavior pattern mining for user identification from a weblog dataset is depicted in Fig. 1. Initially, various numbers of web users activated on the web are stored on the weblog dataset. Through web analytics, the behaviors of the user are examined since it gives information on the number of users to a website and the number of web pages that are browsed by the users. Diverse form of data in the weblog like video, text, images, and audio introduces the complexity to discover relevant patterns for a web user. Thus, the pre-processing of web log files is employed to clean the data by means of removing unwanted data. After that, user behavior on web patterns is analyzed based on the visited web pages. Lastly, frequent web usage patterns are clustered or classified by repeatedly accessed patterns of the same user.

2 Methodology

In our research work, three approaches are proposed and utilized for detecting frequent behavioral patterns of web users to identify the user. Here, weblog datasets are considered as a scenario for detecting frequent user behavior patterns. According to the request sent by the user, the web offers the related required web pages. The user visits the page for a few minutes before proceeding to other web pages provided by the web. The weblog database stores information regarding the viewed web pages at a certain time. The information contained in the weblog database is utilized to observe the behaviors of web users. This is carried out to mining the frequent behavior patterns for finding web user.

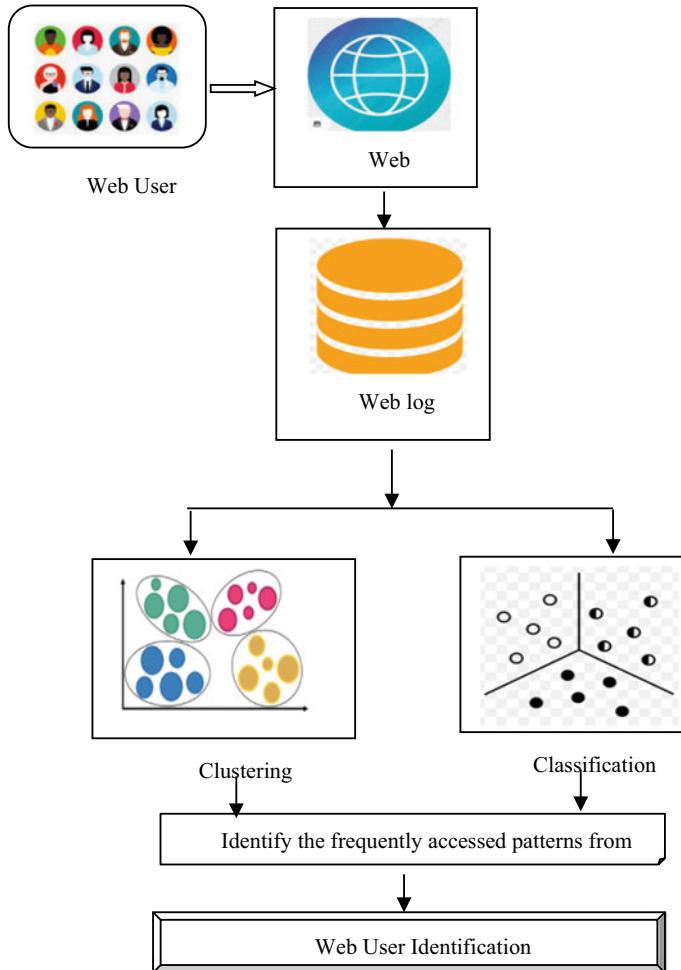


Fig. 1 Architecture of user behavior pattern mining from weblog

Initially, a two-level clustering process (EFK-LBC) is designed in [7] to determine the web user via user behavioral pattern mining. Pre-processing is performed to eliminate unwanted data from input web log files. Primary level cluster is created using Fuzzy k-means, which act as a base learner to detect repeatedly visiting patterns. During the clustering process, a fuzzy membership function is provided to all the web user behavioral patterns which are associated to each cluster centroid. Then, the mean distance of each user behavioral patterns with degree of membership function is considered to fix the cluster centroid. In fuzzy clustering, the distance between the cluster centroid and user behavioral patterns is estimated. Repeated patterns accessed by the same web users are identified as being located close to the centroid. After assigning each web pattern to its own cluster, the cluster's centroid is modified.

This is accomplished by calculating the weighted mean of all web patterns in each cluster. By recalculating the cluster centers, improved clustering results are obtained. In addition, the logit boost clustering technique is used to improve the Fuzzy k-means clustering procedure and reduce fault. In the boosting process, the base learners are combined to produce robust clustering outcomes. With the clustering results, the user ID is correctly detected in the EFK-LBC technique.

In our second approach (NDTS-CFNLC) [8], the classification process helps in finding access patterns that are accessed frequently, with diverse layers. Web user pattern is utilized as input initially. The input layer obtains the number of user patterns. Hidden layer one data processing is applied via Normal Discriminant Pre-processing Analysis (NDPA). Two classes (i.e., relevant, irrelevant) are defined, and for each class, mean value is calculated. Then, the web patterns are mapped to the two classes depending on the mean and variance value with the help of discriminant vector in NDPA. It aids to find whether the relevant and irrelevant patterns are determined. These preprocessed outcomes are provided to the hidden layer two for calculating the similarity of web patterns. Tanimoto similarity is estimated in the NDTS-CFNLC technique recognizing patterns of sites that are frequently visited by the same user during a session. The computed value is given to the output layer in which the classification of web patterns is accomplished. In the output layer, the sigmoid activation function is employed to examine the score value of similarity with the threshold value. If the value of the similarity coefficient is greater than the threshold, the activation function returns the value 1. Otherwise, it returns a value of 0. The value '1' indicates that the user frequently accesses the web patterns, whereas the value '0' indicates that the user rarely accesses the web patterns. As a result, the frequently accessed patterns are detected, thereby identifying the web user with minimal time.

A third proposal was used to increase the precision of pattern mining. The development of the proposed model is aided by Mutual Information-based Pre-processing (MI-P) and Broken-Stick Linear Regression analysis (BLRA). Using Mutual Information-based Pre-processing operation is executed (MI-P). The quantity of web log files is ingested by the web database. The MI-P method eliminates unnecessary and redundant blog information. This, in turn, improves the accuracy and quality of web user behavior pattern mining. In the MI-P method, the mutual information between two web patterns is determined simultaneously. From that, two web patterns are detected as independent when the mutual dependence value is equal to zero. When the mutual dependence value is not equal to zero, two web patterns are dependent and select one web patterns to determine the usage behavior and another one is removed as irrelevant web patterns depending on the computed mutual information value; MI-P method avoids the irrelevant web patterns with minimal item requirements. Thus, the MI-P method enhances the performance of pre-processing to significantly mine the web usage behavior.

After the pre-processing, Broken-Stick Linear Regression Analysis (BLRA) is applied in MIP-BSLR technique. The BLRA is designed to increase the accuracy of web usage behavior mining through partitioning input web patterns into intervals. In the regression analysis, the relationship between the input web patterns is evaluated via BLRA by means of hit ratio. With the help of hit ratio, repeatedly occurred

particular web patterns are determined in weblog files to the total number of web patterns. Depending on this, the input web log files are separated into individual segments (i.e., frequent web patterns or non-frequent web patterns) through BLRA. Then, the linear regression fits with each segment. In BLRA, breakpoint is represented as a threshold value for hit ratio. The optimal breakpoint is determined when BLRA decreases sum of square error. From that, the BLRA considerably finds the frequent web patterns with lower time and maximum accuracy.

3 Experimental Setup

The performance of proposed techniques such as EFK-LBC technique [7], NDTS-CFNLC [8] technique, and MIP-BSLR [9] technique is developed by using Java language. With the help of using two datasets as Apache web log dataset and the NASA dataset, the performance of web user behavior pattern mining is analyzed. Apache web log dataset includes a vast amount of web patterns with user IP address, date, time, method (i.e., HTTP, GET), URL, response code, and bytes. As well as NASA web log dataset comprises web patterns with the attributes such as host, log name, time, method, URL, response, and bytes. From these weblog datasets, web user patterns are considered as input. Then, the proposed methods are applied to identify the frequent web user access patterns. The results of proposed techniques are compared with existing frequent pattern mining-based cross-social network user identification algorithm (FPM-CSNUIA) [10] and Linear-Temporal Logic (LTL)-based model checking technique [11]. The testing parameters of proposed and existing methods are listed as follows:

Pattern mining accuracy, time requirements, and false-positive rate.

3.1 Performance Analysis of Pattern Mining Accuracy

Pattern mining accuracy is the ratio of correctly mined user web patterns to total web patterns. Pattern mining accuracy is a percentage. The method performs better with higher mining accuracy.

The pattern mining accuracy of proposed EFK-LBC technique, NDTS-CFNLC technique, and MIP-BSLR technique using Apache web log dataset is obtained as 84, 88, and 92% when taking number of web patterns as 25. The existing FPM-CSNUIA [12] and LTL-based model checking technique [11] achieve the pattern mining accuracy as 80 and 76% when mining 25 web patterns. Using NASA dataset, 83, 86, and 90% for EFK-LBC technique, NDTS-CFNLC technique, and MIP-BSLR technique are achieved in the first iteration. Also, 79 and 74% of pattern mining accuracy are achieved in existing methods. Graphical results of pattern mining accuracy using Apache web log dataset and NASA dataset are demonstrated in Fig. 2.

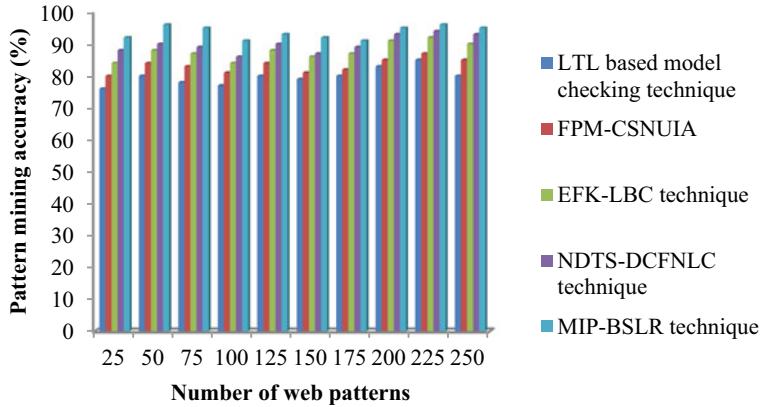


Fig. 2 Measurement of pattern mining accuracy using Apache web log dataset

Figures 2 and 3 demonstrate the experimental outcome of pattern mining accuracy using two datasets. Input patterns in the ranges of 25–250 from Apache web log dataset and 100–1000 from NASA dataset. In Figs. 2 and 3, x-axis shows the inputs, i.e., number of web patterns, and y-axis gives the respective output of pattern mining accuracy. As represented in Figs. 2 and 3, the pattern mining accuracy using proposed and existing methods is gradually improved comparatively.

In contrast to existing works, mutual information-based pre-processing and broken-stick linear regression analysis are carried out in MIP-BSLR technique for increasing the accuracy of pattern mining. By using these processes, irrelevant web patterns are eliminated. In addition, the regression analysis is employed to determine the frequently visited patterns. As a result, the accuracy of web user pattern mining is improved than the other methods. Therefore, the pattern mining accuracy is improved in the proposed MIP-BSLR [7] technique by 13 and 17% to existing FPM-CSNUIA

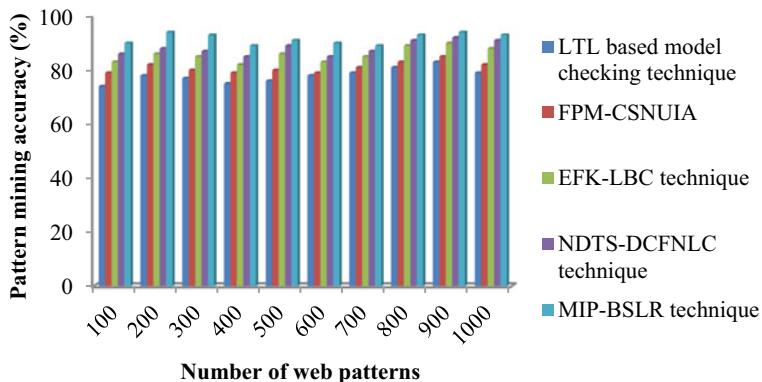


Fig. 3 Measurement of pattern mining accuracy using NASA dataset

[10] and LTL-based model checking technique [11] using dataset Apache web log. Also, the pattern mining accuracy using NASA dataset is enhanced up to 13 and 18% when compared to existing technique correspondingly.

3.2 Performance Analysis of Time Requirements

Time requirements are estimated as the time utilized for obtaining frequent web patterns and thus detecting web user behaviors, which is calculated in the unit of milliseconds (ms). If the time requirements are minimal, then the web user identification performance is more efficient. Ten different iterations are performed in both datasets. Obtained results show that the time requirements of the EFK-LBC technique, NDTS-CFNLC technique, and MIP-BSLR technique are found to be minimum to the other methods.

Proposed MIP-BSLR technique time requirements are provided as 42 ms in the Apache web log dataset for the last iteration (i.e., 250 web patterns) with the difference of 13, 11, 6, and 4 ms for FPM-CSNUIA [10], LTL-based model checking technique [11], EFK-LBC technique [7], and NDTS-CFNLC technique [8], respectively. Similarly, time requirements of NASA dataset using proposed MIP-BSLR technique are obtained as 43 ms in the last run with the difference of 13 ms for FPM-CSNUIA, 12 ms for LTL-based model checking technique, 7 ms for EFK-LBC technique, and 5 ms for NDTS-DCFNL technique, respectively. The results concluded that the proposed MIP-BSLR technique gives lesser time complexity when identifying the frequent web behavior patterns. Based on the above values, the graph is plotted as given in Fig. 4.

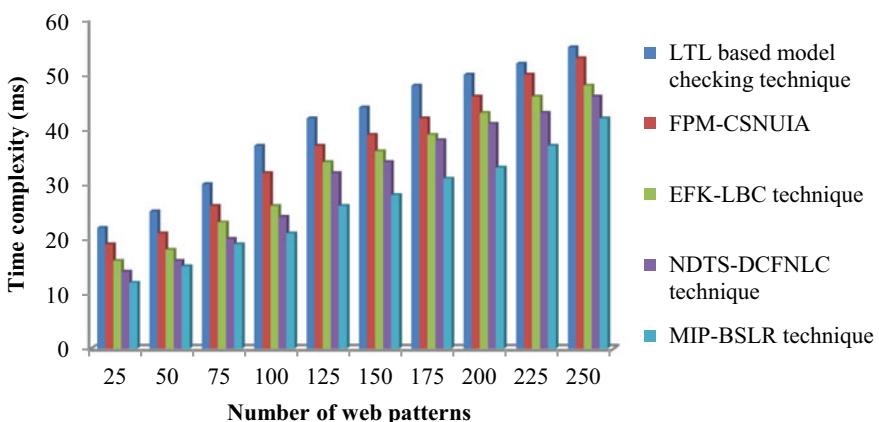


Fig. 4 Measurement of time complexity using Apache web log dataset

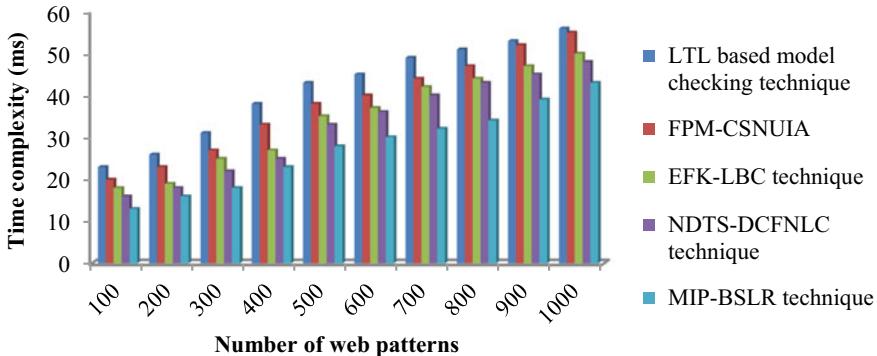


Fig. 5 Measurement of time complexity using NASA dataset

Figures 4 and 5 illustrate the results analysis of time complexity for five different methods according to the different number of web patterns in the ranges of 250–1000 from the datasets. By using these two datasets, the simulation is performed to verify the success of the proposed techniques in terms of time complexity. As given in Figs. 4 and 5, the five methods need inferior time to recognize the web user behaviors depending on the number of web patterns. But, the proposed MIP-BSLR technique effectively minimizes the time requirements.

The minimal time utilization is achieved with the help of using Mutual Information-based Pre-processing and Broken-Stick Linear Regression Analysis in MIP-BSLR. Initially, the mutual dependence between patterns is measured to identify the dependent and independent web patterns. In turn, more related web patterns are chosen and others are removed to decrease the time complexity. Then, the regression analysis is employed to detect the frequently accessed web patterns and thus determines the behaviors of web users. Through the regression analysis, the time requirements are decreased. Hence, the MIP-BSLR technique resulted in 29 and 36% when compared to existing FPM-CSNUIA [10] and existing LTL-based model checking technique [11], respectively, with the time requirements metrics. In addition, the requirements of time in the NASA dataset are curtailed by 28 and 35%.

3.3 False-Positive Analysis

The experimental outcome of the false-positive rate for five methods with two datasets is shown here. The ratio of user-accessed patterns mistakenly mined as frequent to total web pattern input is representing the rate of false positives, which is measured in percentile. A minimal false-positive rate indicates improved web user behavior

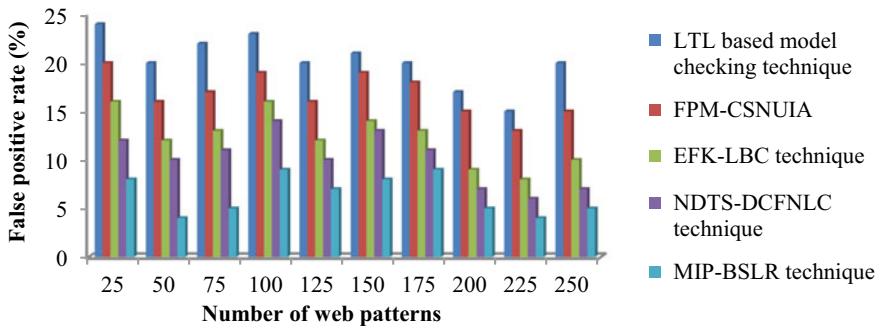


Fig. 6 Measurement of false-positive rate using Apache web log dataset

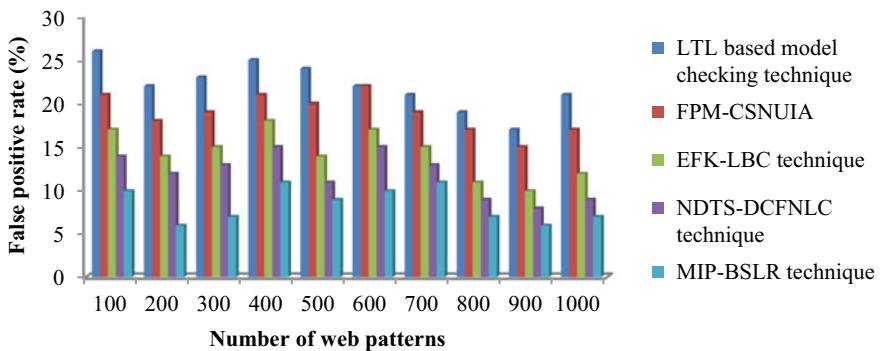


Fig. 7 Measurement of false-positive rate using NASA dataset

identification. Proposed EFK-LBC technique [7], proposed NDTS-DCFNLNC technique [8], and proposed MIP-BSLR technique [9] are compared with existing FPM-CSNUIA [10] and existing LTL-based model checking technique [11]. False positives of the proposed approach are obtained as 10, 7, and 5% for the Apache web log dataset. Besides, the NASA dataset is provided as 17 and 21% for existing FPM-CSNUIA [10] and existing LTL-based model checking techniques [11] by processing 1000 web patterns. Minimal false positives were achieved in the proposed methods. A graphical representation of the false-positive rate is depicted in Figs. 6 and 7.

It is clearly observed that the false-positive rate of the proposed methods is decreased based on the number of web patterns.

4 Conclusion

An efficient technique such as the EFK-LBC technique [7], NDTS-DCFNLNC technique [8], and MIP-BSLR technique [9] is designed to determine the user by mining

frequent user behavior on the web with higher accuracy and minimal complexity. At EFK-LBC technique, two-level clustering process using Fuzzy k-means clustering and logit boost clustering is applied to lessen the error rate and time requirements of behavioral pattern mining. After that, the NDTS-DCFNLNC technique is implemented using similarity measure and classification process. Discriminant pre-processing analysis is employed to eliminate the irrelevant patterns. Besides, similarity between the web patterns is detected to find the web user with higher accuracy. Activation function is also used to effectively classify the patterns with lower error. Lastly, MIP-BSLR technique is proposed to improve behavior pattern mining on web usage with less complexity. With the help of measuring mutual dependence between web patterns using MI-P, relevant patterns are selected. Association between patterns was found by implementing regression analysis for obtaining the frequent web patterns. In turn, the accuracy of pattern mining is improved in the MIP-BSLR technique with a minimum error rate. The experimental results show that the proposed MIP-BSLR technique gives better performance for user identification with lower time, error, and higher accuracy.

References

1. Aldubai AF, Humbe VT, Chowhan SS (2017) Analytical study of intruder detection system in big data environment. In: *Soft computing: theories and applications: proceedings of SoCTA 2016*, vol 1, pp 405–583
2. Anandhi D, Irfan Ahmed MS (2017) Prediction of user's type and navigation pattern using clustering and classification algorithms. *Cluster Comput* 1–10. Springer
3. Adeniyi DA, Wei Z, Yongquan Y (2016) Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Appl Comput Inf* 12(1):90–108. Elsevier
4. Xing L, Deng K, Wu H, Xie P, Gao J (2019) Behavioral habits-based user identification across social networks. *Symmetry* J 11(9):1–19
5. Deng K, Xing L, Zheng L, Wu H, Xie P, Gao F (2019) A user identification algorithm based on user behavior analysis in social networks. *Appl Big Data Soc Sci, IEEE Access* 7:47114–47123
6. Ya J, Liu T, Li Q, Shi J, Zhang H, Lv P, Guo L (2017) Mining host behavior patterns from massive network and security logs. *Procedia Comput Sci* 108:38–47. Elsevier
7. Gokulapriya R, Ganesh Kumar R (2018) Behavioral pattern mining for user identity and access control a cluster-based ensemble model. *Int J Eng Technol* 7(3):438–444
8. Gokulapriya R, Kumar RG (2021) Normal discriminant deep convolution neural classification-based web behavioral pattern mining for user identification. *J Ambient Intell Humaniz Comput* 12(6):6689–6699
9. Raman G, Raj GK (2021) Mutual information pre-processing based broken-stick linear regression technique for web user behaviour pattern mining. *Int J Intell Eng Syst (IJIES)* 14(1)
10. Mishra R, Choubey A (2012) Comparative analysis of Apriori algorithm and Frequent Pattern algorithm for frequent pattern mining in web log data. *Int J Comput Sci Inf Technol* 3(4):4662–4665
11. Hernández S, Álvarez P, Fabra J, Ezpeleta J (2017) Analysis of users' behaviour in structured e-commerce websites. *IEEE Access* 5:11941–11958
12. Wu IC, Yu HK (2020) Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states. *Inf Process Manag* 57(6):1–18. Elsevier

An Imperfect Production System for Non-instantaneous Deteriorating Goods with Preservation Technology Under Cap-and-Trade Policy



Pankaj Narang , Mamta Kumari , and Pijus Kanti De

Abstract An imperfect economic production quantity model is formulated with non-instantaneous deteriorating goods. To lower the impact of deterioration, preservation technology is applied. This suggested model is analyzed under green investment and cap-and-trade policy to lower the carbon emissions created by the manufacturing firm. The defective items made during the production run are discarded. The demand of the product is affected by selling price, green investment, and marketing. The objective of this article is to deduce the optimal production rate and selling price of the product with maximizing the profit. A numerical example has been employed to demonstrate this model. To evaluate the stability of our model, sensitivity analysis with respect to key parameters is carried out. Furthermore, the article is concluded with suggestions for possible future directions.

Keywords EPQ model · Non-instantaneous deterioration · Preservation technology investment · Carbon emission

1 Introduction

Environmental sustainability has grown in importance over the last few decades. The primary cause of global warming is accelerated industrialization, which produces significant carbon emissions. Carbon emissions are now a major concern in the manufacturing sector due to the rise in environmental concerns. The cap-and-trade program is the most important policy among the several on the market to lower emissions. In the cap-and-trade system, the government first provides free emission credits to the company. Then, they might buy or sell in the market for carbon trading. Bai et al. [1] explored a supply chain model with one manufacturer and two competitive retailers for degrading goods under the effect of carbon cap-and-trade laws and green investment. By comparing the impacts of cap-and-trade and low-carbon subsidy schemes on a manufacturing model, Cao et al. [2] determined which was

P. Narang  · M. Kumari · P. K. De

Department of Mathematics, National Institute of Technology Silchar, Silchar, Assam 788010, India

e-mail: pankajnarang830@gmail.com

better for society. A closed-loop supply chain system was formulated by Jauhari et al. [3] that included carbon emissions under the cap-and-trade system account and rework, refurbishing, and waste disposal. For unreliable industrial systems, Entezamnia et al. [4] developed a combined production and carbon trading approach that took into account stochastic and dynamic contexts under the cap-and-trade policy. Dua et al. [5] formulated a production inventory model where the rate of production is dependent on demand and reliability under the effects of carbon emission.

Deterioration is a common aspect of the products like food items, volatile liquids, and radioactive materials. In general, it is discovered that things inevitably lose quality over time, although deterioration can be reduced by using the proper preservation techniques. For instance, keeping fish in a deep refrigerator or utilizing ice can slow down the rate of degradation. Rahaman et al. [6] developed a deteriorated EPQ model in which preservation technology was implemented to restore the significant loss of products during production. Bhunia et al. [7] used tournament genetic algorithm for solving an EPQ model where production rate was the decision variable. Dye [8] evaluated the impact of preservation technology investment on inventory decisions with non-deteriorating products. Pervin et al. [9] formulated an EPQ model with deteriorating goods under the effect of preservation technology where demand function depends on price and stock level. Bardhan et al. [10] formulated an inventory system with demand relying on stock and non-instantaneous deterioration items under the impact of preservation technology.

The classical EPQ model is based on several assumptions, such as the manufacturing system produces only perfect quality items, and the demand rate is constant. However, in reality, these assumptions are rarely satisfied. Over the years, both scholars and managerial decision-makers have focused much attention on market demand. To represent scenarios in the actual world, many demand patterns for various types of products have been investigated and assumed in the literature. Specific parameters that affect the demand rate like selling price, advertisement, quality, and stock level. Khara et al. [11] formulated an imperfect production system where demand relies on the selling price and reliability of the product. Singh [12] created a production inventory model for goods that degrade where the product's demand is dependent on supply, selling price, and time. An integrated imperfect production model that includes a manufacturer and retailer was developed by Khara et al. [13], where demand is dependent on the advertisement. Saxena et al. [14] considered a basic EPQ system with basic setup and item stewardship and investigated lucidly a complete and simple elucidation of item stewardship.

In this article, the rate of demand depends on selling price, marketing, and green investment where these factors have positive as well as negative impacts on buyers, and the production is imperfect where some percentage of defective goods are produced. These defective goods are disposed of at a cost. Carbon cap-and-trade scheme and green investments are implemented to reduce the emissions created by the manufacturing system. The preservation technology for instantaneous deteriorating items is assumed. A numerical illustration and sensitivity analysis are provided to illustrate the effect of parameters on the profit function. In the end, managerial implications and conclusion are provided.

The following is an overview of this article's significant contributions:

1. An imperfect EPQ model for non-instantaneous deteriorating goods with preservation technology under the effect of green investment and carbon cap-and-trade policy is formulated in this study.
2. The production rate and selling price are taken as decision variables.
3. The total average profit of the system is computed.

2 Notations and Assumptions

2.1 Notations

Table 1 presents the notations that have been used in this model.

Table 1 Notations for this model

Notations	Units	Description
C_0	\$/unit	Setup cost per cycle
C_h	\$/unit	Holding cost per unit time
C_p	\$/unit	Production cost per unit time
C_A	\$/frequency	Advertisement cost
C_d	\$/unit	Deterioration cost per unit time
C_{pr}	\$/unit	Preservation cost per unit time
C_f	\$/unit	The trading price of carbon emissions permits
C	Unit	Carbon cap
C_z	\$/unit	Disposal cost per unit time
μ	Unit	Carbon emissions per unit
ω	Unit	Parameter of green technology effect on lowering carbon emissions
A	Unit	Advertisement frequency
P	Unit	Rate of production per cycle
D	Unit	Rate of demand per cycle
ψ	Unit	Percentage of defective items
s	\$/unit	The selling price of the product
g	Unit	Green technology level
ρ	\$	Investment by the manufacturer for adopting green technology
t_1	Month	Length of the time in which item has no deterioration
t_2	Month	Production time
t_3	Month	Total cycle length

2.2 Assumptions

1. The demand function is considered as a function of green investment, selling price, and marketing cost and is defined as $D = a - bs + \kappa g + \lambda A$ where a is the market potential, b , κ , and λ are the sensitivity factors of the selling price, green technology, and frequency of advertisement, respectively.
2. This model is for a single product.
3. The time horizon is limited.
4. The starting and terminal stock level is Nil.
5. The production system is imperfect. After some time, it starts producing defective items.
6. The non-instantaneous deterioration of the product is considered. Also, preservation technology is implemented in the system to decrease the effect of deterioration. The rate of deterioration (θ) and the preservation rate (ϕ) are constants satisfying the relation $\theta + \phi = 1$.
7. The rate of production is always higher than or equal to the sum of the rate of demand and deterioration.
8. The production rate is constant.

3 Model Formulation

The manufacturer begins its production run at time $t = 0$ with production rate P . During time interval $[0, t_1]$, the stock level reduces because of demand rate D . During time interval $[t_1, t_2]$, the system starts manufacturing some defective items due to some unavoidable reasons, and the stock level decreases because of the combined effect of demand, deterioration, and defective items. Additionally, the growth of the inventory level is positively impacted by preservation technology. After time t_2 , the stock level gradually falls because of the customer's demand and deterioration, and at time t_3 , the stock level reduces to zero (Fig. 1).

The following differential equation can represent the changes in stock with time t :

$$\frac{dI}{dt} = P - D ; \quad 0 \leq t \leq t_1 \quad (1)$$

$$\frac{dI}{dt} = P - \psi P - (\theta - \phi)I(t); \quad t_1 \leq t \leq t_2 \quad (2)$$

$$\frac{dI}{dt} = -D - (\theta - \phi)I(t); \quad t_2 \leq t \leq t_3 \quad (3)$$

With the conditions $I(0) = 0$, $I(t_3) = 0$.

By using initial conditions and solving Eqs. (1), (2), and, (3) the resulting equations are as follows:

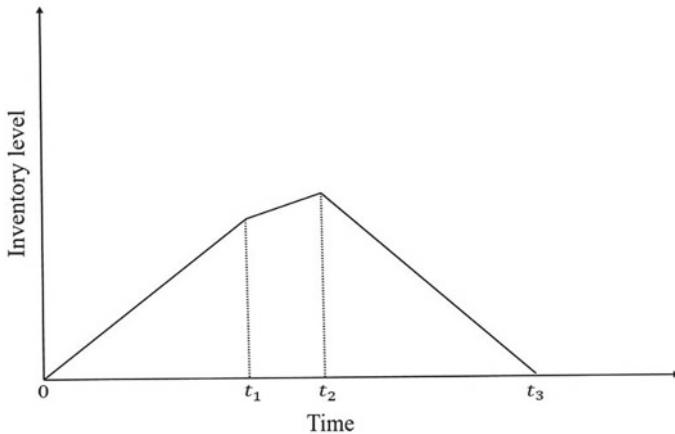


Fig. 1 Graphical representation of inventory model

$$I_1(t) = (P - D)t ; \quad 0 \leq t \leq t_1 \quad (4)$$

$$I_2(t) = \frac{-D + P - P\psi + e^{(\theta+\phi)(-t+t_1)}(D + P(-1 + \psi) + (-D + P)t_1)}{\theta + \phi}; \quad t_1 \leq t \leq t_2 \quad (5)$$

$$I_3(t) = \frac{D(-1 + e^{(\theta+\phi)(-t+t_3)})}{\theta + \phi}; \quad t_2 \leq t \leq t_3 \quad (6)$$

The various costs associated with the proposed model are production cost, setup cost, cost of holding, cost of deterioration, disposal cost, advertisement cost, preservation technology cost, and investment in green technology.

All of these costs are calculated as follows:

1. Production cost (PC)

$$= C_p \left(\int_0^{t_2} P \, dt \right) = C_p P t_2 \quad (7)$$

2. Setup cost (SC)

$$= C_0 \quad (8)$$

3. Holding cost (HC)

$$= C_h \left(\int_0^{t_1} I_1(t) \, dt + \int_{t_1}^{t_2} I_2(t) \, dt + \int_{t_2}^{t_3} I_3(t) \, dt \right) \quad (9)$$

4. Deterioration cost (DC)

$$= C_d \left(\int_0^{t_1} \theta I_1(t) dt + \int_{t_1}^{t_2} \theta I_2(t) dt + \int_{t_2}^{t_3} \theta I_3(t) dt \right) \quad (10)$$

5. Disposal cost (DSC)

$$= C_z \left(\int_{t_1}^{t_2} \varphi P dt \right) \quad (11)$$

6. Preservation technology cost (PTC)

$$= C_{pr} \left(\int_0^{t_1} \phi I_1(t) dt + \int_{t_1}^{t_2} \phi I_2(t) dt + \int_{t_2}^{t_3} \phi I_3(t) dt \right) \quad (12)$$

7. Investment in green technology (GT)

$$= \frac{1}{2} \rho g^2 \quad (13)$$

8. Advertisement cost (AC)

$$= C_A A \quad (14)$$

The emissions from the manufacturing process are determined by multiplying the emissions generated per unit item with the total number of items manufactured.

$$E_m = (\mu - \omega g) P t_2 \quad (15)$$

The total quantity of emissions is restricted by the carbon cap allotted to the manufacturer. The manufacturer must purchase carbon from another company to cover the difference when overall emissions exceed the carbon cap. However, if E_m is lower than C , the manufacturer can sell the extra and make more money. The manufacturer's cost (profit) related to the carbon cap-and-trade policy is described as follows:

$$CC = C_f (E_m - C) \quad (16)$$

Now, the sales revenue is obtained by multiplying the selling price per unit by the demand for the product.

Sales revenue (SR)

$$= s \left(\int_0^{t_3} Dt dt \right) = \frac{1}{2} D s t_3^2 \quad (17)$$

Total average profit =

$$\begin{aligned}
 & \text{SR} - (\text{PC} + \text{SC} + \text{HC} + \text{DC} + \text{DSC} + \text{PTC} + \text{GT} + \text{AC} + \text{CC}) \\
 \text{TAP} = & \frac{1}{2} D s t_3^2 - \left(C_h \left(\frac{1}{2} (-D + P) t_1^2 \right. \right. \\
 & \left. \left. + \frac{Dt_1 - Pt_1 + P\psi t_1 - \frac{(-1 + e^{(\theta+\phi)(t_1-t_2)})(D+P(-1+\psi)+(-D+P)t_1)}{\theta+\phi} - Dt_2 + Pt_2 - P\psi t_2}{\theta+\phi} \right. \right. \\
 & \left. \left. + \frac{D(-1 + e^{-(\theta+\phi)(t_2-t_3)} + (\theta + \phi)t_2 - (\theta + \phi)t_3)}{(\theta + \phi)^2} \right) + C_0 + C_p Pt_2 \right. \\
 & \left. + C_d \left(\frac{\theta \left(Dt_1 - Pt_1 + P\psi t_1 - \frac{(-1 + e^{(\theta+\phi)(t_1-t_2)})(D+P(-1+\psi)+(-D+P)t_1)}{\theta+\phi} \right. \right. \right. \\
 & \left. \left. \left. + D(\theta + \phi)(t_1 - t_2) - Dt_2 + Pt_2 - P\psi t_2}{\theta+\phi} \right) \right. \right. \\
 & \left. \left. - \frac{D\theta \left(\frac{1-e^{-(\theta-\phi)(t_2-t_3)}}{\theta+\phi} - (1 + \theta + \phi)t_2 + (1 + \theta + \phi)t_3 \right)}{\theta+\phi} \right) \right. \\
 & \left. + \phi C_{\text{Pr}} \left(Dt_1 - Pt_1 + P\psi t_1 - \frac{(-1 + e^{(\theta+\phi)(t_1-t_2)})(D+P(-1+\psi)+(-D+P)t_1)}{\theta+\phi} \right. \right. \\
 & \left. \left. + D(\theta + \phi)(t_1 - t_2) - Dt_2}{\theta+\phi} \right) \right. \\
 & \left. + \phi C_{\text{Pr}} \left(Pt_2 - P\psi t_2 - D \left(\frac{1-e^{-(\theta+\phi)(t_2-t_3)}}{\theta+\phi} - (1 + \theta + \phi)t_2 + (1 + \theta + \phi)t_3 \right) \right) \right. \\
 & \left. + P\psi C_z \left(-\frac{t_1^2}{2} + \frac{t_2^2}{2} \right) + C_A A + C_f ((\mu - \omega g) Pt_2 - C) + \frac{g^2 \rho}{2} \right)
 \end{aligned}$$

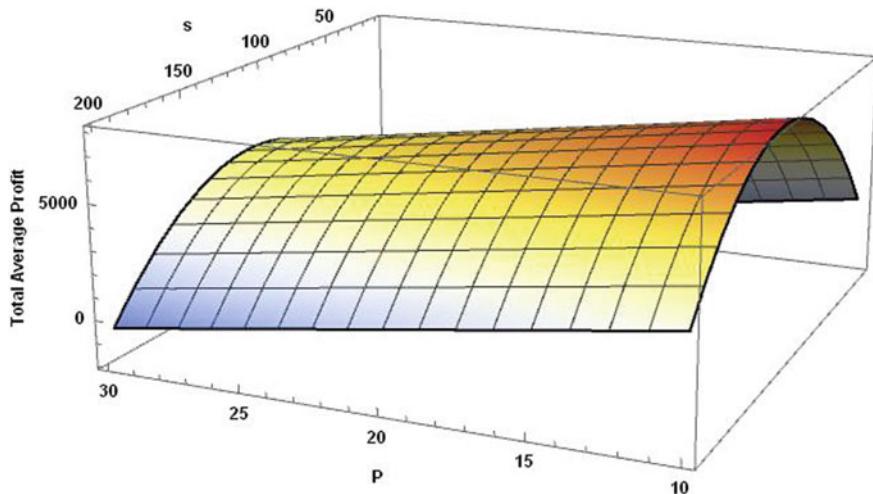


Fig. 2 Total average profit with respect to the production rate and selling price

4 Numerical Illustration

The following parameter values are assumed:

$$\begin{aligned} \mu &= 40; \omega = 0.42; C_h = 3.75; C_0 = 500; C_d = 1.2; \lambda = 0.5; \\ \kappa &= 0.6; b = 0.5; \psi = 0.4; \theta = 0.15; C_z = 8; C_a = 40; C = 2000; \\ C_f &= 2.5; \rho = 25; a = 100; t_1 = 1.2; t_2 = 1.25; t_3 = 1.6; \phi = 0.85 \end{aligned}$$

in appropriate units.

The optimal values of the decision variables obtained by using the graphical method are as follows:

$$P = 10, s = 111.786, \text{ TAP} = 8068.59$$

The concavity of the total average profit has been shown in Fig. 2.

5 Sensitivity Analysis

A detailed sensitivity analysis has been performed to analyze the impact of changes in the numerical values taken in the example.

5.1 Observations

1. From Figs. 3, 4, and 5, it can be concluded that the total average profit of the manufacturer decreases when any cost related to the system increases. This shows that this proposed model satisfies real-life phenomena.
2. In the case of carbon trading price (C_f), Fig. 6 demonstrates that the total average profit increases when the value of C_f increases. This shows that carbon cost acts as profit because the carbon emissions generated in the manufacturing unit are less than the carbon cap.
3. Figure 7 shows that when the initial demand parameter rises, the total average profit increases.
4. Figure 8 represents that the total average profit reduces when the green investment increases.
5. The total average profit reduces when the emissions per unit increase, as shown in Fig. 9. This indicates that the carbon function acts as a cost when the emissions increase.

Fig. 3 Total average profit versus advertisement cost

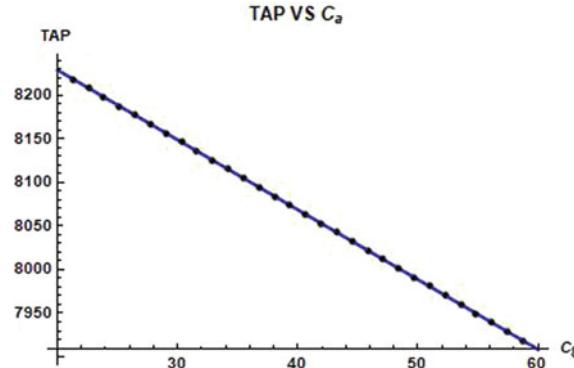


Fig. 4 Total average profit versus disposal cost

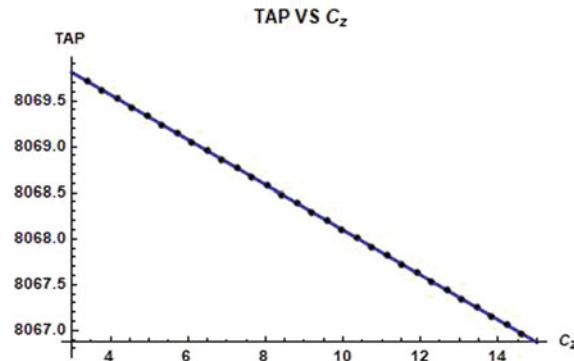


Fig. 5 Total average profit versus setup cost

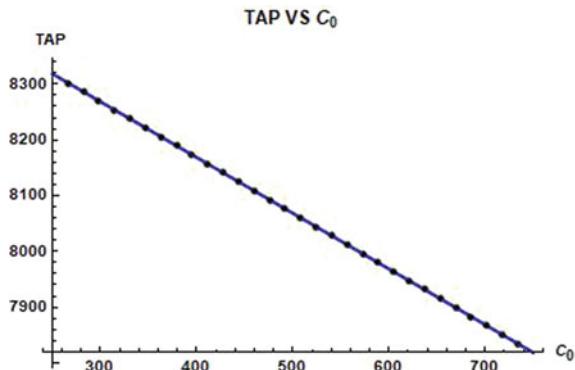


Fig. 6 Total average profit versus trading price

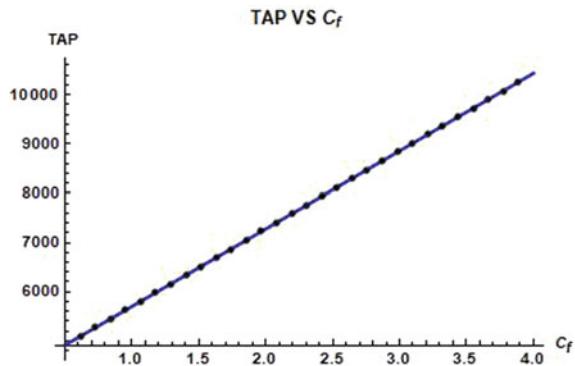
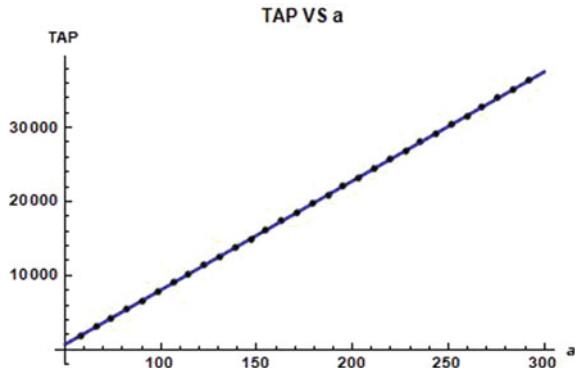


Fig. 7 Total average profit versus market demand



6. In the case of defective items, Fig. 10 demonstrates that when the percentage of defective items increases, the total average profit decreases.

Fig. 8 Total average profit versus green investment

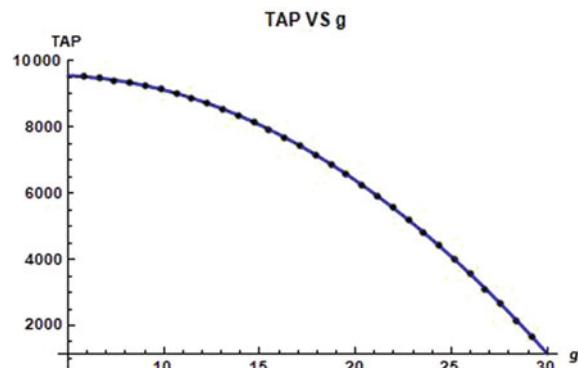


Fig. 9 Total average profit versus emissions

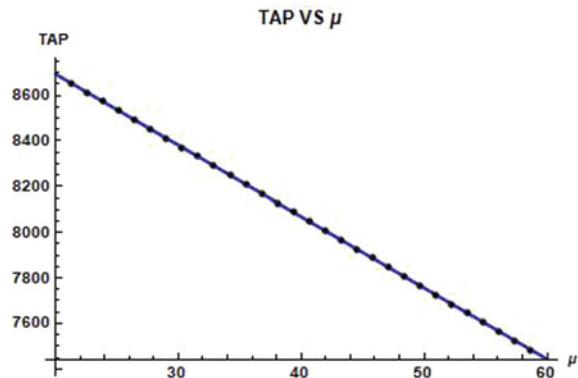
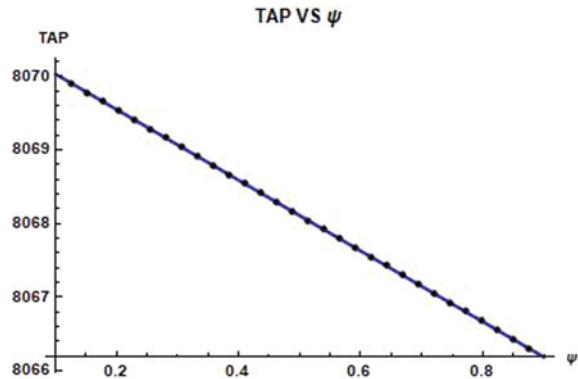


Fig. 10 Total average profit versus defective percentage



6 Conclusion

This article presents an imperfect EPQ model for non-instantaneous deteriorating goods under the cap-and-trade scheme and green investment with preservation technology. The product's demand depends on the selling price, advertisement, and green technology level. Non-instantaneous deterioration items have been considered as many products become partially unstable after some time. The machine produces some defective items after some time due to unavoidable reasons. The carbon cap-and-trade scheme is included to restrict carbon emissions. The preservation technology reduces the effect of deterioration. The novelty of this article is that preservation technology in an imperfect production with carbon emissions has been considered. A profit function has been developed by considering all those factors. This profit function is optimized by using the graphical method. This study presents insightful results on determining the optimal production rate and selling price. In addition, sensitivity analysis has been performed to demonstrate the impact of different factors on total average profit.

This model has several drawbacks that need to be examined in future research. First, the reworking of defective items has not been considered. Second, this model is for a single product. Third, the production rate is considered constant.

This model can be extended by including the reworking of defective items, carbon tax policy, trade credit policy, variable production rate, etc.

References

1. Bai Q, Jin M, Xu X (2019) Effects of carbon emission reduction on supply chain coordination with vendor-managed deteriorating product inventory. *Int J Prod Econ* 208:83–99
2. Cao K, Xu X, Wu Q, Zhang Q (2017) Optimal production and carbon emission reduction level under cap-and-trade and low carbon subsidy policies. *J Clean Prod* 167:505–513
3. Jauhari WA, Adam NAFP, Rosyidi CN, Pujawan IN, Shah NH (2020) A closed-loop supply chain model with rework, waste disposal, and carbon emissions. *Oper Res Perspect* 7:100155
4. Entezaminia A, Gharbi A, Ouhimmou M (2021) A joint production and carbon trading policy for unreliable manufacturing systems under cap-and-trade regulation. *J Clean Prod* 293:125973
5. Dua S, Mogha SK, Dem H (2022) A production inventory system under impact of carbon emission and reliability. In: *Soft computing: theories and applications*. Springer, pp 191–201
6. Rahaman M, Mondal SP, Alam S, De SK (2022) A study of a lock fuzzy EPQ model with deterioration and stock and unit selling price-dependent demand using preservation technology. *Soft Comput* 26(6):2721–2740
7. Bhunia A, Kundu S, Sannigrahi T, Goyal S (2009) An application of tournament genetic algorithm in a marketing oriented economic production lot-size model for deteriorating items. *Int J Prod Econ* 119(1):112–121
8. Dye C-Y (2013) The effect of preservation technology investment on a non-instantaneous deteriorating inventory model. *Omega* 41(5):872–880
9. Pervin M, Roy SK, Weber GW (2020) Deteriorating inventory with preservation technology under price-and-stock-sensitive demand. *J Ind Manag Optim* 16(4):1585
10. Bardhan S, Pal H, Giri BC (2019) Optimal replenishment policy and preservation technology investment for a non-instantaneous deteriorating item with stock-dependent demand. *Oper Res Int J* 19(2):347–368

11. Khara B, Dey JK, Mondal SK (2017) An inventory model under development cost-dependent imperfect production and reliability-dependent demand. *J Manag Anal* 4(3):258–275
12. Singh D (2019) Production inventory model of deteriorating items with holding cost, stock, and selling price with backlog. *Int J Math Oper Res* 14(2):290–305
13. Khara B, Dey JK, Mondal SK (2021) An integrated imperfect production system with advertisement dependent demand using branch and bound technique. *Flex Serv Manuf J* 33(2):508–546
14. Saxena P, Singh C, Sharma K (2018) EPQ model with product stewardship approach. In: *Soft computing: theories and applications*: Springer, pp 107–113

Human Activity Recognition Using a Hybrid Dilated CNN and GRU



Preeti Gupta and Satish Chand

Abstract Human activity recognition is believed to be experiencing rapid growth in the area of computer vision. With advancements in video capturing technologies, more data is being analysed. Study of various activities performed, makes it simple to comprehend the context of the video files, and as a result, various actions may be taken. In this paper, we intend to recognise various human activities by using a hybrid model of dilated convolution neural network (CNN) and recurrent neural network (RNN). We validate its performance on a publically available dataset—HMDB51—and compare its performance with that of the existing models. The model proposed gives an accuracy of 84%.

Keywords Activity recognition · Dilation · GRU · RNN · CNN

1 Introduction

Human activity recognition is about automated understanding of the actions either in video samples or sensor samples which consists of certain activities like human–human interaction, human-object interaction, group activities, certain gestures, actions, etc. The intention is to find out the labels consisting of the actions in order to understand the activities in a better way, by means of an automated system.

Despite considerable advancements in the study of automated activity analysis of humans, recognising human interactions in video is still a challenging problem [1]. Understanding interactions between people takes more than just examining each person's behaviours separately, which is a key component of the challenge. Instead, the coordination of people in both space and time reveals the underlying character of their group activity.

Due to technological advancements, there are millions of video files available in both public and private domains on sites and services like Youtube, Social Media, and other streaming platforms, from which one can easily obtain a variety of data.

P. Gupta (✉) · S. Chand
Jawaharlal Nehru University, New Delhi 67, India
e-mail: preeti62_scs@jnu.ac.in

Yet, these datasets face few challenges like large variation in appearance, occlusions, non-rigid motion, and viewpoint changes, clothing, rare occurrences of few classes like stealing, few action vocabulary which are not well defined. Attempt of solving one such challenge which is occlusion is done in [2]. In today's world, automatic analysis of video activities is required for various applications such as surveillance, automated driving technology, elderly health monitoring, and activity prediction to avoid any mishappenings.

Researchers have been experimenting with and studying various automation approaches for many years; but, with the current progress in deep learning techniques, it is now a classical task. Deep learning is utilised for a variety of difficult tasks, including genre categorization, context analyses of video scenes, and sentiment analysis utilising both text and voice. A set of actions is taken into account whilst classifying a certain activity. The challenges of these datasets include the context and setting in which the acts occurred, as well as infrequent occurrences of particular activities. Our goal is to build a model that can accurately classify the activities in the dataset, such as eat, kick, jump, and climb, using just visual input.

Various action recognition algorithms have recently been developed, and they can be categorised into two: deep learning approaches and manually crafted feature-based methods. The static and motion data in videos was extracted using hand crafted features. For instance, gradients and motion data were extracted using the histogram of oriented gradient and optical flow, HOG, and HOF, respectively. A technique known as the improved dense trajectory (IDT) [3] was created in order to lessen the impact of camera motion. Hand-crafted features and feature encoding techniques like bag-of-words (BOW) were frequently coupled to obtain equivalent performance in action recognition. To simulate various actions, it was challenging to select optimum features and encoding techniques. CNNs gave better performance in image processing such as in [4, 5]. CNNs and RNNs are frequently used in deep learning techniques to learn sophisticated video representation. CNNs showed a strong capacity for learning static visual representation, but struggled to recognise activity in dynamic settings. One of the key causes was that videos include abundant spatial and temporal data, and standard CNNs were unable to extract temporal data.

Convolutional neural networks (CNNs) were used to classify action recognition in video, whilst recurrent neural networks (RNN) were utilised to find the relationship amongst frames in the temporal domain.

Therefore, we have proposed a novel hybrid model with the combination of CNN and RNN. CNNs for learning features in spatial domain and RNNs to find out the relation between frames in the temporal domain which is used to recognise human activity and is evaluated on the publicly available dataset HMDB51 consisting of a wide variety of action labels.

The organisation of this paper is as follows: Section 2 summarises the related works, in brief, Sect. 3, proposed methodology and Sect. 4 presents the result, and Sect. 5 presents the conclusion of the paper.

2 Related Work

Human activity recognition has been explored using several methods in recent years. Prior research mostly employed manually crafted features to explain spatial and motion data employing a variety of local features. Local features, which represent the image using feature descriptors like scale-invariant feature transform (SIFT) [6] and speeded up robust features (SURF) [7] are an efficient tool for image recognition. Image recognition tasks were performing well, so it inspired the researcher's attention towards action recognition tasks focussed on spatiotemporal information. In [8], author further modified HOG descriptor into HOG3D model which was for spatiotemporal gradients for the activity recognition tasks in video data. Just like this, extended methods were produced of some well-known methods like Harris3D [9] is a extended version of Harris corner descriptor [10], and SIFT-3D [11] is of SIFT. A further combination of HOG, motion boundary histogram (MBH), and HOF was discussed which was named as dense trajectories [12]. IDTs features [3] being its enhanced version which takes into account camera motion as well. In [13], author used a combination of IDT and subtensor projection to understand the human actions. It was noticeable that IDT features extraction was computationally complex task and difficult for huge amount of data. Furthermore, encoding methods adoption came into play where bag of visual words (BoVW) [14], VLAD, or FV, local attributes were integrated into global level feature vector to accomplish task of action recognition. These produced better results in high order local features yielding greater accuracy as compare to BoVW. But in this encoding process, temporal information of local features was lost. Meanwhile, conditional random fields [15] were introduced and performed well in extracting long term temporal features to execute the task of action recognition. Drawbacks of these methods remain which are the time-consumption and more manual labour.

Effectiveness in image classification serves as inspiration for several developments in action detection in video [16, 17]. The advancement in the image domain also brought deep learning for video recognition back into the spotlight. The automatic method of learning features is originally addressed in [18], when the input was raw RGB frames. 3D CNN was utilised to recognise human actions. However, because it only accepts fixed-length input frames, this approach was not well suited for variable-length films. In [19], author proposed a largescale video dataset which is named Sports1M and showed many methods to fuse information into the Available CNN models at that time. In [20], authors used convolutional 3D (C3D) model on consecutive frames in order to learn motion and spatial information and produced better accuracy using a $3 \times 3 \times 3$ kernel. In [21], authors discussed a model for expanding the temporal length of the inputs which is termed as long-term temporal convolution (LTC). In [22], authors have introduced a novel architecture having two streams for acquiring spatial and motion features, and further, fusion was done by employing either an average pooling or a linear SVM model. In [23, 24], additional information sources were proposed for learning the motion features and spatial features for the task of action recognition which was based on two stream method.

3 Methodology

For the proposed human activity recognition task, a method is proposed, which is a combination of the two models CNN and RNN. The proposed architecture is described in Fig. 1.

The network is motivated from [25]. Here, authors have used a combination of CNN and RNN to classify sports in the video.

The inputs to the model are the sequences of RGB colour frames. Each frame is input to a separate convolutional layer. The weights are being shared by all the convolutional layers. The function which is used is ReLU, which is applied at the output layer of the model. Convolved layers shown in Fig. 2 and parameters used in each layer have been shown in Table 1.

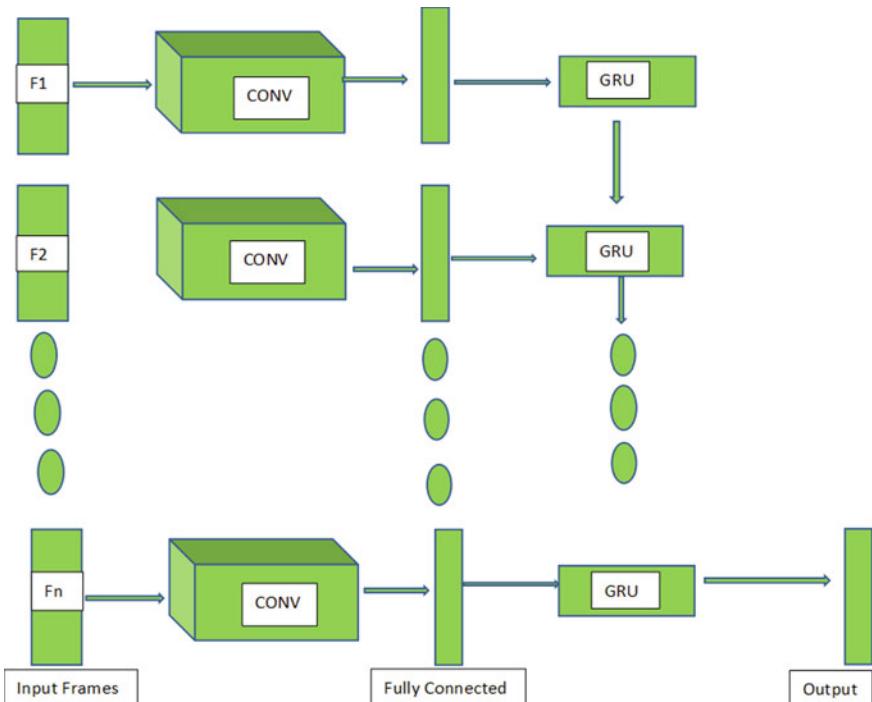


Fig. 1 Illustration of the model used

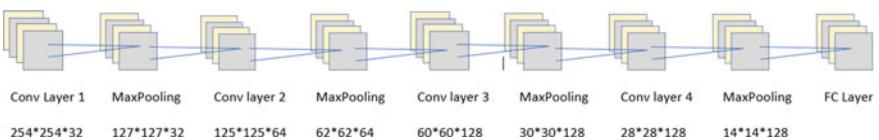


Fig. 2 Description of convolution network layers

Table 1 Parameters used in each layer

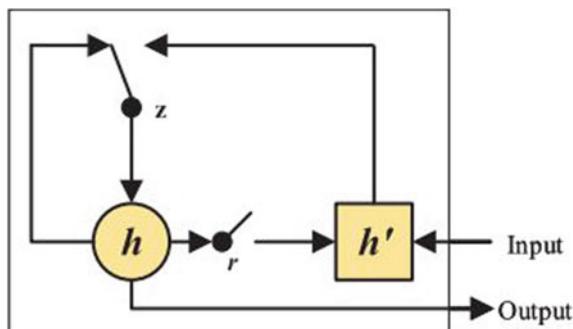
Layers	1st	2nd	3rd	4th
Filter size	3 * 3	3 * 3	3 * 3	3 * 3
Dilation	–	(1, 2)	(2, 1)	(2, 2)
Feature map	32	64	128	128
Non-linearity	ReLU	ReLU	ReLU	ReLU
Dimensionality reduction	MaxPooling	MaxPooling	MaxPooling	MaxPooling

The convolution neural network consists of 32 features map with 3 by 3 receptive field, which is further followed by the MaxPooling layer with a stride of 2. Subsequent layers consist of a context module which is necessary for checking wider area. It has been accomplished by using dilated convolutions which is a process in which kernels are expanded by skipping some pixels. By including the context module, we basically are enabling the model to learn human activity sequences and background more accurately.

Zero padding was done on dilated convolutions. Lastly, a fully connected layer that had 64 neurons was applied, and the results produced were fed to the recurrent unit. A dropout of 0.2 is used. It prevents overfitting and helps in producing better results. After each activation, batch normalisation was done. For the optimization part, ReLU was used. The learning rate was set to ‘0.0001’ throughout the process. The loss function used was categorical cross-entropy. The experiment is performed on 50 epochs.

Lastly, gated recurrent units were applied to find the relation between the frames in a time domain, and it is somewhat similar to LSTM, only it requires fewer parameters.

The GRU design, which couples the input and forget gate into an update gate [26], is a more simplified form of the LSTM architecture [27]. The GRU reduces the gating signals to two as opposed to the three units of the LSTM design. The GRU architecture is demonstrated in Fig. 3, which incorporates of an update gate z and a reset gate r . The update gate controls how quickly the information from the previous

Fig. 3 GRU architecture [26]

state can be added to the current state. The reset gate, on the other hand, is used to regulate how much status information from the previous moment can be disregarded. The forward propagation information can be computed at time step t as follows:

$$Z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (1)$$

$$Z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (2)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (3)$$

$$\tilde{h}_t = \tan h(W x_t + U_r (r_t \odot h_{t-1})) \quad (4)$$

$$h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t \quad (5)$$

where the sigmoid function and element-wise multiplication are denoted by the symbols σ and \odot , respectively. With the help of the GRU's reset gate t_r , update gate t_z , and candidate hidden layer t_h , each recurrent unit is able to adaptively extract dependencies on various time scales. The update gate t_z selects how much of the prior memory should be retained, and the reset gate t_r determines how to merge the current input with the old memory.

4 Experiments

Configurations of the hardware for the experimentation: AMD ryzen 7 1800x 3.6 GHz processor, 64 GB RAM, Nvidia Geforce GTX 1650 ti of 8 GB GPU. TensorFlow backend Keras was used to implement the model and was employed on Windows 10 Professional operating system. Dataset taken for experimentation is the HMDB51 dataset [28]. It consists of 51 classes which have a total of nearly 7000 clips. Each class contains around 101 clips. These 51 classes can be further broken down into 5 distinct categories, including general facial actions, facial actions with object manipulation, general body motions, body movements with object interaction, and body movements for human interactions. These video snippets were gathered from a variety of sources, largely films, with a tiny amount coming from open databases like the Prelinger archive, YouTube, and Google videos (Fig. 4).

The training was done on 70% of the dataset, and 30% of the dataset was separated for testing. Testing was carried out for four frames in a sequence ranging to 20 frames in a sequence and evaluated to achieve a better model with higher accuracy. As keeping more frames in a sequence requires having more GPU memory, we did not proceed to include more frames. The best accuracy 0.84 is calculated as correctly classified classes divided by a total number of classes.

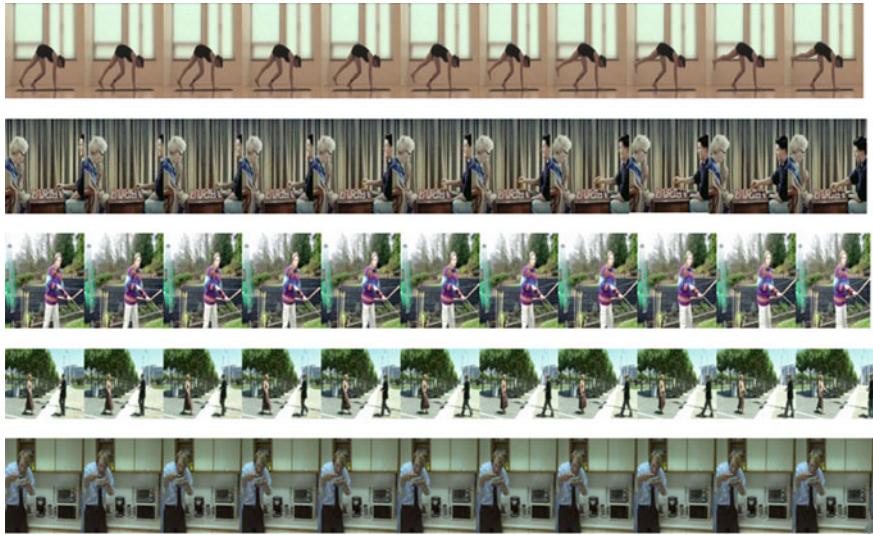


Fig. 4 Frames used for training from 5 different classes (handstand, pick, shoot-bow, walk, eat), extracted from the dataset HMDB51

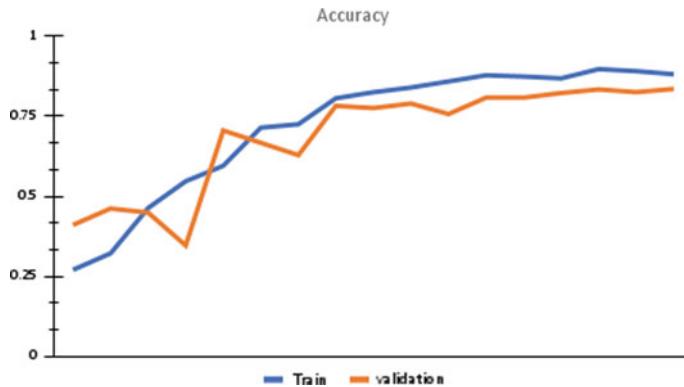


Fig. 5 Training and validation graphs of our proposed method

On analysing the present best accuracy of classifying the dataset with our model, our model outperformed all other methods.

The training and validation graphs are shown in Fig. 5. Result comparison amongst various proposed methods and our proposed methods is shown in Table 2.

Table 2 Results comparison of various methods

Method	Accuracy (%)
IDTs [26]	57.2
MIFS [3]	65.1
Two-stream ConvNet [29]	40.5
Two-stream fusion [22]	47.1
I3D + PoTion [30]	80.9
ST-ResNet [31]	43.4
C3D [32]	54.9
CoViAR [33]	59.1
ARTNet [34]	67.6
Fusion of 2D + pseudo-3D CNN [33]	68.6
Ours (CNN + RNN)	83.7

5 Conclusion

We need to have high-level features to classify any problem accurately. It could be achieved by going deep into the model or selecting higher-resolution images. Both could be explored further if hardware specifications are met. Our proposed hybrid CNN and RNN model consider both spatial and temporal information. This proposed model works and performs exceptionally well for the dataset taken with a classification accuracy of 84%. Further, the model could be improvised to improve significantly.

References

1. Bhardwaj R, Dang K, Gupta SC, Kumar S (2018) Review on human activity recognition using soft computing. *Soft computing: theories and applications*. Springer, Singapore, pp 783–790
2. Putra PU, Shima K, Shimatani K (2022) A deep neural network model for multi-view human activity recognition. *PLoS ONE* 17(1):e0262181
3. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: *Proceedings of the IEEE international conference on computer vision*, Dec 2013, pp 3551–3558
4. Pathak KC, Kundaram SS (2020) Accuracy-based performance analysis of Alzheimer’s disease classification using deep convolution neural network. *Soft computing: theories and applications*. Springer, Singapore, pp 731–744
5. Bordia B, Nishanth N, Patel S, Anand Kumar M, Rudra B (2020) Automated traffic light signal violation detection system using convolutional neural network. *Soft computing: theories and applications*. Springer, Singapore, pp 579–592
6. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
7. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: *Proceedings of the European conference on computer vision*. Springer, Berlin, Germany, 2006, pp 404–417

8. Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the 19th British machine vision conference, 2008, pp 1–275
9. Laptev I (2005) On space-time interest points. *Int J Comput Vis* 64(2–3):107–123
10. Harris CG, Stephens M (1988) A combined corner and edge detector. *Proc Alvey Vis Conf* 15(50):10–5244
11. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on multimedia, 2007, pp 357–360
12. Wang H, Kläser A, Schmid C, Liu C-L (2013) Dense trajectories and motion boundary descriptors for action recognition. *Int J Comput Vis* 103(1):60–79
13. Carmona JM, Climent J (2018) Human action recognition by means of subtensor projections and dense trajectories. *Pattern Recognit* 81:443–455
14. Li F-F, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05), Jul 2005
15. Wang Y, Mori G (2009) Max-margin hidden conditional random fields for human action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Jun 2009, pp 872–879
16. Crasto N, Weinzaepfel P, Alahari K, Schmid C (2019) MARS: motion-augmented RGB stream for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Jun 2019
17. Yang H, Yuan C, Li B, Du Y, Xing J, Hu W, Maybank SJ (2019) Asymmetric 3D convolutional neural networks for action recognition. *Pattern Recognit* 85:1–12
18. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell* 35(1):221–231
19. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Jun 2014, pp 1725–1732
20. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE international conference on computer vision (ICCV), Dec 2015, pp 4489–4497
21. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. *IEEE Trans Pattern Anal Mach Intell* 40(6):1510–1517
22. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
23. Yang X, Molchanov P, Kautz J (2016) Multilayer and multimodal fusion of deep neural networks for video classification. In: Proceedings of the 24th ACM international conference on multimedia, pp 978–987
24. Shi Y, Tian Y, Wang Y, Huang T (2017) Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Trans Multimed* 19(7):1510–1520
25. Russo MA, Filonenko A, Jo K (2018) Sports classification in sequential frames using CNN and RNN. In: 2018 international conference on information and communication technology robotics (ICT-ROBOT), pp 1–3. <https://doi.org/10.1109/ICT-ROBOT.2018.8549884>
26. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) [Online]
27. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J (2017) LSTM: a search space odyssey. *IEEE Trans Neural Netw Learn Syst* 28(10):2222–2232
28. Kuehne H et al (2011) HMDB: a large video database for human motion recognition. In: 2011 international conference on computer vision. IEEE
29. Lan Z, Lin M, Li X, Hauptmann AG, Raj B (2015) Beyond Gaussian pyramid: multiskip feature stacking for action recognition. In: CVPR, pp 204–212
30. Zhu W, Hu J, Sun G, Cao X, Qiao Y (2016) A key volume mining deep framework for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Jun 2016, pp 1991–1999

31. Sun L, Jia K, Yeung D-Y, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: ICCV
32. Feichtenhofer C, Pinz A, Wildes RP (2016) spatiotemporal residual networks for video action recognition. In: NIPS
33. Wang L, Li W et al (2018) Appearance-and-relation networks for video classification. In: CVPR
34. Wu C-Y, Zaheer M et al (2018) Compressed video action recognition. In: CVPR

A New Framework for Disease Prediction: Using Dimensionality Reduction and Feature Selection



Shreya Sahu, Pranesh Das, and A. Binu Jose

Abstract Clinical databases often consist of large number of attributes and dimensions. For clinical data analysis, some parameters are not helpful for diagnosis and prediction purposes. Therefore, applying feature selection is necessary as it removes unimportant features and improves effectiveness of the model. In this paper, a framework for disease prediction with dimensionality reduction and feature selection is proposed. The framework also uses ensemble learning methods along with some well-known classification techniques. The proposed framework is validated with five standard datasets with four different types of diseases. The comparative performance analysis shows that the proposed approach performs well with respect to some of the existing approaches. The proposed approach has the potential to be used by the practitioners for an effective prediction of other medical diseases as well.

Keywords Feature selection (FS) · Dimensionality reduction · Ensemble learning · Classification · Disease prediction

1 Introduction

Nowadays, the digital healthcare systems generate huge amount of medical data. As practitioners, it is difficult to obtain valuable information from this massive data. Advanced statistical approaches are frequently used to gain insights or identify significant information using data mining, artificial intelligence, machine learning, and deep learning [1]. Data mining and knowledge exploration in the medical field represent a relatively new research area that is of great interests to many researchers [2].

Since the medical data gathering is gradually increasing, physicians are capable of diagnosing the disease better. But the presence of large amount of data makes the diagnosis really complex and time-consuming. This is mainly due to high dimensionality of data that contains irrelevant and redundant data. To solve these problems,

S. Sahu · P. Das · A. Binu Jose (✉)

Department of Computer Science and Engineering, National Institute of Technology Calicut,
Kozhikode, Kerala, India

e-mail: binujose_p200050cs@nitc.ac.in

feature selection and dimensionality reduction methods play an important role. These methods extract important features from the high dimensional data so that the dataset can be transformed into a lower dimension [3].

Moreover, it not only supports in training the model faster but also lowers the complexity of the model. It makes it easier to improve the performance and reduce the complexity. Researchers are using various machine learning techniques to predict diseases using vast amounts of medical data [4].

A cluster-based decision tree learning (CDTL) approach was proposed in [5]. Significant features are obtained using entropy for each combination of class-set. Finally, on all the entropy clusters, machine learning models including random forest (RF), decision tree (DT), linear model (LM), and support vector machine (SVM) were applied. RF classifier outperformed other models by improving the accuracy from 76.70 to 89.30% on full and reduced dataset.

In [6], sub-bands for electrocardiogram (ECG) signals were computed using discrete wavelet transform (DWT) technique. Information based on time resolution and frequency could be easily computed with these sub-bands. For dimensionality reduction, three standard algorithms including Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) were applied on sub-bands. Further, probabilistic neural network (PNN), artificial neural network (ANN), and support vector machine (SVM) were used for classification. Results showed that ICA in combination with PNN produced highest accuracy.

A heart disease diagnosis system was proposed in [7], where a feature subset of reduced dimension was used. The distance measure was used to rank features in forward inclusion, forward selection, and backward elimination methods. The forward selection method generated smaller subsets and increased the accuracy as compared to back-elimination and forward inclusion techniques.

In [8], correlation-based feature selection was used where Random Forest with Stratified KFold cross-validation method in the reduced subsets outperformed and obtained 86.94% accuracy, whereas Hoeffding tree method obtained 85.43% accuracy.

In [9], chi squared method with PCA are used to obtain fourteen feature out of seventy four features. CHI-PCA with Random Forest classifier obtained highest accuracy of 98.7% on Cleveland datasets.

In [10], the data of 50 cervical cancer patients were used for prognosis. The important features were selected based on the decease in gini impurity. In this work, six classifiers including logistic regression, k-nearest neighbors, random forest, Naive Bayes, support vector machine, and neural network were compared using AUC curve. Fisher's score and Relief methods were used for feature selection, and variational autoencoder was used for feature extraction for Parkinson's Disease diagnosis in [11]. The optimum features were trained on a multi-kernel support vector machine classifier which obtained an accuracy of 0.918. Principal Component Analysis and Singular Value Decomposition were used for transforming the high dimensional data, and decision tree importance is used for feature selection. After feature reduction, a classification for thyroid disease was performed using neural network (NN) and K-nearest neighbor (KNN) classifiers in [12]. A score-based artificial fish swarm

algorithm (SAFSA) for feature selection is introduced to reduce the vocal features for Parkinson's Disease classification in [13]. Classification of PD is done by using ensemble learning techniques with hybrid classifiers of random forest, fuzzy K-nearest neighbor, fuzzy convolutional neural network, and kernel support vector machines. A modified grey wolf optimization is proposed in [14] since the GWO is very famous for its characteristics such as simplicity, tuning parameters, scalability, etc. The enhanced GWO is proposed with support vector machine (SVM) to obtain the reduced subset in Wisconsin Diagnostic Breast Cancer (WDBC) database. The classification accuracy obtained is 98.24%. In [15], a Promoted Crow Search Algorithm (PCSA) is introduced to improve the crow's search performance both locally and globally. The convergence rate of the algorithm has been increased by introducing the concept of chaos. On an average, the algorithm has performed 2.5% better in fitness index and 20%. The main focus of the analysis is to explore different feature reduction methods as some of the existing work appears to lack better feature selection methods. Also, an attempt is made to increase accuracy of predicted model using various ensemble learning methods. Another motivation is to get the best model with high predictions and less losses. The proposed framework is described in the next section.

2 Proposed Framework

The proposed study focuses on increasing classification accuracy by reducing the number of features and making use of ensemble learning methods in the various disease dataset. The framework for classifying diseases is depicted in Fig. 1. The key components of the framework include data collection, data pre-processing, feature selection, feature extraction, data splitting, model training with classifiers, ensemble learning, and model evaluation. Firstly, the original dataset is pre-processed and applied directly to all the classification models. Secondly, the pre-processed data is applied to dimensionality reduction and feature selection methods. The classification models were then trained with the reduced datasets. Model accuracy, F1 score, and loss were recorded for each models, and comparative performance analysis is made. The following sections describe the building blocks of the suggested framework.

2.1 Data Collection

This study gathered UCI Cleveland heart disease dataset [16], Breast Cancer Wisconsin dataset [17], UCI Dermatology dataset [18], UCI Cervical Cancer dataset [19], and Kaggle Framingham dataset [20], all of which can be accessed online for testing purposes (Table 1).

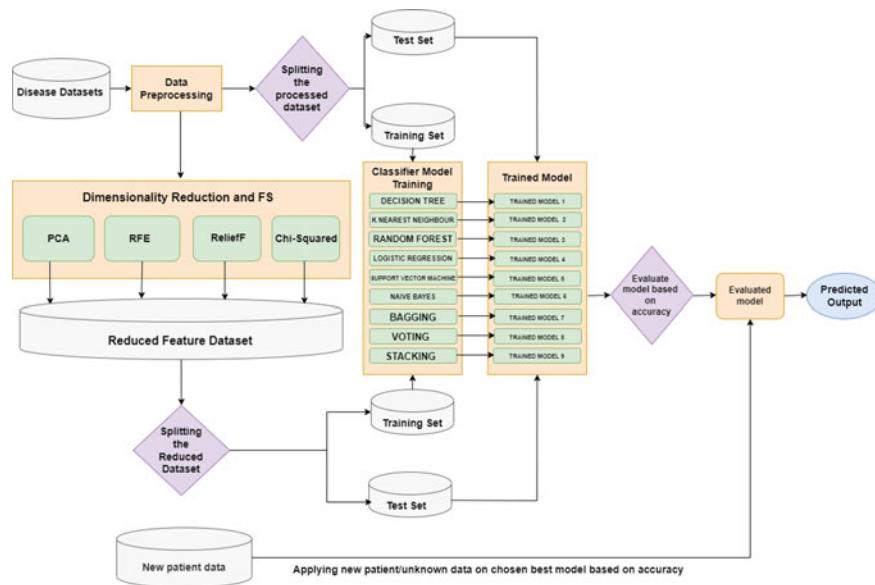


Fig. 1 Proposed framework

Table 1 Disease classification datasets

S. No.	Datasets	Number of features	Number of instances	Number of classes
1	Cleveland heart disease dataset	14	303	2
2	Framingham heart disease dataset	16	4133	2
3	Breast cancer Wisconsin dataset	33	569	2
4	Cervical cancer dataset	36	858	2
5	Dermatology dataset	35	358	6

2.2 Data Pre-processing

Pre-processing the dataset is necessary for the effective representation of the quality of data. The dataset has been pre-processed using techniques such as eliminating missing values from features, removing the redundant features, and normalizing the dataset. Missing values were handled by replacing it with the mean value of the attribute.

2.3 Dimensionality Reduction and Feature Selection

Many datasets are very difficult to deal with because of the large number of features. Higher number of features make it difficult to classify the dataset with better accuracy. Many studies have shown that using PCA as a reduction technique improves the performance of model. PCA creates a new set of features from the original dataset which are called principal components. The features are arranged in the decreasing amount of variance calculated in a particular feature. The higher the variance, higher the informative a feature is [21]. The components with higher amount of explained variance are retained for model training.

Feature selection methods are used to identify the best subset from the overall dataset that can provide the best classification accuracy. Feature selection methods are applied before training any model to eliminate the irrelevant and redundant data to improve the performance of a model. Three standard feature selection methods are used in the framework are Recursive Feature Elimination (RFE), ReliefF, and Chi-squared method of independence.

RFE is a wrapper style feature selection method that uses another machine learning model at the core to select features. Here, logistic regression model is used, and the features are recursively selected by ranking features using logistic regression coefficients as feature importance score. The method discards the least important features and then refits the model with reduced features. The process continues till it finds the optimal number of features. ReliefF feature selection method works by detecting feature dependencies. Instead of searching through feature combinations, it uses the concept of nearest neighbors to obtain feature statistics that account for correlations indirectly. The process is repeated for a number of times with the nearest neighbors, and the weights of every feature are calculated. Based on the feature weights, the features are selected. Chi-squared test of independence is a univariate feature selection method that calculates the correlation between two features.

2.4 Classification Models

In this framework, six different supervised classification algorithms are used. The classifiers are chosen for their diverse nature of classifying data so that there should not be any biasing or correlation in predicting the output value. The algorithms chosen based on different characteristics are random forest, k-nearest neighbor, support vector machine, decision tree, Gaussian Naive Bayes, and logistic regression.

To improve the predictive power of a model, ensemble learning methods such as voting, stacking, and bagging methods are chosen. Multiple models' predictions are combined in a voting ensemble, which can be used to classify or predict data. It improves the model performance by taking majority vote of many individual classifiers performances. Bagging method is useful in lowering the variance. In stacking, the output of individual estimators is stacked, and the final prediction is computed

Table 2 Selected parameters for classifier models

Classifier models	Parameter settings for the models
Decision tree	Criterion = 'gini,' Presort = False, MinImpuritySplit = 1e-07, MinSamplesSplit = 2, MinSamplesLeaf = 1, RandomState = 42, Splitter = 'best'
Random forest	Criterion = 'gini,' MaxFeatures = 1, MinSamplesLeaf = 3, MinSamplesSplit = 10, NoOfEstimators = 100, RandomState = 42, Verbose = 0
Logistic regression	Penalty = L2, InterceptScaling = 1, FitIntercept = True, RandomState = 42, C = 1.0, Solver = 'liblinear,' MaxIteration = 100, MultiClass = 'auto,' Verbose = 0
GaussianNB	Priors = None, VariableSmoothing = 1e-09
SVM	C = 1.0, Degrees = 3, Gamma = 'scale,' Kernel = 'rbf,' Shrinking = True, Probability = True, CacheSize = 200, max_iter = 200, RandomState = 42
K-nearest neighbor	NoOfNeighbors = 5, Algorithm = 'auto,' Weights = 'uniform,' LeafSize = 30, Power = 2, Metric = 'minkowski'
Bagging	BaseEstimator = LogisticRegression(), RandomState = 42, Solver = 'liblinear,' MaxIteration = 100
Voting	Estimators = {LR, SVC, RFC}, TypeOfVoting = 'soft'
Stacking	Estimators = {LR, SVC, RFC}, FinalEstimator = GaussianNB()

using a classifier. The purpose of the selected classifier techniques is to find how different feature selection methods perform on different models. The parameter settings for the classifiers can be seen in Table 2.

2.5 Performance Metric

For performance evaluation, accuracy, recall, F1 score, precision, and loss were recorded. The mentioned performance metric was calculated from true negative (TN), true positive (TP), false negative (FN), and false positive (FP) values that are generated using confusion matrix. The loss function used for the analysis is log loss because of the binary or multi-class classification.

3 Result and Discussion

This section consists the experimental results obtained using the proposed framework. A comparative analysis is done for each model based on accuracy and loss. Tables 3, 4, 5, 6, and 7 show the tabular representation of accuracy, F1 score, and loss of all the classifiers with respect to different feature reduction techniques. Figure 2

presents the graphical representation of accuracy values of each models for all the datasets. The results are compared for two scenarios, i.e., on original and reduced datasets. Below results show how different feature selection model (check) increases the accuracy of different classification models. In Tables 3, 4, 5, 6, and 7, it is clear that applying classification models after FS and dimensionality reduction methods produce better results as compared to applying models on raw data. Tables 3, 4, 5, 6, and 7 show the number of features selected by PCA, RFE, ReliefF, and Chi-sq method. For Cleveland Dataset in Table 3, the performance is improved using ensemble methods in raw data. Using PCA, performance of LR, SVM, and KNN is improved, whereas other classifier did not perform better than raw data. RFE improved the performance of most of the classifiers. DT performed best using RFE method with 83.6% accuracy. Stacking classifier using ReliefF with 10 features performed best among all the methods with 91.8% and loss 0.83. Also in Chi-sq most of the models improved the performance, and stacking outperformed with 91.8% with 8 features and loss 0.82. In Framingham dataset, in Table 4, PCA reduced the components to 12, it improved the performance of LR, GNB, KNN, and bagging. Number of features selected in RFE is 10 where only DT improved the performance from 75.9 to 76.3%. Using ReliefF with top 10 features improved the performance of DT and KNN classifiers. Chi-squared method with 6 features outperformed in all the classifiers with the best performance in SVM and voting classifier with 85.64% accuracy in both. In Table 5, using PCA, RFE, ReliefF, and Chi-sq method with reduced number of features improved the performances of almost all the classifiers. LR obtained the best result of 97.67% using PCA. In Table 6, the components are reduced to 20 using PCA in Dermatology Dataset. GNB, SVM, and KNN performed better using PCA. Using ReliefF with reduced features improved the performance of all the classifiers with voting giving the highest accuracy of 100% with loss 0.1. DT didn't show any improvement using Chi-sq, whereas improved the performances in other classifiers with RF outperforming with 99.49% accuracy. Table 7 shows the performance of classifiers on Breast Cancer Dataset where PCA reduced the number of components to 13. RFE selected top 12 features, where DT, LR, GNB, and bagging improved the performance with 94.73%, 97.07%, 94.73%, and 96.5% accuracy, respectively. ReliefF selected top 10 features and improved accuracy of DT, LR, and Bagging. Reduced subset with 11 features in Chi-sq method obtained best accuracy in LR and voting classifier with 98.24%.

It can be seen that applying dimensionality reduction in dataset with large number of features improves the result significantly as compared to applying it on a lower dimensional dataset. Ensemble learning methods by gathering individual models into ensembles gave better predictive performances such as lower loss and high classification accuracy. The performance of classification models are better in case of feature selection techniques as compared to PCA because PCA sometimes eliminates features that has low variance but is relevant for predicting target. The performances of proposed study are compared with some of the recent studies for all the included datasets in Table 8.

Table 3 Performance of classifiers on Cleveland heart disease dataset

Feature selection	Performance metric	Cleveland dataset								
		DT	RF	LR	GNB	SVM	KNN	BAG	VOT	STK
Raw data (features = 14)	Accuracy	77.04	83.6	88.24	85.24	85.24	81.96	90.1	88.5	90.16
	F1 score	79.6	83.5	88.5	85.2	85.2	81.9	90.1	90.1	
	Loss	7.9	0.43	0.4	0.69	0.4	3.07	0.38	0.44	0.83
PCA (components = 10)	Accuracy	73.7	80.3	90.1	85.24	88.52	83.6	86.52	83.6	86.8
	F1 score	73.6	80.3	90.1	85.2	88.5	83.6	86.8	83.5	86.8
	Loss	9.0	0.48	0.38	0.4	0.4	1.38	0.4	0.4	0.91
RFE (features = 11)	Accuracy	83.6	85.24	88.52	86.88	85.2	83.6	88.52	90.16	90.1
	F1 score	83.5	85.2	88.5	86.8	85.2	83.5	88.5	90.1	
	Loss	5.66	0.42	0.41	0.67	0.42	3.06	0.41	0.41	0.99
ReliefF (features = 10)	Accuracy	78.68	85.24	86.8	86.88	83.6	80.32	88.52	88.52	91.8
	F1 score	78.6	85.2	86.8	86.8	83.5	80.3	88.5	88.5	91.7
	Loss	7.3	0.4	0.41	0.59	0.4	3.64	0.41	0.4	0.83
Chi-sq (features = 8)	Accuracy	81.96	86.88	86.88	86.88	86.88	83.6	86.88	90.16	91.8
	F1 score	81.9	86.8	86.8	86.8	86.8	83.5	86.8	88.5	91.7
	Loss	6.24	0.44	0.41	0.6	0.37	2.48	0.42	0.39	0.82

Table 4 Performance of classifiers on Framingham heart disease dataset

Feature selection	Performance metric	Framingham dataset								
		DT	RF	LR	GNB	SVM	KNN	BAG	VOT	STK
Raw data (features = 16)	Accuracy	75.96	83.3	83.87	81.0	83.3	81.5	83.5	83.2	83.06
	F1 score	75.6	76	77.3	79.5	76.2	76.9	77.38	76	79.4
	Loss	8.3	0.42	0.4	1.4	0.44	2.5	0.4	0.41	0.69
	Accuracy	75.5	83.38	84.03	83.1	83.4	82.5	84.2	83.2	82.5
	F1 score	74.9	75.9	77.8	80.3	76.1	77.8	78.2	75.9	79.4
	Loss	8.4	0.43	0.4	0.57	0.43	2.6	0.4	0.41	0.61
RFE (features = 10)	Accuracy	76.3	83.3	83.7	81.6	83.3	81.9	83.7	83.3	82.7
	F1 score	75.9	76.2	77	79.4	76.4	77.2	77	76.2	77.5
	Loss	8.16	0.42	0.4	1.39	0.45	2.3	0.4	0.4	0.87
	Accuracy	82.98	83.14	83.2	81.7	83.2	83.7	83.2	83.2	82.66
	F1 score	76.6	75.8	75.6	78.2	76.2	75.6	75.6	76	77.4
	Loss	1.18	0.45	0.43	1.19	0.45	3.28	0.42	0.43	0.68
Chi-sq (features = 6)	Accuracy	85	85.48	85.56	83.38	85.64	82.09	85.48	85.64	82.74
	F1 score	80	80.3	79.7	80.3	79.5	78.8	79.7	79.5	80
	Loss	1.1	0.4	0.38	1.2	0.4	2.3	0.38	0.38	0.59

Table 5 Performance of classifiers on cervical cancer dataset

Feature selection	Performance metric	Cervical cancer dataset								
		DT	RF	LR	GNB	SVM	KNN	BAG	VOT	STK
Raw data (features = 34)	Accuracy	94.57	95.34	96.5	94.5	94.1	93.7	95.7	94.5	94.9
	F1 score	94.57	94.7	96.3	94.5	91.3	91.1	95.6	93.2	95.2
	Loss	1.87	0.09	0.11	2.98	0.19	1.53	0.11	0.11	0.57
PCA (components = 17)	Accuracy	92.63	96.51	97.67	92.63	95.34	95.7	97.28	96.9	95.34
	F1 score	92.7	96	97.5	92.7	94.8	95.1	97.1	96.5	95.4
	Loss	2.5	0.1	0.1	1.12	0.13	0.7	0.1	0.11	0.37
RFE (features = 7)	Accuracy	97.28	97.28	97.28	97.28	96.12	97.28	97.28	96.51	96.12
	F1 score	97.14	97.1	97.1	97.1	96.4	97.1	97.1	96.7	96.4
	Loss	0.33	0.2	0.098	0.82	0.09	0.31	0.097	0.093	0.9
Relieff (features = 12)	Accuracy	97.28	96.51	97.28	97.28	97.28	97.28	97.28	96.14	96.12
	F1 score	97.14	96.45	97.14	97.14	97	97.14	97.14	96.4	96.4
	Loss	0.33	0.2	0.09	1.26	0.09	0.31	0.09	0.09	0.81
Chi-sq (features = 10)	Accuracy	96.89	96.51	96.51	96.9	96.12	97.28	96.51	96.12	96.12
	F1 score	96.6	96.4	96.4	96.6	96.4	97.14	96.45	96.42	96.42
	Loss	0.33	0.21	0.1	1.37	0.1	1.34	0.09	0.09	0.99

Table 6 Performance of classifiers on dermatology dataset

Feature selection	Performance metric	Dermatology dataset								
		DT	RF	LR	GNB	SVM	KNN	BAG	VOT	STK
Raw data (features = 34)	Accuracy	94.44	97.22	98.6	84.72	70.83	84.72	98.14	97.22	97.22
	F1 score	94.2	97.2	98.6	83.3	67.0	85.8	98.6	97.2	97.2
	Loss	1.91	0.48	0.07	0.58	1.23	0.08	0.15	0.31	0.14
PCA (components = 20)	Accuracy	91.66	95.83	98.61	97.22	94.44	86.91	98.61	97.22	97.22
	F1 score	91.3	95.7	98.6	97.2	94.5	87.08	98.6	97.2	97.3
	Loss	2.8	0.69	0.14	0.04	0.36	1.21	0.15	0.21	0.19
RFE (features = 18)	Accuracy	91.67	99.07	99.07	80.55	99.07	97.22	98.14	98.14	98.14
	F1 score	91.8	99.07	99.08	76.08	99.06	97.2	98.11	98.14	98.14
	Loss	2.87	0.38	0.09	5.62	0.13	0.06	0.11	0.12	0.26
Relieff (features = 24)	Accuracy	96.29	99.07	99.07	99.07	99.07	99.07	99.07	100	99.074
	F1 score	96.12	99.06	99.07	96.12	99.06	99.06	99.07	100	99.08
	Loss	1.29	0.43	0.12	5.32	0.14	0.04	0.12	0.11	0.31
Chi-sq (features = 24)	Accuracy	87.96	99.49	99.07	87.96	99.28	99.07	99.07	99.07	99.07
	F1 score	87.8	99.5	99.07	87.8	99.3	99.06	99.07	99.07	99.07
	Loss	2.9	0.41	0.11	5.38	0.14	0.07	0.12	0.12	0.31

Table 7 Performance of classifiers on breast cancer dataset

Feature selection	Performance metric	Breast cancer dataset					
Raw data (features = 32)	Accuracy	93.56	97.6	95.9	93.5	98.2	95.9
	F1 score	93.6	97.65	95.86	93.6	98.2	95.8
	Loss	2.2	0.09	0.15	0.45	0.06	0.16
PCA (components = 13)	Accuracy	96.49	95.9	98.24	96.49	97.07	96.49
	F1 score	96.5	95.9	98.2	96.5	97.08	96.4
	Loss	1.21	0.18	0.06	0.4	0.07	0.27
RFE (features = 12)	Accuracy	94.73	97.66	97.07	94.73	97.66	95.32
	F1 score	94.74	97.07	97.04	94.7	97.6	97.06
	Loss	1.81	0.09	0.16	0.07	0.07	0.45
Relieff (features = 10)	Accuracy	95.32	96.49	97.07	95.32	96.49	95.32
	F1 score	95.33	96.47	97.66	95.33	96.45	95.3
	Loss	1.61	0.09	0.15	0.37	0.09	0.29
Chi-sq (features = 11)	Accuracy	94.15	96.49	98.24	94.15	94.15	95.9
	F1 score	94.15	96.5	98.24	94.15	94.05	95.88
	Loss	2.01	0.09	0.07	0.32	0.13	0.09

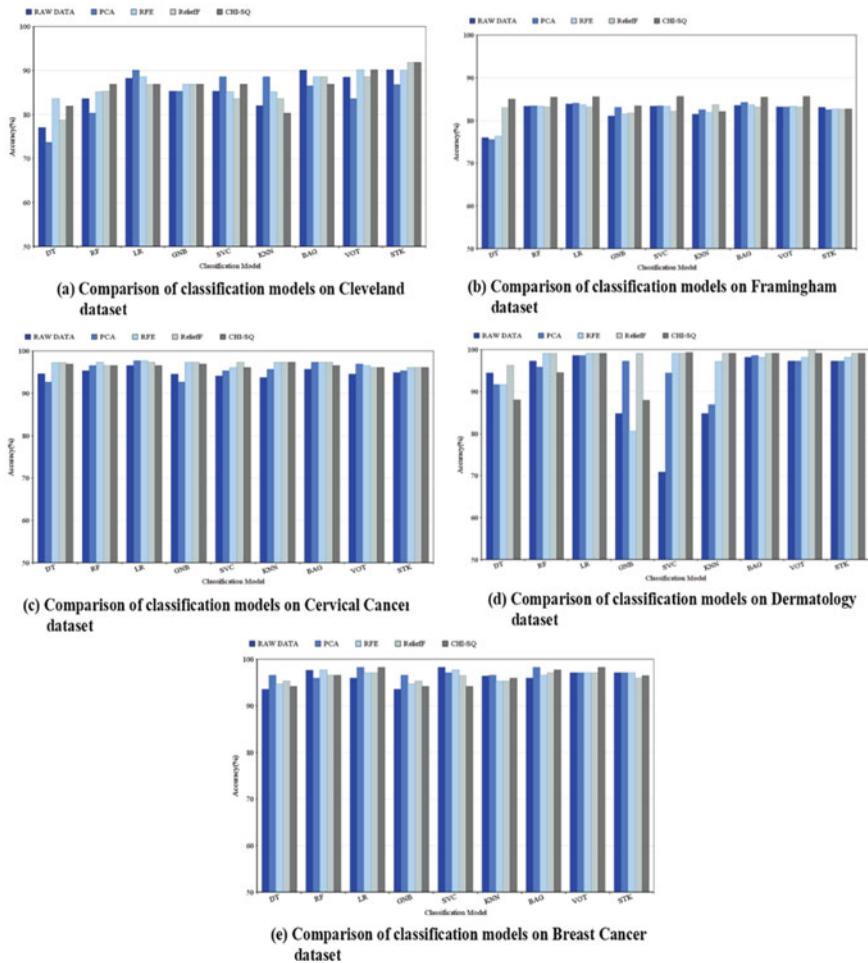


Fig. 2 Comparison of classification models on different datasets

4 Conclusion and Future Work

Very large amount of data in the field of medical science is still a matter of great concern especially for the practitioners for prognosis of many disease. Dimensionality reduction can be a great milestone to resolve this problem. The paper introduces four dimensionality reduction and feature selection techniques along with the classification models for prediction of different types of diseases. Tables 3, 4, 5, 6, and 7 show the comparative analysis between these feature reduction techniques. The analysis clearly shows that eliminating unimportant features can significantly improve

Table 8 Comparative performance analysis with existing models

Dataset	Previous studies	Accuracy (%)
Cleveland heart disease	CDTL + RF [5]	89.30
	KNN [2]	87
	Fuzzy AHP + FFNN [22]	83
	RF + stratified KFold [8]	86.94
	Feature importance + bagging [23]	89
	ReliefF + stacking	91.80
Framingham heart disease	Chi-sq + stacking	91.80
	Risk score + RST [24]	85.11
Cervical cancer	Chi-sq + voting	85.64
	SVM [25]	94
	Stacked autoencoder [26]	97
	Deep CNN [27]	95
	Random forests [28]	95
Dermatology	Bagging + ReliefF/RFE	97.28
	SVM + correlation matrix [29]	99.86
	Stacking [30]	99.67
	Feature importance + gradient boosting [31]	99.68
Wisconsin diagnosis breast cancer	ReliefF + voting	100
	Logistic regression [32]	95
	P-boosted C5.0 [33]	97.65
	K-SVM [34]	97.38
	ANN [35]	97.3
	PCA + bagging	98.24

the performance of most of the predicting models. For further enhancement in this study, deep learning techniques can be explored for the classification of diseases.

References

1. Coccia M (2020) Deep learning technology for improving cancer care in society: new directions in cancer imaging driven by artificial intelligence. *Technol Soc* 60:101198
2. Srivastava K, Choubey DK (2020) Heart disease prediction using machine learning and data mining. *Int J Recent Technol Eng* 9(1):212–219
3. Fodor IK (2002) A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Laboratory, Livermore, CA
4. Chen M, Hao Y, Hwang K, Wang L, Wang L (2017) Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* 5:8869–8879
5. Gopu M, Swarnalatha P (2021) Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evol Intell* 14

6. Martis RJ, Rajendra Acharya U, Min LC (2013) ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. *Biomed Signal Process Control* 8(5):437–448
7. Shilaskar S, Ghatol A (2013) Feature selection for medical diagnosis: evaluation for cardiovascular diseases. *Expert Syst Appl* 40(10):4146–4153
8. Alim MA, Habib S, Farooq Y, Rafay A (2020) Robust heart disease prediction: a novel approach based on significant feature and ensemble learning model. In: 2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET), pp 1–5
9. Gárate-Escamila AK, El Hassani AH, Andrès E (2020) Classification models for heart disease prediction using feature selection and PCA. *Inform Med Unlocked* 19:100330
10. Nakajo M, Jinguiji M, Tani A, Yano E, Hoo CK, Hirahara D, Togami S, Kobayashi H, Yoshiura T (2022) Machine learning based evaluation of clinical and pretreatment 18F-FDG-PET/CT radiomic features to predict prognosis of cervical cancer patients. *Abdom Radiol* 47(2):838–847
11. Gunduz H (2021) An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomed Signal Process Control* 66:102452
12. Jha R, Bhattacharjee V, Mustafi A (2022) Increasing the prediction accuracy for thyroid disease: a step towards better health for society. *Wireless Pers Commun* 122(2):1921–1938
13. Gafoor SHA, Theagarajan P (2022) Intelligent approach of score-based artificial fish swarm algorithm (SAFSA) for Parkinson's disease diagnosis. *Int J Intell Comput Cybern*
14. Kumar S, Singh M (2021) Breast cancer detection based on feature selection using enhanced grey wolf optimizer and support vector machine algorithms. *Vietnam J Comput Sci* 8(02):177–197
15. Samieyan B, MohammadiNasab P, Mollaei MA, Hajizadeh F, Kangavari M (2022) Solving dimension reduction problems for classification using promoted crow search algorithm (PCSA). *Computing* 1–30
16. Aha DW (1988) UCI machine learning repository
17. Mangasarian OL, Wolberg WH, Street WN (1995) UCI machine learning repository
18. Guvenir HA, İlter N (1998) UCI machine learning repository
19. Fernandes J, Fernandes K, Cardoso JS (2017) UCI machine learning repository
20. UCI machine learning repository
21. Abdi H, Williams LJ (2010) Principal component analysis. *Wiley Interdiscip Rev Comput Stat* 2(4):433–459
22. Vivekanandan T, Sriman NC, Iyengar N (2017) Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Comput Biol Med* 90:125–136
23. Oswald, Sathwika GJ, Bhattacharya A (2022) Prediction of cardiovascular disease (CVD) using ensemble learning algorithms. In: 5th joint international conference on data science & management of data (9th ACM IKDD CODS and 27th COMAD). Association for Computing Machinery, New York, NY, pp 292–293
24. Chen Y-S, Cheng C-H, Chen S-F, Jhuang J-Y (2020) Identification of the Framingham risk score by an entropy-based rule model for cardiovascular disease. *Entropy* 22(12):1406
25. Wu W, Zhou H (2017) Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access* 5:25189–25195
26. Adem K, Kılıçarslan S, Cömert O (2019) Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Syst Appl* 115:557–564
27. Zahras D, Rustam Z (2018) Cervical cancer risk classification based on deep convolutional neural network. In: 2018 international conference on applied information technology and innovation (ICAITI). IEEE, pp 149–153
28. Abdoh SF, Rizka MA, Maghraby FA (2018) Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access* 6:59475–59485
29. Pal S, Verma AK (2019) Prediction of skin disease with three different feature selection techniques using stacking ensemble method. *Appl Biochem Biotechnol*

30. Verma AK, Pal S, Tiwari BB (2020) Skin disease prediction using ensemble methods and a new hybrid feature selection technique. *Iran J Comput Sci* 3(4):207–216
31. Pal S, Kumar S, Verma AK (2019) Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study. *Appl Biochem Biotechnol*
32. Dinesh P, Kalyanasundaram P (2022) Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest, and decision tree to measure accuracy. *ECS Trans* 107(1):12681–12691
33. Tian J-X, Zhang J (2022) Breast cancer diagnosis using feature extraction and boosted c5.0 decision tree algorithm with penalty factor. *Math Biosci Eng* 19(3):2193–2205
34. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Syst Appl* 41(4):1476–1482
35. Aalaei S, Shahraki H, Rowhani manesh A, Eslami S (2016) Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran J Basic Med Sci* 19(5):476

Review of Metaheuristic Techniques for Feature Selection



Sanat Jain, Ashish Jain, and Mahesh Jangid

Abstract Due to the rise of high-dimensional datasets in various sectors, feature selection is one of the most important issues. The primary goal of the feature selection problem is to lower the dimension of the feature set while retaining the performance accuracy. Several techniques of feature selection have been developed to obtain the optimal subset of features. To obtain optimal features, metaheuristics are modern optimization techniques that are used by the research community. In this paper, four groups of metaheuristic techniques have been identified based on their behavior. The classifier name, datasets, and assessment metrics for the metaheuristic methods used to solve the feature selection challenges are provided. After reviewing the papers, difficulties and problems are also observed while trying to use various metaheuristic methods to find the best subset of features. For those researchers who desire to continue their work on creating or refining metaheuristic techniques for feature selection, several research gaps are also mentioned.

Keywords Feature selection · Metaheuristic · Wrapper method

1 Introduction

In today's scenario, most researchers facing problems due to lots of data available in the market. Handling this data is a very complicated and challenging task because there are a lot of attributes/features present in the dataset. To find useful information, it is not necessary to select all features because some sets of data may be redundant, and contain missing and irrelevant information, which decreases the result of the model. For that reason, diminishing the original dataset size while persisting the precision of execution is the primary aim to reduce the number of features. To reduce the features, the process of extractions and process of selection of features are important. From

S. Jain · A. Jain (✉)

Department of Information Technology, Manipal University Jaipur, Jaipur, India
e-mail: ashishjn.research@gmail.com

M. Jangid

Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India

the original dataset, feature construction or extraction constructs a new group of attributes, while relevant features are selected by the feature selection process. This article mainly focuses on the feature selection process.

Nowday's machine learning faces major challenges to solve the problem of feature selection. If a set contains n number of elements, complete 2^n subsets are conceivable from which the best subset must be picked. It will be extremely tough when n approaches a huge number because the model's performance cannot be evaluated at each subset. As a result, several strategies have been developed to cope with the situation. Some techniques such as random search, exhaustive search, and greedy search. are used to discover the optimum subset for feature selection problems. In most approaches, premature convergence, extreme complexity, and high computing costs are common problems. To deal with these types of problems, metaheuristic optimization algorithms are given a lot of attention. To find the optimum subset of attributes they are the most efficient and productive techniques while keeping the model's accuracy [1].

Over the few decades, various kinds of optimization methodologies have been proposed to address different types of problems. Feature selection has evolved as an interesting research issue due to its many uses in different areas such as bioinformatics, text mining, industrial applications, computer vision, and image processing.

In recent years, to solve optimization problem, various metaheuristic algorithms applied successfully because it obtains optimal solution with a higher level of heuristics. The reason behind using metaheuristic techniques is easy to implement, simple structure, avoiding local optima, and their ability to explore various disciplines [2].

In the past few years, many nature-inspired metaheuristic techniques have been developed that are attracting the attention of numerous researchers. This study gives the following contribution:

- This paper introduces basic knowledge of metaheuristic techniques.
- The categorization of metaheuristic techniques and the list of algorithms that comes under this category are given.
- The literature includes key information about wrapper feature selection strategies and assessment measures.
- It describes the obstacles and difficulties in creating an algorithm for feature selection problems. Additionally, it provides the assessment metrics formula for examining performance.
- To improve the research effort, the research limitations as well as the suggestions for further studies are also presented.

The arrangement of this paper is as follows: Sect. 2 summarize the metaheuristic and feature selection preliminaries. In Sect. 3, the problems and difficulties are discussed. Research gap and future work focusing on the wrapper feature approach are suggested in Sect. 4, and the conclusion is highlighted in Sect. 5.

2 Background

This section provides a thorough explanation of the feature selection using a statistical model, as well as definitions, conceptual frameworks, and categorizations of metaheuristic algorithms.

2.1 Feature Selection

To deal with irrelevant, inappropriate, and unnecessary features, feature selection methods are used. It is a method for extracting the best attributes from datasets [3]. One of the most important and difficult issues in machine learning is feature selection. Feature selection is used in various fields like it is used as biomedical problems to searching the best available gene from the candidates [4], image processing [5], text mining [6], etc. The framework of the feature selection method is shown in Fig. 1. The data set, feature subset selection, evaluation metrics, appropriate feature evaluation, and verification are the five essential parts of the feature selection process.

To find the optimal set of features, various feature selection techniques have been developed. The methods are typically divided into three groups: filter, wrapper, and embedding methods [8, 9]. The filter method works independently as compared to classification or learning algorithms because it always focuses on the basic characteristics of the information [10]. In wrapper techniques, the classification method is always present, and it communicates with the classifier. In comparison to filter approaches, these methods are more computationally costly and produce better results. When combine above both techniques embedded method forms. In embedded techniques, the selection of features is the basic phase of the training procedure, which is carried out alongside the classifier [11].

Wrapper-based feature selection method provides better outcomes as compared to the filter approach. Wrapper approaches are based on the modeling methodology in which each subset is created and then assessed. To generate a subset in this method, various search strategies have been used. These search strategies are categorized

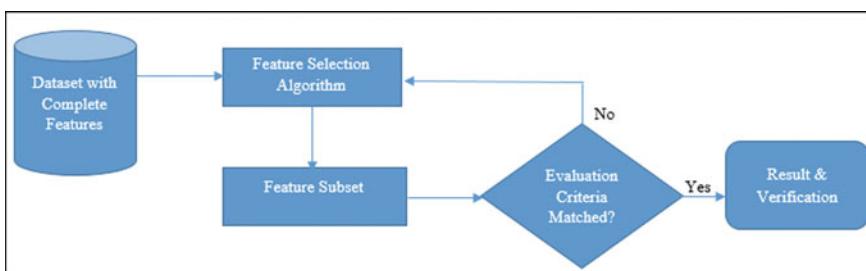


Fig. 1 Complete process of feature selection [7]

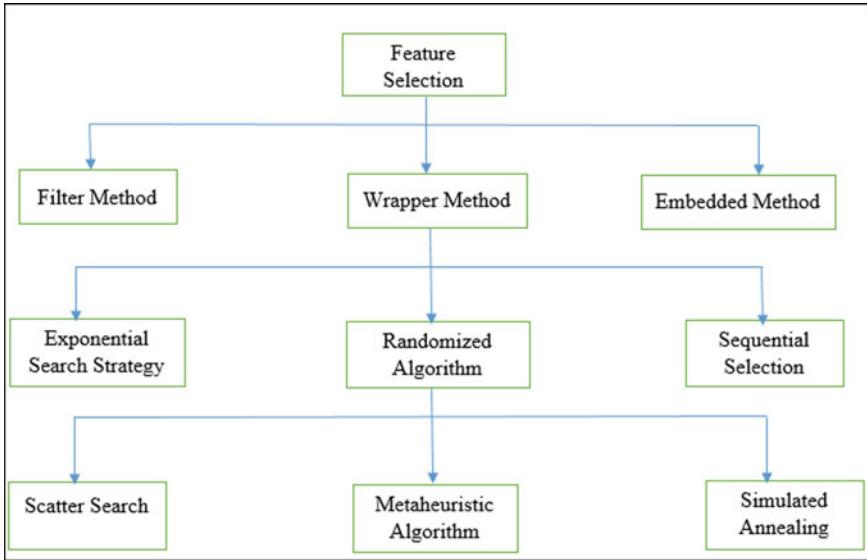


Fig. 2 Categorization of feature selection techniques [7]

into sequential, exponential, and randomized methods [12]. The sequential approach works as sequentially to add or delete features. After this, the feature has been added or deleted from the chosen subset because it cannot be modified in a way that results in local optima. The sequential approach includes floating forward selection or floating backward selection, best first, linear forward selection, and others. In the exponential technique, assessed features are increasing exponentially with feature size. Despite producing correct outcomes, this method is impractical to utilize due to its high computing cost [13, 14]. Randomness is used in randomized algorithms to avoid getting stuck in local optima and to help them in expanding the search space. Randomized algorithms such as random generation, simulated annealing, and other metaheuristic techniques are frequently referred to as population-based techniques [15].

We do not provide a thorough explanation of each technique used in the process of the feature selection process. The explanation of these techniques is available in [16]. Figure 2 displays a flow chart that classifies the approaches which solve the feature selection problem.

2.2 *Metaheuristic Algorithms*

The metaheuristic technique is a way to resolve optimization issues by finding the best (or very best) answer. These methods are simple, adaptable, and capable of avoiding local optima [17]. They are also derivative-free approaches. Metaheuristic algorithms

exhibit stochastic behavior; they begin the optimization method by producing random solutions. In contrast to gradient search approaches, this does not need the computation of the search space derivative. Due to their simple design and easy execution, metaheuristic algorithms are adaptable and customizable. The primary characteristic of metaheuristic methodology is its exceptional capacity to delay the algorithms' premature convergence. These methodologies perform as a "black box," to avoid local optima, and efficiently explore the search space because of the stochastic nature of algorithms. Exploration and exploitation, the two key components, are traded off by the algorithms [18, 19]. These algorithms extensively explore the promising search space during the exploration stage, and the exploitation stage involves the local search of any potential area that was discovered during the exploration stage. They have been successfully used to solve a variety of engineering and scientific issues, such as in power electronics to determine the best method for generating electricity, in manufacturing fields to plan work schedules, shipping, vehicle scheduling, and facility locations, and in geotechnical works to design bridges and buildings, in communication for radar design and networking, in data mining to find classification, clustering, prediction, and modeling, etc. The following two major categories best describe metaheuristic algorithms:

1. Metaheuristic algorithms using a single solution: these techniques begin their optimization process with a single solution and update it as they move through iterations. Furthermore, it doesn't fully explore the search space, and it could result in trapping within local optima.
2. Metaheuristic algorithms are based on a set of solutions: these algorithms start their optimization process by generating a set of solutions. With each generation or iteration, the set of solutions changes. As numerous solutions cooperate and have a great search space exploration, the algorithms help to avoid local optima. Additionally, they possess the ability to jump to the promising area of the search space.

Therefore, most issues in the actual world are solved using population-based strategies. Because of these qualities, metaheuristic techniques are highly regarded by researchers. Numerous algorithms have been developed to address various challenges. The metaheuristic algorithms are classified into four groups according to their behavior: swarm intelligence-based, evolution-based, physics-based, and human-related [20]. Figure 3 shows how the algorithms were categorized.

It is important to note that numerous metaheuristic techniques have been created since 1966. This work presents literature on such techniques that have been created or suggested in the last two decades. Various metaheuristic algorithms are shown in Tables 1, 2, 3, and 4 according to category. The table's first column displays the name of the method. The table's second column displays the abbreviation. The table's third column displays the algorithm developed year. These methods have been used to handle several practical problems; however, the scope of this study is limited to highlighting their use in feature selection issues. The algorithms that have been created and used to solve feature selection challenges in the last twenty years are listed here.

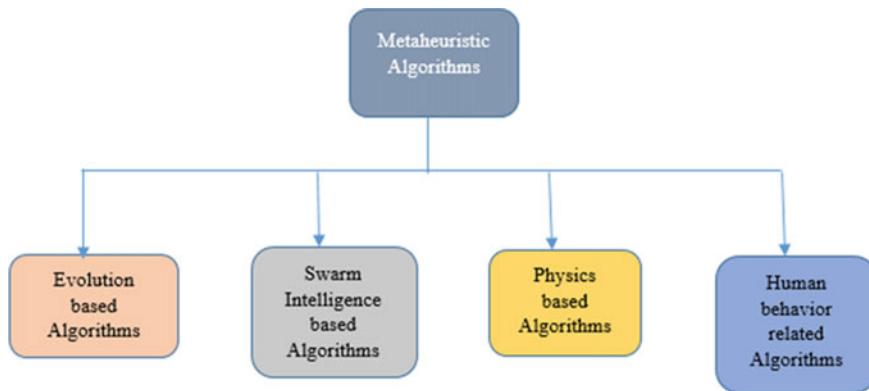


Fig. 3 Categorization of metaheuristic algorithms [7]

Table 1 List of evolution-based algorithms used in the last two decades for feature selection

Name of algorithm	Abbreviation	Year
Artificial immune algorithm	AIA	2002
Imperialist competitive algorithm	ICA	2007
Ecology inspired optimization	ECO	2011
Differential search algorithm	DS	2011
Flower pollination algorithm	FPA	2012
Back tracking search optimization	BSA	2013
stochastic fractal search	SFS	2014
Artificial algae algorithm	AAA	2015
Monkey king evolutionary	MKE	2016
synergistic fibroblast optimization	SFO	2018
Sun and leaf optimization	SLO	2018
Neural network algorithm	NNA	2018
Seasons optimization algorithm	SO	2020
Coronavirus optimization algorithm	COA	2020

Evolution-based Algorithms: It is taking inspiration from natural evolution and begins with a set of individuals that are generated at random. The greatest solutions in these algorithms are combined to produce new individuals. The optimum option is chosen while using crossover and mutation to create the new individuals. The genetic algorithm (GA), which is based on the Darwinian theory of evolution, is the most widely used in this category [21]. Other algorithms include differential evolution [22], tabu search [23], evolution strategy [24], genetic programming [25], etc.

Swarm Intelligence-based Algorithms: Algorithms of this category were influenced by the social interactions of many insects, fish, animals, and birds among

Table 2 List of swarm intelligence-based algorithms used in the last two decades for feature selection

Name of algorithm	Abbreviation	Year
Bacterial foraging optimization algorithm	BFOA	2002
Cat swarm optimization	CSO	2006
Artificial bee colony	ABC	2007
Bumblebees algorithm	B	2009
Paddy field algorithm	PFA	2009
Cuckoo search	CS	2009
Bat algorithm	BA	2010
Eagle strategy	ES	2010
Hierarchical swarm optimization	HSO	2010
Firefly algorithm	FA	2010
Fruit fly optimization algorithm	FOA	2011
Eco inspired evolutionary algorithm	ECO	2011
Flower pollination algorithm	FPA	2012
Artificial cooperative search algorithm	ACS	2012
The optbees algorithm	OptBees	2012
Wolf search algorithm	WSA	2012
Dolphin echolocation	DE	2013
Swallow swarm optimization algorithm	SSO	2013
Chicken swarm optimization	CSO	2014
Grey wolf optimization	GWO	2014
Ant lion optimizer	ALO	2015
Bird swarm algorithm	BSA	2015
Dragonfly algorithm	DA	2015
Dolphin swarm optimization algorithm	DSOA	2016
Crow search algorithm	CSA	2016
Whale optimization algorithm	WOA	2016
Grasshopper optimization algorithm	GOA	2017
Spotted hyena optimizer	SHO	2017
Emperor penguin optimizer	EPO	2018
Squirrer search algorithm	SSA	2018
Butterfly optimization algorithm	BOA	2019
Emperor penguin colony	EPC	2019

(continued)

Table 2 (continued)

Name of algorithm	Abbreviation	Year
Manta ray foraging optimization algorithm	MRFOA	2020
Black widow optimization algorithm	BWOA	2020
Tunicate swarm algorithm	TSA	2020
African vultures optimization algorithm	AVOA	2021
Red colobuses monkey	RCM	2021
Artificial gorilla troops optimizer	GTO	2021
Dingo optimizer	DOX	2021
Honey badger algorithm	HBA	2022

others. The widely used method is particle swarm optimization (PSO), invented by Kennedy and Eberhart [26]. It takes cues from a flock of birds' behavior which includes flying around a search area until they locate the optimal spot (position). Swarm intelligence techniques include ant colony optimization [27], honeybee swarm optimization method [28], monkey optimization [29], etc.

Physics-based Algorithms: These type of algorithms influenced by the laws of physics that govern the natural world. Physics-based algorithms include simulated annealing [30], harmony search [31], and others.

Human Behavior-Related Algorithms: These methods are mainly influenced by how people behave. Every human has a unique manner of carrying out tasks that has an impact on how well they succeed. This inspires academics to create algorithms. The league championship method [32], teaching learning-based optimization technique (TLBO) [33], and others are well-known algorithms.

Based on the existing literature review, FS-NBGSK technique performs better as compared to other techniques with twenty-two standard datasets from the UCI repository. Foundation of the FS-NBGSK technique is the concept of human gaining and sharing. By integrating binary junior and senior gaining and sharing stages [20], Agrawal et al. [34] introduced GSK approach to address the issue of feature selection available in binary form. A KNN classifier was used to test the FS-NBGSK algorithm.

FS-NBGSK technique provides better performance on the following basis: work on lowest average fitness values with excellent accuracy and a limited number of extracted features. It takes less calculation time than others [7]. Furthermore, it solves binary space concerns with a novel initialization, introduces the working mechanisms with two stages such as beginners-intermediate and intermediate-expert gaining-sharing stages, and modifies the dimensions of the stages. Additionally, it prevents the algorithm from early convergence and being trapped in local optima.

Table 3 List of physics-based algorithms used in the last two decades for feature selection

Name of algorithm	Abbreviation	Year
Small world optimization algorithm	SWOA	2006
Big bang big crunch	BBBC	2006
Central force optimization	CFO	2007
Gravitational search algorithm	GSA	2009
Charged system search	CSS	2010
Galaxy based search algorithm	GbSA	2011
Electro magnetism optimization	EMO	2011
Artificial chemical reaction optimization algorithm	ACROA	2011
Black hole algorithm	BH	2012
Water cycle algorithm	WCA	2012
Curved space optimization	CSO	2012
Ray optimization	RO	2012
Mine blast algorithm	MBA	2013
Atmosphere clouds model	ACMO	2013
Kinetic gas molecule optimization	KGMO	2014
Colliding bodies optimization	CBO	2014
Weighted super position attraction	WSA	2015
Lightning search algorithm	LSA	2015
Sine cosine algorithm	SCA	2016
Water evaporation optimization	WEO	2016
Multi-verse optimizer	MVO	2016
Electro-search algorithm	ES	2017
Thermal exchange optimization	TEO	2017
Find fix finish exploit analyze	F3EA	2019
Projectiles optimization	PRO	2020
Gradient-based optimizer	GBO	2020
Levy flight distribution	LFD	2020
Solar system algorithm	SSA	2020
Material generation algorithm	MGA	2021
Crystal structure algorithm	CryStAl	2021

3 Issue and Challenges

Despite having a significant deal of success using metaheuristic algorithms to solve problems of feature selection some difficulties and challenges also arise which are discussed as follows:

Table 4 List of human behavior-related algorithms used in the last two decades for feature selection

Name of algorithm	Abbreviation	Year
Harmony search	HS	2001
League championship algorithm	LCA	2009
Human inspired algorithm	HIA	2009
Social emotional optimization	SEOA	2010
Brain storm optimization	BSO	2011
Teaching learning based optimization	TLBO	2011
Exchange market algorithm	EMA	2014
Group counseling optimization	GCO	2014
Jaya algorithm	JA	2016
Volleyball premier league algorithm	VPL	2018
Supply demand-based optimization	SDO	2019
Group teaching optimization algorithm	GTOA	2020
Dynastic optimization algorithm	DOA	2020
Search and rescue optimization algorithm	SAR	2020
Tiki-taka algorithm	TTA	2021
Cooperation search algorithm	CSA	2021

- **Scalability and Stability:** Scalability is a requirement for the suggested technique to handle huge datasets in feature selection problems. A highly scalable classifier that can handle a huge dataset must be part of the algorithm's design [35]. Stability is another important consideration when the algorithm is developing. If a feature selection technique consistently identifies the same feature subset across many samples of datasets, it is considered to be stable. Most of the time, the feature selection method gets unstable when struggling to find the optimum categorization. Instability occurs when features with a high correlation are removed to achieve the best accuracy of classification. So, scalability and stability are just as crucial as classification accuracy.
- **Choice of Classifier:** the selection of a classifier is critical in the design of a wrapper feature selection algorithm since it greatly affects the quality of the results.
- **Construction of Objective Function:** To resolve the feature selection issue, the multi-objective function is built to merge the two conflicting objectives. Multi-objective function is used to improve the fitness function that is very productive and useful.
- **Evaluation Criteria to Check the Performance:** The effectiveness of the wrapper feature selection method has been examined in the literature using a variety of evaluation measures. Most common evaluation measures are Sensitivity/ Recall/ True positive rate [36–38], Specificity / True negative rate (TNR) [36, 38], Precision/ Positive predictive value [37, 39], F-score [38, 40], etc. Based on these parameters, African vultures [41], red colobuses monkey [42], artificial gorilla [43], dingo optimizer [44], honey badger [45] algorithms, material generation

[46], crystal structure algorithm [47], tiki-taka [48] and cooperation search [49] algorithms for solving feature selection problem are also exist in the literature.

4 Research Gap

The shortcomings found in this work are highlighted as there are very few evolutions-based and human behavior-related techniques have been developed. Based on human activities and natural evolution, new algorithms can be created. Different components of meta-heuristics optimization techniques require theoretical and mathematical foundations such as the capacity to search locally versus globally, exploration versus exploitation, and convergence. Additionally, metaheuristic algorithms exhibit a delayed convergence rate because of the random generating movement, and without being aware of the search direction, they explore the search region. The parameter values used in the metaheuristic algorithms must be changed because they might converge too soon or become stuck in local optima.

5 Conclusion

This paper provides a comprehensive review of metaheuristic techniques that are developed in the last two decades to solve feature selection challenges. To resolve feature selection challenges, researchers may find it easy to understand if they have access to a thorough explanation of the mathematical model and its various techniques. As a result, the fundamental definition, significance, and classification of metaheuristic algorithms are provided. The review addressed several issues, including the categories of metaheuristic and how it influences datasets and classification accuracy. In order to provide a broad description of the research for selecting techniques, this paper focuses on recent advancements in the specific areas that need to be addressed to provide improved answers to select features. This study is required because there is a lack of precise information and complexities on feature selection techniques, which affects the precision, applicability, and overall effectiveness of predictive models. Hence, determining the best methods to select meaningful and important features from a specific situation using a metaheuristic optimization approach requires understanding the impact, effect, and significance of selecting features in classification as well as researching, examining, and perusing the existing literature to understand where each technique stands now.

References

1. Sureka V, Sudha L, Kavya G, Aruna KB (2020) Nature inspired meta-heuristic optimization algorithms capitalized. In: 2020 6th International conference on advanced computing and communication systems (ICACCS), pp 1029–1034. IEEE
2. Dokeroglu T, Sevinc E, Kucukyilmaz T, Cosar A (2019) A survey on new generation metaheuristic algorithms. *Comput Ind Eng* 137:106040
3. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans Knowl Data Eng* 17(4):491–502
4. Ahmed S, Zhang M, Peng L (2013) Enhanced feature selection for biomarker discovery in LC-MS data using GP. In: 2013 IEEE congress on evolutionary computation, pp 584–591. IEEE
5. Ghosh A, Datta A, Ghosh S (2013) Self-adaptive differential evolution for feature selection in hyperspectral image data. *Appl Soft Comput* 13(4):1969–1977
6. Aghdam MH, Ghasem-Aghaee N, Basiri ME (2009) Text feature selection using ant colony optimization. *Expert Syst Appl* 36(3):6843–6853
7. Agrawal P, Abutarboush HF, Ganesh T, Mohamed AW (2021) Metaheuristic algorithms on feature selection: a survey of one decade of research (2009–2019). *IEEE Access* 9:26766–26791
8. Dash M, Liu H (1997) Feature selection for classification. *Intell Data Anal* 1(1–4):131–156
9. Hoque N, Bhattacharyya DK, Kalita JK (2014) MIFS-ND: a mutual information-based feature selection method. *Expert Syst Appl* 41(14):6371–6385
10. Xu Z, King I, Lyu MRT, Jin R (2010) Discriminative semi-supervised feature selection via manifold regularization. *IEEE Trans Neural Netw* 21(7):1033–1047
11. Tang J, Aleyani S, Liu H (2014) Feature selection for classification: a review. In: Data classification: algorithms and applications, p 37
12. Jović A, Brkić K, Bogunović N (2015) A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp 1200–1205. IEEE
13. Sun Z, Bebis G, Miller R (2004) Object detection using feature subset selection. *Pattern Recogn* 37(11):2165–2176
14. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell* 22(1):4–37
15. Liu H, Motoda H (eds) (1998) Feature extraction, construction and selection: a data mining perspective, vol 453. Springer Science & Business Media
16. Xue B, Zhang M, Browne WN, Yao X (2015) A survey on evolutionary computation approaches to feature selection. *IEEE Trans Evol Comput* 20(4):606–626
17. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. *Adv Eng Softw* 69:46–61
18. Olorunda O, Engelbrecht AP (2008) Measuring exploration/exploitation in particle swarms using swarm diversity. In: 2008 IEEE congress on evolutionary computation (IEEE world congress on computational intelligence), pp 1128–1134
19. Lin L, Gen M (2009) Auto-tuning strategy for evolutionary algorithms: balancing between exploration and exploitation. *Soft Comput* 13(2):157–168
20. Mohamed AW, Hadi AA, Mohamed AK (2020) Gaining-sharing knowledge based algorithm for solving optimization problems: a novel nature-inspired algorithm. *Int J Mach Learn Cybern* 11(7):1501–1529
21. Holland JH (1992) Genetic algorithms. *Sci Am* 267(1):66–73
22. Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Global Optim* 11(4):341–359
23. Glover F (1986) Future paths for integer programming and links to artificial intelligence. *Comput Oper Res* 13(5):533–549
24. Rechenberg I (1978) Evolutionsstrategien. In: *Simulationsmethoden in der Medizin und Biologie*. Springer, Berlin, pp 83–114
25. Holland JH (1992) *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press (1992)

26. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, vol 4, pp 1942–1948
27. Dorigo M, Maniezzo V, Colorni A (1996) Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern Part B (Cybern)* 26(1):29–41
28. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization, vol 200, pp 1–10. Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department
29. Yi TH, Li HN, Zhang XD (2012) A modified monkey algorithm for optimal sensor placement in structural health monitoring. *Smart Mater Struct* 21(10):105033
30. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
31. Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. *Simulation* 76(2):60–68
32. Kashan AH (2009) League championship algorithm: a new algorithm for numerical function optimization. In: 2009 international conference of soft computing and pattern recognition, pp 43–48. IEEE
33. Rao RV, Savsani VJ, Vakharia DP (2012) Teaching–learning-based optimization: an optimization method for continuous non-linear large scale problems. *Inf Sci* 183(1):1–15
34. Agrawal P, Ganesh T, Mohamed AW (2021) A novel binary gaining–sharing knowledge-based optimization algorithm for feature selection. *Neural Comput Appl* 33(11):5989–6008
35. Bolón-Canedo V, Rego-Fernández D, Peteiro-Barral D, Alonso-Betanzos A, Guijarro-Berdíñas B, Sánchez-Marcano N (2018) On the scalability of feature selection methods on high-dimensional data. *Knowl Inf Syst* 56(2):395–442
36. Ewees AA, El Aziz MA, Hassanien AE (2019) Chaotic multi-verse optimizer-based feature selection. *Neural Comput Appl* 31(4):991–1006
37. Belazzoug M, Touahria M, Nouioua F, Brahimi M (2020) An improved sine cosine algorithm to select features for text categorization. *J King Saud Univ Comput Inf Sci* 32(4):454–464
38. Han X, Chang X, Quan L, Xiong X, Li J, Zhang Z, Liu Y (2014) Feature subset selection by gravitational search algorithm optimization. *Inf Sci* 281:128–146
39. Zhang K, Yuan Z, Yang T, Lu Z, Cao Q, Tian Y, Zhu Y, Cao W, Liu X (2020) Chlorophyll meter-based nitrogen fertilizer optimization algorithm and nitrogen nutrition index for in-season fertilization of paddy rice. *Agron J* 112(1):288–300
40. Martínez-Álvarez F, Asencio-Cortés G, Torres JF, Gutiérrez-Avilés D, Melgar-García L, Pérez-Chacón R, Rubio-Escudero C, Riquelme JC, Troncoso A (2020) Coronavirus optimization algorithm: a bioinspired metaheuristic based on the COVID-19 propagation model. *Big Data* 8(4):308–322
41. Abdollahzadeh B, Gharehchopogh FS, Mirjalili S (2021) African vultures optimization algorithm: A new nature-inspired metaheuristic algorithm for global optimization problems. *Comput Ind Eng* 158:107408
42. AL-kubaisy WJ, Yousif M, Al-Khateeb B, Mahmood M, Le DN (2021) The red colobuses monkey: a new nature–inspired metaheuristic optimization algorithm. *Int J Comput Intell Syst* 14(1):1108–1118
43. Abdollahzadeh B, Soleimanian Gharehchopogh F, Mirjalili S (2021) Artificial gorilla troops optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *Int J Intell Syst* 36(10):5887–5958
44. Bairwa AK, Joshi S, Singh D (2021) Dingo optimizer: a nature-inspired metaheuristic approach for engineering problems. *Math Prob Eng*
45. Hashim FA, Houssein EH, Hussain K, Mabrouk MS, Al-Atabany W (2022) Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems. *Math Comput Simul* 192:84–110
46. Talatahari S, Azizi M, Gandomi AH (2021) Material generation algorithm: a novel metaheuristic algorithm for optimization of engineering problems. *Processes* 9(5):859
47. Talatahari S, Azizi M, Tolouei M, Talatahari B, Sareh P (2021) Crystal structure algorithm (CryStAl): a metaheuristic optimization method. *IEEE Access* 9:71244–71261

48. Zamli KZ, Kader A, Din F, Alhadawi HS (2021) Selective chaotic maps Tiki-Taka algorithm for the S-box generation and optimization. *Neural Comput Appl* 33(23):16641–16658
49. Feng ZK, Niu WJ, Liu S (2021) Cooperation search algorithm: A novel metaheuristic evolutionary intelligence algorithm for numerical optimization and engineering optimization problems. *Appl Soft Comput* 98:106734

New Type of Degenerate Changhee–Genocchi Polynomials of the Second Kind



Azhar Iqbal, Waseem A. Khan, and Mohd Nadeem

Abstract The λ -analogue type degenerate Changhee–Genocchi numbers and polynomials of the second kind are presented. In this paper, the proposed polynomials are not the same from the previously considered degenerate Changhee–Genocchi numbers and polynomials. Properties of present numbers and polynomials are investigated. In addition, we provide a new identities and relations between the proposed polynomials.

Keywords Degenerate Changhee–Genocchi numbers and polynomials · Higher-order degenerate Changhee–Genocchi polynomials and numbers

1 Introduction

Carlitz began by studying degenerate numbers and polynomials, and these polynomials are connected to Bernoulli and Euler numbers and polynomials [1, 2]. Degenerate polynomials based on special polynomials have been presented by a number of researchers in various of fields [3]. Recently, Kim et al. [3–6], Khan et al. [7–13], Sharma et al. [14, 15], Muhiuddin et al. [16–18] introduced the new identities of degenerate special numbers and polynomials. These polynomials are developed with non-differential equations. The identities and technical method are helpful for problems in mathematical physics.

A. Iqbal (✉) · W. A. Khan

Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, Al Khobar 31952, Kingdom of Saudi Arabia
e-mail: aiqbal@pmu.edu.sa

W. A. Khan

e-mail: wkhan1@pmu.edu.sa

M. Nadeem

Department of Natural and Applied Sciences, Glocal University, Saharanpur, Uttar Pradesh 247121, India

The generalized Euler polynomials of order r in the case of $r \in \mathbb{C}$ are defined by (see [1, 8])

$$\left(\frac{2}{e^\omega + 1} \right)^r e^{\xi\omega} = \sum_{v=0}^{\infty} \mathbb{E}_v^{(r)}(\xi) \frac{\omega^v}{v!} \quad |\omega| < \pi. \quad (1.1)$$

When $\xi = 0$, $\mathbb{E}_v^{(r)} = \mathbb{E}_v^{(r)}(0)$.

The Changhee–Genocchi polynomials are represented as follows [19]:

$$\frac{2 \log(1 + \omega)}{2 + \omega} (1 + \omega)^\xi = \sum_{v=0}^{\infty} CG_v(\xi) \frac{\omega^v}{v!}. \quad (1.2)$$

When $\xi = 0$, $CG_v = CG_v(0)$.

Carlitz [1] studied the degenerate Euler polynomials.

$$\frac{2}{(1 + \lambda\omega)^{\frac{1}{\lambda}} + 1} (1 + \lambda\omega)^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} \mathbb{E}_{v,\lambda}(\xi) \frac{\omega^v}{v!} \quad (\lambda \in \mathbb{R}). \quad (1.3)$$

When $\xi = 0$, $\mathbb{E}_{v,\lambda} = \mathbb{E}_{v,\lambda}(0)$. The higher-order degenerate Euler polynomials are represented as follows [1]:

$$\left(\frac{2}{(1 + \lambda\omega)^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda\omega)^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} \mathbb{E}_{v,\lambda}^{(r)}(\xi) \frac{\omega^v}{v!}. \quad (1.4)$$

Note that $\lim_{\lambda \rightarrow 0} \mathbb{E}_{v,\lambda}^{(r)} = \mathbb{E}_v^{(r)}(0)$ ($v \geq 0$), where $\mathbb{E}_v^{(r)}(\xi)$ are the higher-order Euler polynomials.

Using (1.1) and (1.4), we see

$$\begin{aligned} \sum_{v=0}^{\infty} \lim_{\lambda \rightarrow 0} \mathbb{E}_{v,\lambda}^{(r)}(\xi) \frac{\omega^v}{v!} &= \lim_{\lambda \rightarrow 0} \left(\frac{2}{(1 + \lambda\omega)^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda\omega)^{\frac{\xi}{\lambda}} \\ &= \left(\frac{2}{e^\omega + 1} \right)^r e^{\xi\omega} = \sum_{v=0}^{\infty} \mathbb{E}_v^{(r)}(\xi) \frac{\omega^v}{v!}. \end{aligned} \quad (1.5)$$

Thus, by (1.4), we can written as

$$\lim_{\lambda \rightarrow 0} \mathbb{E}_{j,\lambda}^{(r)}(\xi) = \mathbb{E}_j^{(r)}(\xi), \quad (j \geq 0). \quad (1.6)$$

The degenerate Genocchi polynomials $\mathbb{G}_v(\xi; \lambda)$ are studied as follows [20, 21]:

$$\frac{2\omega}{e_\lambda(\omega) + 1} e_\lambda^\xi(\omega) = \sum_{v=0}^{\infty} \mathbb{G}_v(\xi; \lambda) \frac{\omega^v}{v!}. \quad (1.7)$$

For $\xi = 0$, $\mathbb{G}_v(\lambda) = \mathbb{G}_v(0; \lambda)$, it is represented as the degenerate Genocchi numbers.

For $\lambda \in \mathbb{R}$, the logarithm function $\log_\lambda(1 + \omega)$ represents the inverse of the degenerate function $e_\lambda(\omega)$. The definition is as follows (see [22]):

$$\log_\lambda(1 + \omega) = \sum_{v=1}^{\infty} \lambda^{v-1} (1)_{v, 1/\lambda} \frac{\omega^v}{v!}. \quad (1.8)$$

It is simple to demonstrate this.

$$\lim_{\lambda \rightarrow 0} \log_\lambda(1 + \omega) = \sum_{v=1}^{\infty} (-1)^{v-1} \frac{\omega^v}{v!} = \log(1 + \omega).$$

Note that $e_\lambda(\log_\lambda(1 + \omega)) = \log_\lambda(e_\lambda(1 + \omega)) = 1 + \omega$.

The first kind degenerate Stirling numbers [5] are written by

$$\frac{1}{\sigma!} (\log_\lambda(1 + \omega))^\sigma = \sum_{v=\sigma}^{\infty} S_{1,\lambda}(v, \sigma) \frac{\omega^v}{v!} \quad (k \geq 0). \quad (1.9)$$

It can be seen that $\lim_{\lambda \rightarrow 0} S_{1,\lambda}(v, \sigma) = S_1(v, \sigma)$, where $S_1(v, \sigma)$ representing the Stirling numbers of the first kind as follows [2–12, 16, 17, 19, 20, 22–24]:

$$\frac{1}{\sigma!} (\log(1 + \omega))^\sigma = \sum_{v=\sigma}^{\infty} S_1(v, \sigma) \frac{\omega^v}{v!} \quad (k \geq 0).$$

The degenerate Stirling numbers of the second kind [6] are written as

$$\frac{1}{\sigma!} (e_\lambda(\omega) - 1)^\sigma = \sum_{v=\sigma}^{\infty} S_{2,\lambda}(v, \sigma) \frac{\omega^v}{v!} \quad (k \geq 0). \quad (1.10)$$

It is noticed that $\lim_{\lambda \rightarrow 0} S_{2,\lambda}(v, \sigma) = S_2(v, \sigma)$, where $S_2(v, \sigma)$ are called the Stirling numbers of the second kind as follows (see [14, 15, 17, 18, 21, 25–27]):

$$\frac{1}{\sigma!} (e^\omega - 1)^\sigma = \sum_{v=\sigma}^{\infty} S_2(v, \sigma) \frac{\omega^v}{v!} \quad (k \geq 0).$$

The second kind of degenerate Bernoulli polynomials are presented by [11]

$$\frac{\omega}{\log_\lambda(1+\omega)}(1+\omega)^\xi = \sum_{v=0}^{\infty} b_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \quad (1.11)$$

It's important to remember that $\lim_{\lambda \rightarrow 0} b_{v,\lambda}(\xi) = b_v(\xi)$, ($v \geq 0$).

Kim et al. [5] presented the modified type degenerate Daehee polynomials as

$$\frac{\log_\lambda(1+\omega)}{\omega}(1+\omega)^\xi = \sum_{v=0}^{\infty} \mathbb{D}_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \quad (1.12)$$

For $\xi = 0$, $\mathbb{D}_{v,\lambda} = \mathbb{D}_{v,\lambda}(0)$ are known as degenerate Daehee numbers.

The following polynomials of the second kind are represented by [19]

$$\frac{2 \log(1+\omega)}{(1+\lambda \log(1+\omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log(1+\omega))^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} J_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \quad (1.13)$$

For $\xi = 0$, $J_{v,\lambda} = J_{v,\lambda}(0)$ is representing the degenerate Changhee–Genocchi numbers of the second kind.

In 2020, Sharma et al. [14] considered the second type of degenerate Daehee polynomials are written as

$$\frac{\log_\lambda(1+\omega)}{(1+\lambda \log_\lambda(1+\omega))^{\frac{1}{\lambda}} - 1} (1 + \lambda \log_\lambda(1+\omega))^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} \mathbb{D}_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \quad (1.14)$$

When $\xi = 0$, $\mathbb{D}_{v,\lambda} = \mathbb{D}_{v,\lambda}(0)$ are called the second kind of the new type of Daehee numbers.

The λ -analogue type degenerate Changhee–Genocchi numbers and polynomials of the second kind are considered. These polynomials and numbers are different from the previously introduced degenerate Changhee–Genocchi numbers and polynomials, as the main subject of this article. In order to develop new properties with the λ -analogue type degenerate Changhee–Genocchi polynomials of the second kind and Carlitz's degenerate Euler polynomials, various characteristics of these numbers and polynomials are developed.

2 A New Type of Degenerate Changhee–Genocchi Polynomials of the Second Kind

For $\lambda \in \mathbb{R}$, we consider a new type of degenerate Changhee–Genocchi polynomials of the second kind which is defined by the generating function given below.

$$\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} J_{v,\lambda,2}(\xi) \frac{\omega^v}{v!}. \quad (2.1)$$

When $\xi = 0$, $J_{v,\lambda,2} = J_{v,\lambda,2}(0)$ are known as the second type of the new type of degenerate Changhee–Genocchi numbers, where $\lim_{\lambda \rightarrow 0} \log_\lambda(1 + \omega) = \log(1 + \omega)$.

By (1.13) and (2.1), we see

$$\begin{aligned} \sum_{v=0}^{\infty} \lim_{\lambda \rightarrow 0} J_{v,\lambda,2}(\xi) \frac{\omega^v}{v!} &= \lim_{\lambda \rightarrow 0} \frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \frac{2 \log(1 + \omega)}{(1 + \lambda \log(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \sum_{v=0}^{\infty} J_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \end{aligned} \quad (2.2)$$

As a result,

$$\lim_{\lambda \rightarrow 0} J_{v,\lambda,2}(\xi) = J_{v,\lambda}(\xi) \quad (v \geq 0).$$

Theorem 2.1 Consider $v \geq 0$, we write

$$J_{v,\lambda,2}(\xi) = \sum_{\eta=0}^{v-1} \sum_{\mu=0}^{\eta} \left(\frac{\lambda^{v-\eta-1} (1)_{v-\eta, 1/\lambda} \mathbb{E}_{\mu, \lambda}(\xi) v!}{(v-l)! \eta!} \right) S_{1, \lambda}(\eta, \mu).$$

Proof Using (1.2), (1.8), (1.9), and (2.1), we noticed that

$$\begin{aligned} &\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \log_\lambda(1 + \omega) \left(\sum_{\mu=0}^{\infty} \mathbb{E}_{\mu, \lambda}(\xi) \frac{1}{\mu!} (\log_\lambda(1 + \omega))^\mu \right) \\ &= \log_\lambda(1 + \omega) \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu, \lambda}(\xi) \sum_{v=\mu}^{\infty} S_{1, \lambda}(v, \mu) \frac{\omega^v}{v!} \\ &= \log_\lambda(1 + \omega) \sum_{v=0}^{\infty} \left(\sum_{\mu=0}^v \mathbb{E}_{\mu, \lambda}(\xi) S_{1, \lambda}(v, \mu) \right) \frac{\omega^v}{v!} \\ &= \left(\sum_{v=1}^{\infty} \lambda^{v-1} (1)_{v, 1/\lambda} \frac{\omega^v}{v!} \right) \left(\sum_{\eta=0}^{\infty} \left(\sum_{\mu=0}^{\eta} \mathbb{E}_{\mu, \lambda}(\xi) S_{1, \lambda}(\eta, \mu) \right) \frac{z^\eta}{\eta!} \right) \end{aligned}$$

$$= \sum_{v=1}^{\infty} \left(\sum_{\eta=0}^{v-1} \sum_{\mu=0}^{\eta} \left(\frac{\lambda^{v-\eta-1} (1)_{v-\eta, 1/\lambda} \mathbb{E}_{\mu, \lambda}(\xi)}{(v-l)! \eta!} \right) S_{1, \lambda}(\eta, \mu) \right) \omega^v, \quad (2.3)$$

yields the proof. \square

Theorem 2.2 Consider $v \geq 0$, we write

$$J_{v, \lambda, 2}(\xi) = \sum_{\mu=0}^{v-1} \mathbb{E}_{\mu, \lambda}(\xi) (\mu + 1) S_{1, \lambda}(v, \mu + 1).$$

Proof By (1.2), (1.9), and (2.1), we note that

$$\begin{aligned} & \frac{2 \log_{\lambda}(1 + \omega)}{(1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu, \lambda}(\xi) \frac{1}{\mu!} (\log_{\lambda}(1 + \omega))^{\mu+1} \\ &= \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu, \lambda}(\xi) \frac{\mu + 1}{(\mu + 1)!} (\log_{\lambda}(1 + \omega))^{\mu+1} \\ &= \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu, \lambda}(\xi) (\mu + 1) \sum_{v=\mu+1}^{\infty} S_{1, \lambda}(v, \mu + 1) \frac{\omega^v}{v!} \\ &= \sum_{v=1}^{\infty} \left(\sum_{\mu=0}^{v-1} \mathbb{E}_{\mu, \lambda}(\xi) (\mu + 1) S_{1, \lambda}(v, \mu + 1) \right) \frac{\omega^v}{v!}. \end{aligned} \quad (2.4)$$

In view of (2.1) and (2.4), the required result is obtained. \square

Theorem 2.3 Consider $v \geq 0$, we can write as

$$J_{v, \lambda, 2}(\xi) = \sum_{\eta=0}^v \sum_{\mu=0}^{\eta} \binom{v}{\eta} (\xi)_{\mu, \lambda} S_{1, \lambda}(\eta, \mu) J_{v-\eta, \lambda, 2}.$$

Proof By (1.8) and (2.1), we get

$$\begin{aligned} & \frac{2 \log_{\lambda}(1 + \omega)}{(1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \frac{2 \log_{\lambda}(1 + \omega)}{(1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{1}{\lambda}} - 1} \sum_{\mu=0}^{\infty} \left(\frac{\xi}{\lambda} \right)_{\mu} \lambda^{\mu} \frac{1}{\mu!} (\log_{\lambda}(1 + \omega))^{\lambda} \\ &= \left(\sum_{\eta=0}^{\infty} J_{\eta, \lambda, 2} \frac{z^{\eta}}{\eta!} \right) \left(\sum_{\mu=0}^{\infty} (\xi)_{\mu, \lambda} \sum_{\eta=0}^{\infty} S_{1, \lambda}(\eta, \mu) \frac{\omega^{\eta}}{\eta!} \right) \end{aligned}$$

$$\begin{aligned}
&= \left(\sum_{v=0}^{\infty} J_{v,\lambda,2} \frac{\omega^v}{v!} \right) \left(\sum_{\eta=0}^{\infty} \left(\sum_{\mu=0}^{\eta} (\xi)_{\mu,\lambda} S_{1,\lambda}(\eta, \mu) \right) \frac{\omega^{\eta}}{\eta!} \right) \\
&= \sum_{v=0}^{\infty} \left(\sum_{\eta=0}^v \sum_{\mu=0}^{\eta} \binom{v}{\eta} (\xi)_{\mu,\lambda} S_{1,\lambda}(\eta, \mu) J_{v-\eta,\lambda,2} \right) \frac{\omega^v}{v!}, \tag{2.5}
\end{aligned}$$

yields the proof. \square

Theorem 2.4 Consider $v \geq 0$, then

$$\mathbb{G}_{v,\lambda}(\xi) = \sum_{\mu=0}^v J_{\mu,\lambda,2}(\xi) S_{2,\lambda}(v, \mu).$$

Proof By substituting ω with $e_{\lambda}(\omega) - 1$ in (2.1) and using (1.7), we obtain

$$\begin{aligned}
\sum_{\mu=0}^{\infty} J_{\mu,\lambda,2}(\xi) \frac{1}{\mu!} (e_{\lambda}(\omega) - 1)^{\mu} &= \frac{2\omega}{(1 + \lambda\omega)^{\frac{1}{\lambda}} + 1} (1 + \lambda\omega)^{\frac{\xi}{\lambda}} \\
&= \sum_{v=0}^{\infty} \mathbb{G}_{v,\lambda}(\xi) \frac{\omega^v}{v!}. \tag{2.6}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sum_{\mu=0}^{\infty} J_{\mu,\lambda,2}(\xi) \frac{1}{\mu!} (e_{\lambda}(\omega) - 1)^{\mu} &= \sum_{\mu=0}^{\infty} J_{\mu,\lambda,2}(\xi) \sum_{\mu=v}^{\infty} S_{2,\lambda}(v, \mu) \frac{\omega^v}{v!} \\
&= \sum_{v=0}^{\infty} \left(\sum_{\mu=0}^v J_{\mu,\lambda,2}(\xi) S_{2,\lambda}(v, \mu) \right) \frac{\omega^v}{v!}, \tag{2.7}
\end{aligned}$$

yields the proof. \square

Theorem 2.5 For all $v \geq 0$, we write

$$\frac{1}{2} [J_{v,\lambda,2}(1) + J_{v,\lambda,2}] = \begin{cases} 0, & \text{if } v = 0, \\ \lambda^{v-1} (1)_{v,1/\lambda}, & \text{if } j \geq 1. \end{cases}$$

Proof The following result is obtained from Eq. (2.1),

$$\begin{aligned}
 2 \log_\lambda(1 + \omega) &= \left((1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1 \right) \sum_{v=0}^{\infty} J_{v,\lambda,2} \frac{\omega^v}{v!} \\
 &= \frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} \\
 &\quad - \sum_{v=0}^{\infty} J_{v,\lambda,2} \frac{\omega^v}{v!} = \sum_{v=1}^{\infty} (J_{v,\lambda,2}(1) + J_{v,\lambda,2}) \frac{\omega^v}{v!}. \quad (2.8)
 \end{aligned}$$

Alternatively,

$$2 \log_\lambda(1 + \omega) = 2 \sum_{\omega=1}^{\infty} \lambda^{\omega-1} (1)_{\omega,1/\lambda} \frac{\omega^v}{v!}, \quad (2.9)$$

yields the proof. \square

Theorem 2.6 Let $d \in \mathbb{N}$ and $v \geq 0$. Then

$$J_{v,\lambda,2}(\xi) = \sum_{\mu=0}^v d^{\mu-1} S_{1,\lambda}(v, \mu) \sum_{a=0}^{d-1} \beta_{\mu, \frac{\lambda}{d}, 2} \left(\frac{a + \xi}{d} \right).$$

Proof From (1.9) and (2.1), we get

$$\begin{aligned}
 &\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\
 &= \frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \sum_{a=0}^{d-1} (1 + \lambda \log_\lambda(1 + \omega))^{\frac{a+\xi}{\lambda}} \\
 &= \frac{1}{d} \left(\frac{2d \log_\lambda(1 + \omega)}{\left(1 + \frac{\lambda}{d} (d \log_\lambda(1 + \omega))^{\frac{d}{\lambda}} \right) + 1} \right) \sum_{a=0}^{d-1} \left(1 + \frac{\lambda}{d} (d \log_\lambda(1 + \omega))^{\frac{d(a+\xi)}{\lambda d}} \right) \\
 &= \frac{1}{d} \sum_{a=0}^{d-1} \sum_{\mu=0}^{\infty} J_{\mu, \frac{\lambda}{d}, 2} \left(\frac{a + \xi}{d} \right) \frac{1}{\mu!} (d \log_\lambda(1 + \omega))^{\mu} \\
 &= \frac{1}{d} \sum_{a=0}^{d-1} \sum_{\mu=0}^{\infty} J_{\mu, \frac{\lambda}{d}, 2} \left(\frac{a + \xi}{d} \right) \sum_{v=\mu}^{\infty} d^{\mu} S_{1,\lambda}(v, \mu) \frac{\omega^v}{v!} \\
 &= \sum_{v=0}^{\infty} \left(\sum_{\mu=0}^v \sum_{a=0}^{d-1} d^{\mu-1} J_{\mu, \frac{\lambda}{d}, 2} \left(\frac{a + \xi}{d} \right) S_{1,\lambda}(v, \mu) \right) \frac{\omega^v}{v!}, \quad (2.10)
 \end{aligned}$$

yields the proof. \square

For $r \in \mathbb{N}$, a new type of higher-order degenerate is introduced. The generating function gives the Changhee–Genocchi polynomials of the second kind.

$$\left(\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} = \sum_{v=0}^{\infty} J_{v,\lambda,2}^{(r)}(\xi) \frac{\omega^v}{v!}. \quad (2.11)$$

When $\xi = 0$, $J_{v,\lambda,2}^{(r)} = J_{v,\lambda,2}^{(r)}(0)$ are referred to as a new type higher-order degenerate Changhee–Genocchi numbers of the second kind.

We can see this in (2.11) as

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \sum_{v=0}^{\infty} J_{v,\lambda,2}^{(r)}(\xi) \frac{\omega^v}{v!} &= \lim_{\lambda \rightarrow 0} \left(\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \left(\frac{2 \log(1 + \omega)}{2 + \omega} \right)^r (1 + \omega)^{\xi} = \sum_{v=0}^{\infty} CG_v^{(r)}(\xi) \frac{\omega^v}{v!}, \end{aligned}$$

where $CG_v^{(r)}(\xi)$ represents the higher-order Changhee–Genocchi polynomials.

Theorem 2.7 *Let $v \geq 0$. Then*

$$J_{v,\lambda,2}^{(r)}(\xi) = \sum_{\mu=0}^{v-1} \mathbb{E}_{\mu,\lambda}^{(r)}(\mu + 1) S_{1,\lambda}(v, \mu + 1).$$

Proof From (1.4), (1.9), and (2.11), we note that

$$\begin{aligned} &\left(\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\ &= \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu,\lambda}^{(r)}(\xi) \frac{1}{\mu!} (\log_\lambda(1 + \omega))^{\mu+1} \\ &= \sum_{\mu=0}^{\infty} \mathbb{E}_{\mu,\lambda}^{(r)}(\xi) (\mu + 1) \sum_{v=\mu+1}^{\infty} S_{1,\lambda}(v, \mu + 1) \frac{\omega^\mu}{\mu!} \\ &= \sum_{v=1}^{\infty} \left(\sum_{\mu=0}^{v-1} \mathbb{E}_{\mu,\lambda}^{(r)}(\mu + 1) S_{1,\lambda}(v, \mu + 1) \right) \frac{\omega^v}{v!}, \end{aligned} \quad (2.12)$$

yields the proof. \square

Theorem 2.8 *Consider $v \geq 0$, we have*

$$J_{v,\lambda,2}^{(r)}(\xi) = \sum_{\eta=0}^v \binom{v}{\eta} J_{\eta,\lambda,2}^{(r-k)} J_{v-\eta,\lambda,2}^{(k)}(\xi). \quad (2.13)$$

Proof By (2.11), we get

$$\begin{aligned}
& \left(\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\
&= \left(\frac{\log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^{r-k} \left(\frac{\log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} - 1} \right)^k \\
&\quad (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} = \left(\sum_{\eta=0}^{\infty} J_{\eta, \lambda, 2}^{(r-k)} \frac{\omega^\eta}{\eta!} \right) \left(\sum_{\nu=0}^{\infty} J_{\nu, \lambda, 2}^{(k)}(\xi) \frac{\omega^\mu}{\mu!} \right) \\
&= \sum_{\nu=0}^{\infty} \left(\sum_{\eta=0}^{\nu} \binom{\nu}{\eta} J_{\eta, \lambda, 2}^{(r-k)} J_{\nu-\eta, \lambda, 2}^{(k)}(\xi) \right) \frac{\omega^\nu}{\nu!}, \tag{2.14}
\end{aligned}$$

yields the proof (2.13). \square

Theorem 2.9 Let $\nu \geq 0$. Then

$$J_{\nu, \lambda, 2}^{(r)}(\xi) = \sum_{\eta=0}^{\nu} \sum_{\mu=0}^{\eta} \binom{\nu}{\eta} (\xi)_{\mu, \lambda} S_{1, \lambda}(\eta, \mu) J_{\nu-\eta, \lambda, 2}^{(r)}. \tag{2.15}$$

Proof From (1.9) and (2.11), we note that

$$\begin{aligned}
& \left(\frac{2 \log_\lambda(1 + \omega)}{(1 + \lambda \log_\lambda(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r (1 + \lambda \log_\lambda(1 + \omega))^{\frac{\xi}{\lambda}} \\
& \quad \left(\sum_{\nu=0}^{\infty} J_{\nu, \lambda, 2}^{(r)} \frac{\omega^\nu}{\nu!} \right) \left(\sum_{\eta=0}^{\infty} \left(\sum_{\mu=0}^{\eta} (\xi)_{\mu, \lambda} S_{1, \lambda}(\eta, \mu) \right) \frac{\omega^\eta}{\eta!} \right) \\
&= \sum_{j=0}^{\infty} \left(\sum_{\eta=0}^{\nu} \sum_{\mu=0}^{\eta} \binom{\nu}{\eta} (\xi)_{\mu, \lambda} S_{1, \lambda}(\eta, \mu) J_{\nu-\eta, \lambda, 2}^{(r)} \right) \frac{\omega^\nu}{\nu!}, \tag{2.16}
\end{aligned}$$

yields the proof. \square

Theorem 2.10 Let $\nu \geq 0$. Then

$$J_{\nu, \lambda, 2}^{(r)}(\xi + \eta) = \sum_{\eta=0}^{\nu} \sum_{\mu=0}^{\eta} \binom{\nu}{\eta} J_{\nu-\eta, \lambda}^{(r)}(\xi) (\eta)_{\mu, \lambda} S_{1, \lambda}(\eta, \mu). \tag{2.17}$$

Proof In (1.9) and (2.11), we see

$$\begin{aligned}
\sum_{v=0}^{\infty} J_{v,\lambda,2}^{(r)}(\xi + \eta) \frac{\omega^v}{v!} &= \left(\frac{2 \log_{\lambda}(1 + \omega)}{(1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{1}{\lambda}} + 1} \right)^r \\
&\quad (1 + \lambda \log_{\lambda}(1 + \omega))^{\frac{\xi + \eta}{\lambda}} \\
&= \left(\sum_{l=0}^{\infty} J_{l,\lambda,2}^{(r)}(\xi) \frac{t^l}{l!} \right) \left(\sum_{k=0}^{\infty} \left(\sum_{m=0}^k (\eta)_{m,\lambda} S_{1,\lambda}(k, m) \right) \frac{z^k}{k!} \right) \\
&= \sum_{j=0}^{\infty} \left(\sum_{\eta=0}^v \sum_{\mu=0}^{\eta} \binom{v}{\eta} J_{v-\eta,\lambda}^{(r)}(\xi) (\eta)_{\mu,\lambda} S_{1,\lambda}(\eta, \mu) \right) \frac{\omega^v}{v!},
\end{aligned} \tag{2.18}$$

yields the proof. \square

3 Conclusion

The λ -analogue type degenerate Changhee–Genocchi numbers and polynomials of the second kind are constructed. A few properties of these numbers and polynomials are considered. The results are found in good agreement. In addition, we have given new identities and relations for the λ -analogue type degenerate Changhee–Genocchi polynomials of the second kind and Carlitz’s degenerate Euler polynomials.

References

1. Carlitz L (1979) Degenerate Stirling Bernoulli and Eulerian numbers. *Util Math* 15:51–88
2. Carlitz L (1956) A degenerate Staud–Clausen theorem. *Arch Math* 7:28–33
3. Kim T, Kim DS (2017) Degenerate Changhee numbers and polynomials of the second kind, vol 2021, pp 1–8. Article ID 7172054. [arXiv:1707.09721v1](https://arxiv.org/abs/1707.09721v1) [math. NT]
4. Kim BM, Jang L-C, Kim W, Kwon HI (2017) Degenerate Changhee–Genocchi numbers and polynomials. *J Inequal Appl* 294
5. Kim T, Kim DS, Kim H-Y, Kwon J (2020) Some results on degenerate Daehee and Bernoulli numbers and polynomials. *Adv Differ Equ* 311
6. Kim T (2018) A note on degenerate Stirling numbers of the second kind. *Proc Jangjeon Math Soc* 21(4):589–598
7. Khan WA, Acikgoz M, Duran U (2020) Note on the type 2 degenerate multi-poly-Euler polynomials. *Symmetry* 12. Article ID 1691
8. Khan WA, Haroon H (1920) Some symmetric identities for the generalized Bernoulli, Euler and Genocchi polynomials associated with Hermite polynomials. *Springer Plus* 2016:5
9. Khan WA, Nisar KS, Duran U, Acikgoz M, Araci S (2018) Multifarious implicit summation formulae of Hermite-based poly-Daehee polynomials. *Appl Math Inf Sci* 12(2):305–310
10. Khan WA, Muhiuddin G, Muhyi A, Al-Kadi D (2021) Analytical properties of type 2 degenerate poly-Bernoulli polynomials associated with their applications. *Adv Differ Equ* 420
11. Khan WA, Muhyi A, Ali R, Alzobydi KAH, Singh M, Agarwal P (2021) A new family of degenerate poly-Bernoulli polynomials of the second kind with its certain related properties. *AIMS Math* 6(11):12680–12697

12. Khan WA, Ali R, Alzobydi KAH, Ahmed N (2021) A new family of degenerate poly-Genocchi polynomials with its certain properties. *J Funct Spaces* 2021:1–8. Article ID 6660517
13. Khan WA, Alatawi MS (2022) Analytical properties of degenerate Genocchi polynomials the second kind and some of their applications. *Symmetry* 14:1500
14. Sharma SK, Khan WA, Araci S, Ahmed SS (2020) New type of degenerate Daehee polynomials of the second kind. *Adv Differ Equ* 428
15. Sharma SK, Khan WA, Araci S, Ahmed SS (2020) New construction of type 2 of degenerate central Fubini polynomials with their certain properties. *Adv Differ Equ* 587
16. Muhiuddin G, Khan WA, Muhyi A, Al-Kadi D (2021) Some results on type 2 degenerate poly-Fubini polynomials and numbers. *Comput Model Eng Sci* 29(2):1051–1073
17. Muhiuddin G, Khan WA, Al-Kadi D (2021) Construction on the degenerate poly-Frobenius–Euler polynomials of complex variable. *J Funct Spaces* 2021:1–9. Article ID 3115424
18. Muhiuddin G, Khan WA, Al-Kadi D (2022) Some identities of the degenerate poly-Cauchy and unipoly Cauchy polynomials of the second kind. *Comput Model Eng Sci*. <https://doi.org/10.32604/cmes.2022.017272>
19. Kwon H-I, Kim T, Park JW (2016) A note on degenerate Changhee–Genocchi polynomials and numbers. *Glob J Pure Appl Math* 12(5):4057–4064
20. Lim D (2016) Some identities of degenerate Genocchi polynomials. *Bull Korean Math Soc* 53(2):569–579
21. Park JW, Jang GW, Kwon J (2017) The λ -analogue degenerate Changhee polynomials and numbers. *Glob J Pure Appl Math* 13(3):893–900
22. Kim DS, Kim T (2020) A note on a new type of degenerate Bernoulli numbers. *Russ J Math Phys* 27(2):227–235
23. Khan WA, Kamarujjama M (2021) Some identities on type 2 degenerate Daehee polynomials and numbers. *Indian J Math* 63(3):433–477
24. Khan WA, Kamarujjama M (2022) A note on type 2 degenerate multi-poly Bernoulli polynomials of the second kind. *Proc Jangjeon Math Soc* 25(1):59–68
25. Roman S (1984) The umbral calculus. Pure Appl Math 111. Academic Press, New York
26. Sharma SK (2020) A note on degenerate poly-Genocchi polynomials. *Int J Adv Appl Sci* 7(5):1–5
27. Alatawi MS, Khan WA (2022) New type of degenerate Changhee–Genocchi polynomials. *Axioms* 11:355

Temperature Aware Bi-partitioning Multi-level Logic Synthesis



Apangshu Das, Vivek Kumar Singh, and Sambhu Nath Pradhan

Abstract The integration of a large number of transistors and device scaling results in the development of high-power density within Very Large-Scale Integrated (VLSI) circuits. Power density is directly proportional to chip temperature and grows exponentially as package density increases. As a result, considering temperature impacts at all levels of the VLSI design cycle becomes a fundamental phenomenon because it reduces a circuit's performance and dependability. To decrease the power density of the circuit, a bi-partitioning heuristic approach based on Shannon's expansion is provided in this study to offset the effect of temperature. Following bi-partitioning, the co-factored sub-circuits are transformed into cutting-edge AND-Inverter Graphs (AIGs) based on multi-level logic. The MCNC benchmark synthesis results indicate a maximum improvement of 36% area and 21.51% power density over the espresso tool-based decomposition.

Keywords AIGs · Power density · Thermal aware · Shannon's expansion

1 Introduction

Temperature has become an important parameter in modern electronic devices. As devices are becoming portable and wireless, they demand low area and power utilization for hardware design. To address the issue of area, the circuit functions are performed as multi-level logic, and various low-power techniques are used for high-power requirements. Over the last two decades, significant progress has been made in establishing thermal models and computational strategies for temperature reduction at the micro-architectural level [1, 2]. Due to considerable temperature generation, uneven scaling has caused a sharp increase in power density and, as a result, hotspot problems in current generation chips. Peak temperatures can reduce performance and lifetime or even damage the chip. According to [3], increasing the temperature by 10–15 °C will reduce the average lifetime of integrated circuits (ICs) by half. To

A. Das · V. K. Singh (✉) · S. N. Pradhan

Department of Electronics and Communication Engineering, NIT, Agartala, India

e-mail: erviveksingh77@gmail.com

reduce the temperature of the die at the physical design level, a number of thermally conscious approaches have been developed. However, each method has the potential to elevate the cost of the system by using a heat sink to limit any potential rise in the system's heat dissipation. For high-performance CPUs, cooling system costs are rising rapidly at a rate of 1–3 dollar or more per watt of power dissipation [4]. Design time thermal aware solutions can be created to mitigate the effects of the power and thermal properties of ICs, which would lower the cost of cooling. At the level of logic synthesis, thermal awareness approaches can be used more effectively. As a result, the logic level is crucial in circuit optimization to save cooling expenses. Prior studies focused on finding solutions to the portability issue that results from feature size growth, which significantly increases the overall power used by ICs. As a result, heat generation has a significant impact on the performance and efficiency of the circuits, which can occasionally result in self-heating burnout. As a result, power density, which closely correlates to temperature [5, 6], is a crucial parameter for reducing the thermal effect in VLSI circuits. In order to reduce the generation of heat, a circuit must be optimized using power density as a fitness criterion. The equation below can be used to determine the link between chip temperature and power density:

$$\text{Temp}_{\text{chip}} = \text{Temp}_{\text{Ambient}} + R_{\theta} \frac{\text{Power}_{\text{Total}}}{\text{Area}_{\text{Total}}} \quad (1)$$

The terms “ $\text{Temp}_{\text{chip}}$ ” and “ $\text{Temp}_{\text{Ambient}}$ ” in Eq. (1) refer to the average chip temperature and the ambient temperature, respectively ($\text{Temp}_{\text{Ambient}} = 45^{\circ}\text{C}$). R is the equivalent thermal resistance ($\text{m}^2 \text{ }^{\circ}\text{C}/\text{W}$) of the package, heat sink, and substrate (Si) layer. The total amount of power consumed is $\text{Power}_{\text{Total}}$ (in W). The total chip area is $\text{Area}_{\text{Total}}$ (in m^2). In order to reduce the generation of heat, a circuit must be optimized using power density as a fitness criterion. When a specific packaging method for Eq. (1) is targeted, it is found that the power and area change during logic synthesis. As a result, the only factor affecting IC temperature generation is power density.

We used power density as a metric for temperature because of this, and we also thought about power density minimization while choosing the variables for bipartitioning logic synthesis. The remaining manuscript is structured as follows. The Shannon Expansion and fundamental terminologies used in AIGs-based multi-level logic synthesis are illustrated in Sect. 2. The calculation of the objective parameter and fitness function is described in Sect. 3. Section 4 discusses the experimental findings, while Sect. 5 concludes with recommendations for further work.

2 Shannon Expansion and AIGs Basic Terminologies

This section includes a brief overview of Shannon's expansion as well as the fundamental terms and characteristics of AIGs [7, 8].

2.1 Shannon's Expansion

A Boolean function is expressed using Shannon's expansion as the sum of two sub-functions of the original function. The Shannon decomposition theorem is another name for it. According to Shannon's Theorem, the function $f(W_0, W_1, W_2, \dots, W_i, \dots, W_n)$ may be expressed as the sum of two terms, one with the particular variable (let's say, W_i) set to 0, and one with it set to 1 [9].

The mathematical expression for Shannon's expansion can be written as:

$$\begin{aligned}
 f(W_0, W_1, W_2, \dots, W_i, \dots, W_n) &= W_i \cdot f(W_1, W_2, \dots, W_{i=1}, \dots, W_n) \\
 &\quad + \overline{W_i} \cdot f(W_1, W_2, \dots, W_{i=0}, \dots, W_n) \\
 &= W_i \cdot f_{W_i} + \overline{W_i} \cdot f_{\overline{W_i}} \\
 &= W_i \cdot \text{Cofactor}_1 + \overline{W_i} \cdot \text{Cofactor}_2
 \end{aligned} \quad (2)$$

Here, W_i is the control variable, and f_{W_i} and $f_{\overline{W_i}}$ are called positive cofactor and negative cofactor with respect to variable W_i .

2.2 AND-Inverter Graphs Preliminaries

A flexible and effective Directed Acyclic Graph (DAG) representation of Boolean functions is the And-Inverter Graph (AIG). Primary inputs (PIs), two-input AND gates, inverters, and constants make up the realized graph. Primary inputs and constants serve as the representation for terminal nodes. The representation of nodes is an AND gate with two inputs and one output. An edge that has been complimented indicates inversion. To create shared AIGs, several AIGs may be employed simultaneously [8].

For node and logic level minimization in multi-level logic networks, the quick and greedy AIG rewriting method is employed. Rewriting often begins at the PIs and moves through the topological order to the principal outputs using k -feasible cuts (POs). The set of cuts for a PI only includes the trivial cut. The cuts (p) are computed by combining the cuts of 'x' and 'y' for an internal node 'p' with two fan-ins, 'x' and 'y', as shown in [8]:

$$\Psi(p) = \{\{p\}\} \cup \{u \cup v | u \in \Psi(x), u \in \Psi(y), |u \cup v| \leq k\} \quad (3)$$

Equation (3) demonstrates that combining two sets of cuts (u and v) keeps k -feasible cuts while adding the trivial cut of the node to the set of pair-wise unions of cuts belonging to the fan-in.

AIG Refactoring is a special type of rewriting used for larger cuts. For each AIG node, in a topological order, one K -feasible cut is computed with $10 \leq k \leq 20$.

Heuristically, the cut leaves are minimized to maximize the amount of re-convergent routes that are covered by the cut. Using BDDs, the cut's Boolean function is calculated and converted into a SOP. The SOP is taken into consideration, and the resulting AIG sub-graph is handled using the standard AIG rewriting [8].

AIG balancing is a method performed in one linear-time to elicit the Boolean network in topological order. The node uses the associative transformation to minimize the AIG levels as much as possible. For instance, three different AIG nodes are x , y , and z . Associativity is depicted in [8] as follows:

$$x(yz) = (xy)z = (xz)y \quad (4)$$

3 Objective Parameter Estimation and Fitness Calculation

The main objective of the suggested work is to reduce temperature. Since the absolute temperature of a system is unknown at the logic level, power density is used to estimate the temperature. Area occupied by logic is another objective function. For a logic level that is independent of technology, the area's absolute value is also unknown. Therefore, after AIG's transformation and any potential node reduction via AIG's rewriting, refactoring, and balancing approach, the number of node counts is taken into account to estimate the area.

3.1 Area Estimation

The sub-functions are translated into AIG after being decomposed into Shannon's bi-partitioning graphs with respect to a chosen input variable, and then the node minimization algorithm, such as rewriting, refactoring, and balancing, is used to achieve the reduced AIG structure. The number of AIG nodes is taken into account as the corresponding logic area in this work.

3.2 Temperature Estimation

In this study, power density is taken into account to evaluate a logic circuit's temperature. Total power dissipation is estimated, and the power density is determined by dividing total power by total area [9–11]. Switching activity is calculated to determine the power dissipation. The charging and discharging of load and parasitic capacitors causes switching power consumption. To determine the switching activity of the circuit, the SIS tool [12] is used. To determine how many switching transitions each node is anticipated to experience, the instructions "power_estimate" and "print_power" are used. Equation 5 is used to compute power density.

$$\text{Power - density} = \frac{\text{Total_Power}}{\text{Total_Area}} \quad (5)$$

3.3 Fitness Function Calculation

A variable's eligibility as an input for the Shannon expansion decomposition is determined by the fitness function. Equation 6 can determine the fitness of a specific input variable as in work [13].

$$\text{fitness}[\text{input_variable}(i)] = w_1 \frac{\text{area}(i)}{\text{Original_area}} + w_2 \frac{\text{power - density}(i)}{\text{Original_power - density}} \quad (6)$$

Equation 6 calculates the fitness of a specific input variable for the suggested method. Here, “area (i) and power density (i)” stand for the area and power density that a specific input variable, ‘ i ’ has incurred, respectively. The terms “Original area” and “Original power density” relate to the area that the original circuit occupied and the power density it produced. The goal functions’ weight factors are w_1 and w_2 , and the value of $w_1 + w_2 = 1$.

4 Results and Discussion

The suggested bi-partitioning for multi-level logic synthesis is implemented in C on a Pentium IV system with 3.4GHz clock frequency and 4 GB RAM memory using the LINUX platform. It is based on Shannon expansion and AIGs decomposition. The benchmark circuits for MCNC are used for the simulation. To determine the best solution for a temperature-aware circuit realization, a thorough search heuristic technique was used. Each variable is taken into account separately, and the resulting sub-circuits are further reduced using a multi-level realization based on AIGs. By using rewriting, refactoring, and balancing approaches, the AIGs are further diminished. Area and power consumption are evaluated following the removal of nodes from each level, and power density is determined by dividing total power consumption by total circuit area. By using the SIS tool, which takes into account the clock frequency of 20MHz and Vdd as 5 V, the power is determined. Table 1 compares the suggested method to the espresso-based circuit breakdown in terms of space and power density [14].

Table 1 demonstrates that the suggested optimum power density aware realization of bi-partitioning for multi-level logic synthesis improves in area and power density by 24.55% and 21.51%, respectively, over the existing original circuit. In comparison with other decomposed solutions, the best area aware realization shows a 36% area

Table 1 Area and power density analysis and comparison of proposed approach with respect to Espresso decomposed circuit realization

Benchmark circuits	w ₁ =0 w ₂ =1 (Best power density aware solution)		w ₁ =1 w ₂ =0 (Best area aware solution)		Optimal solution for area and power density		Max_delay		Espresso decomposed realization [14]	
	Area	Pow_Den	Area	Pow_Den	Area	Pow_Den	Area	Pow_Den	Area	Pow_Den
5xpl	596.47	76.7	559.88	82.12	559.88	82.12	0.36	678.59	96.57	0.32
alu2	2646.63	64.05	2583.16	72.57	2583.16	73.98	0.54	2611.2.2	136.31	0.63
apex42a	3997	96.12	3997	96.12	3997	96.12	0.57	13967.55	104.57	1.08
clip	229.73	77.86	229.73	77.86	229.73	77.86	0.57	507.9	102.74	0.58
cm82a	26.61	65.29	26.61	65.29	26.61	65.29	0.54	139.71	91.94	0.56
duke2	1350.52	84.51	868.19	96.57	868.19	96.57	0.58	2298.54	99.91	0.65
ex5	1470.27	98.5	705.2	109.36	705.2	109.36	0.59	2558	135.41	0.66
f51m	283.16	77.91	219.75	84.34	219.75	84.34	0.54	675.26	103.02	0.57
inc	683.16	67.93	559.81	80.44	543.04	82.3	0.38	728.48	112.25	0.39
misex1	329.94	73.02	276.51	77.66	239.71	78.91	0.54	385.86	103.7	0.57
pho2	1252.81	80.32	1169.65	85.14	1169.65	85.14	0.56	1270.69	119.33	0.6
pcler8	656.34	67.93	606.24	70.1	606.24	70.1	0.55	695.22	78.91	0.57
rd53	199.98	84.9	149.87	89.32	149.87	89.32	0.55	212.89	91.8	0.54
rd73	446.57	78.41	446.57	78.41	446.57	78.41	0.54	508.94	108.97	0.57
square5	249.89	73.98	243.24	75.94	243.24	75.94	0.53	352.6	95.66	0.56
tt2	975.47	97.08	898.96	98.04	898.96	98.04	0.43	1024.5	95.66	0.4
x2	299.38	96.57	166.53	73.42	166.53	73.42	0.51	299.38	96.57	0.55
Z5xpl	646.57	87.93	539.92	94.45	539.92	94.45	0.54	661.95	100.87	0.57
Average % improvement W.r.t. espresso	24.55	21.51	36.00	18.68	35.86	18.12	6.26			

and 18.68% power density improvement. To determine the best solution in terms of area and power density, all decompositions are compared using the fitness function to obtain the lowest cost factor and highest fitness value. The ideal solution for area and power density utilizing the proposed approach improves by 35.86% in area and 18.12% in power density. It is also found that the proposed solution reduces the maximum delay within the logic by 6.26%.

5 Conclusion and Future Work

Using Shannon's expansion, the circuit can be bi-partitioned by selecting the appropriate control variable. Shannon's expansion approach can be used to reduce circuit power and power density while increasing area overhead. The bi-partitioned sub-circuits are then transferred to AIGs-based multi-level circuit synthesis to reduce area overhead. Sub-circuits are reduced further using the rewriting, refactoring, and balancing methods. The suggested study addresses the issue of correct control variable selection for bi-partitioning logic synthesis sub-functions for temperature minimization. Based on the control input variable, the bi-partitioning Shannon's algorithm is used. AND-Inverter Graphs are used to represent the positive and negative co-factors. In order to minimize the area and its accompanying power, node reduction methods, rewriting, refactoring, and balancing are used. Power density is calculated by dividing total power by total area. Fitness is calculated for all control input variables in order to find the ideal control variable with the smallest area and power density. An exhaustive search heuristic technique for bi-partitioning logic synthesis based on Shannon's expansion is listed and tabulated. The representative temperature is reduced in this work by lowering the power density. By using physical design, this approach can be extended to acquire the absolute temperature of circuits.

References

1. Sankaranarayanan K, Velusamy S, Stan M, Skadron K (2005) A case for thermal-aware floor-planning at the microarchitectural level. *J Instr-Level Parallelism* 7(1):1–16
2. Huang W, Sankaranarayanan K, Skadron K et al (2008) Accurate, pre-RTL temperature-aware design using a parameterized, geometric thermal model. *IEEE Trans Comput* 57(9):1277–1288
3. Shang L, Dick RP (2006) Thermal crisis: challenges and potential solutions. *IEEE Potentials* 25(5):31–35
4. Gunther, Binns SF, Carmean DM, Hall JC (2001) Managing the impact of increasing microprocessor power consumption. *Intel Technol J* 5(1):1–9
5. Pedram M, Nazarian S (2006) Thermal modeling, analysis, and management in VLSI circuits: principles and methods. *Proc IEEE* 94(8):1487–1501
6. Das A, Hareesh YC, Pradhan SN (2020) NSGA-II based thermal-aware mixed dual reed-muller network synthesis using parallel tabular technique. *J Circ Syst Comput*
7. Kandasamy N, Ahmad F, Telagam N (2018) Shannon logic based novel QCA full adder design with energy dissipation analysis. *Int J Theor Phys* 57(12):3702–3715

8. Das A, Pradhan SN (2019) Thermal-aware output polarity selection based on and-inverter graph manipulation. *Recent Adv Electr Electron Eng* 12(1):30–39
9. Lavagno L (1995) Timed Shannon circuits: a power-efficient design style and synthesis tool. In: 32nd Design automation conference. IEEE, pp 254–260
10. Das A, Choudhury SR, Kumar BK, Pradhan SN (2012) An elitist area-power density trade-off in VLSI floorplan using genetic algorithm. In: 7th International conference on electrical and computer engineering. IEEE, pp 729–732
11. Das A, Kumar Singh V, Nath Pradhan S (2022) Shared reduced ordered binary decision diagram-based thermal-aware network synthesis. *Int J Circ Theory Appl* 50(6):2271–2286
12. Sentovich EM, Singh KJ, Lavagno L, Moon C, Murgai R, Saldanha A, Savoj H, Stephan PR, Brayton RK, Sangiovanni-Vincentelli A (1992) SIS: a system for sequential circuit synthesis
13. Singh VK, Sarkar T, Pradhan SN (2021) Power-aware testing for maximum fault coverage in analog and digital circuits simultaneously. *IETE Tech Rev* 1–15
14. Brayton RK, Hachtel GD, McMullen C, Sangiovanni-Vincentelli A (1984) Logic minimization algorithms for VLSI synthesis, vol 2. Springer Science & Business Media

Determination Human Behavior Prediction Supported by Cognitive Computing-Based Neural Network



Jyoti Parashar, Virendra Singh Kushwah, and Munishwar Rai

Abstract By the development of image processing systems, human behavior prediction has been grown for research. Many research outcomes are previously available to recognize human behavior based on different features in form of image sequences. The combination of human and machine learning is very interesting to determine human behavior without much complexity. Nevertheless, there was little have a look at concerning the mixture of presently identified conduct statistics with conduct prediction. The dataset is shaped with the aid of using very own captured pictures accompanied with the aid of using anger, happy, sad, disgust, etc. TensorFlow framework in conjunction with CNN version make this device higher for end result generation. Using numerous length of epoch defines accuracy of prediction and performances of the proposed device. In this work, cognitive neural community is used to make prediction easily and generate higher results.

Keywords Human behavior · Cognitive computing · Machine learning · Convolutional neural network

1 Introduction

Artificial Intelligence is what's to come. For the most part computerized reasoning is depicted as a biggest element amazing to impact the innovation. Cognitive machine learning is one of the sciences behind the biggest force of unreal insight and cognitive processing is the engine that propels the science. Perhaps the most encouraging utilization of cognitive machine learning is to improve association with the com-

J. Parashar

Dr. Akhilesh Das Gupta Institute of Technology Management, New Delhi, India

V. S. Kushwah (✉)

VIT Bhopal University, Kothrikalan, Sehore, Madhya Pradesh 466114, India

e-mail: virendra.kushwah@vitbhopal.ac.in

M. Rai

Maharishi Markandeshwar Deemed-to-be University, Ambala, India

e-mail: munishwar.rao@mmu.edu.in

puters followed by human behavior. Cognitive computing are frameworks that learn at scale, reason with reason and connect with people normally. It is a combination of software engineering and cognitive science—that is, the comprehension of the human mind and how it functions. Through self-instructing calculations that utilization information mining, visual acknowledgment, and regular language preparing, the computer can take care of issues and accordingly improve human cycles. This infers that association of computerized reasoning with people will request some capacity of something very similar to predict human behavior.

Previously, individuals' behavior was capricious on the grounds that there was no method for gathering and breaking down information that goes into the dynamic cycles. Presently with the passage of sophisticated computer framework and cloud which assists with putting away colossal measure of information supplement the advancement of cognitive machine deep learning. The conventional behavior model is to foster a factual or logical model of human behavior and to allocate a conveyance fitting to approve the model. The cognitive machine learning approach is very not the same as the ordinary model. Cognitive machine learning framework figures out how to predict the results by noticing the human behavior following by cognitive science.

The facial recognition (FR) framework is a more normal biometric data measure with preferable variety over some other strategy. Hence, FR has become a new theme in software engineering identified with biometrics and machine learning [3, 14]. In addition, the FR framework is suggesting huge advantages contrasted with other biometric security arrangements, for example, palm prints and fingerprints. The framework catches biometric estimations of an individual from a particular distance without cooperating with the individual. Accordingly, this innovation is turning out to be fundamental for various private structures and corporate associations. This procedure depends on the capacity to perceive a human face and afterward look at the changed features of the face with recently recorded appearances. This element additionally expands the significance of the framework and empowers it to be generally utilized across the world. It is created with easy to use features and tasks that incorporate unique nodal points of the face. There are around 80–90 remarkable nodal points of a face. From these nodal focuses, the FR framework estimates critical viewpoints including the distance between the eyes, length of the facial structure, state of the cheekbones, and profundity of the eyes.

Machine learning is a software engineering field that gives computers the capacity to learn without further explicit programming. The principle focal point of machine learning is giving calculations to preparing to play out an undertaking machine learning identified with the field of computational insights what's more, numerical streamlining. Machine learning incorporates different techniques, for example, support learning, regulated learning, nearly supervised learning, and unsupervised learning. Machine learning can be utilized on numerous errands that individuals figure no one but they can do, for example, messing around, learning subjects, and acknowledgment [9]. Contrasted with different models, artificial neural networks require an additional arrangement of specialized abilities and calculated information. The most significant of these specialized abilities is the capacity to utilize a profound learning system. A

deep learning structure speeds up the improvement interaction what's more, gives productive information handling, perception, and deployment tools [18].

2 Background

A staggered theoretical model that describes the user behavior utilizing activities, exercises, intra-movement behavior and between action behavior using this calculated model, a deep learning design dependent on LSTMs that models between action behaviors are introduced [2]. Two distinct Long Short-Term Memory (LSTM) networks are developed that take into account various suppositions about the information and accomplish distinctive demonstrating complexities and forecast exactnesses. The networks are trained and tested with two genuine world datasets, one being openly accessible while the other gathered from a field test. Displaying on the portion level public dataset mitigates the cold beginning issue. Experiments demonstrate that contrasted with customary methodologies dependent on succession mining or secret Markov displaying, LSTM networks perform essentially better. The ability of LSTM networks to detect long term correlations in activity data is also demonstrated. The trained models are each under 500 KB in estimate and can be sent to run continuously on a mobile phone with no conditions on the cloud. This can help applications like mobile personal assistant by giving predictive context [11].

People are progressively coming into contact with computerized reasoning and AI frameworks. On occasion it is obvious, as on account of Siri, Alexa, Cortana, or Google Assistant. It is likewise apparent on account of self-driving cars or non-player characters in computer games. At times it is less evident, as on account of calculations that work in the background to suggest products, and approve bank loans. Given the potential for intelligent framework to affect individuals' lives, designed-based frameworks with this in mind [13].

Human-focused AI is likewise in acknowledgment of the way that people can be similarly mysterious to intelligent systems. At the point when we consider intelligent frameworks getting people, we generally consider regular language and discourse handling whether a intelligent system can react suitably to expressions. Natural language processing, speech processing, and pastime reputation are critical demanding situations in constructing useful, interactive structures. To be sincerely effective, AI and ML structures want a principle of thoughts approximately humans.

3 Material and Methods

3.1 Deep Learning

Cognitive machine learning alludes to the blend of AI and brain cognitive system, explicitly, joining the accomplishments of AI. Three examination headings are proposed by creators and in this crisis of learning, integral learning framework and development of learning [16]. Deep learning finds many-sided structure in enormous informational collections by utilizing the backpropagation calculation to demonstrate how a machine should change its interior boundaries that are utilized to process the portrayal in each layer from the portrayal in the past layer. Deep convolutional nets have achieved leap forwards in preparing pictures, video, discourse, and sound, though repetitive nets have focused light on successive information like content and discourse [12, 15].

Predicting the conduct of human members in essential settings is a significant issue in numerous spaces. Most existing work either accepts that members are entirely reasonable, or endeavors to straightforwardly demonstrate every member's intellectual cycles dependent on experiences from cognitive psychology science and trial financial matters. In this work, we present another option, a deep learning approach that consequently performs cognitive modeling without depending on such master information. Authors presented an original design that permits a solitary organization to sum up across various info and yield measurements by utilizing grid units as opposed to scalar units, and show that its exhibition altogether outflanks [10].

Deep neural networks (DNN) models can possibly give new experiences in the investigation of human dynamic, because of their high limit and information driven plan. While these models might have the option to go past hypothesis driven models in predicting human conduct, their hazy nature restricts their capacity to clarify how an activity is completed. This clarify capacity issue stays annoying. Here authors showed the utilization of a DNN model as an exploratory instrument to recognize unsurprising and reliable human conduct in esteem-based dynamic past the extent of hypothesis driven models. Creators are additionally proposed utilizing hypothesis driven models to portray the activity of the DNN model and furthermore prepared a DNN model to foresee human choices in a four-armed bandit task. We tracked down that this model was more precise than a support learning reward-arranged model equipped toward picking the most remunerating choice [7].

Authors developed a brought together structure for movement acknowledgment-based conduct examination and activity expectation. For this reason, first they proposed part combination technique for exact action acknowledgment and afterward distinguish the huge consecutive practices of occupants from perceived exercises of their day by day schedules. Besides, practices designs are additionally used to foresee the future activities from past exercises [6].

In the period of massive information, deep learning shows exceptionally fascinating points of view. It is encountering remarkable achievement in numerous applications. DL utilizes AI procedures including managed as well as solo methodologies

to learn various leveled portrayals in deep models. DL utilizes profound models to manage complex connections between the information and the class label. DL and gathering-based calculations were extremely famous and proficient for multi-source and multi-temporal far off detecting picture characterization. They perform better compared to Support Vector Machine (SVM) as the DL can bargain with optical pictures as well as with radar pictures. DL showed a superior exhibition in removing highlights from hyper-spectral and multi-spectral pictures like separating sorts of names, pixel-based grouping, semantic division, and acknowledgment of articles and classes [5].

3.2 TensorFlow Framework

The most renowned system that we can use in deep learning for our case is TensorFlow [4]. TensorFlow (TF) is an open-source structure bound for AI. Tensorflow has inherent help for DL and neural network (NN), it is easy to assemble an organization, allot boundaries, and run the preparation interaction. Tensorflow additionally incorporates a library of basic, teachable mathematic capacities valuable for NN. In addition, TensorFlow utilizations Central Processing Units (CPUs) and Graphics Processing Units (GPUs) for registering and that enlivens the aggregate time [8].

Tensorflow is a structure for communicating ML calculations and an execution for executing it [1]. To be sure, it is utilized for building, preparing, and inducing a few DL models. TensorFlow additionally offers various tasks that are appropriate for neural networks, for example, softmax, sigmoid, relu, convolution2D, and maxpool [1]. Subsequently, this system is utilized in various applications, for example, picture preparing, picture characterization, object recognition, and semantic division [17]. In view of the intricacy of enormous satellite information, it is very troublesome, to assemble a direct neural organization learning model for picture arrangement. We chose to exploit the advantages of profound learning and conventional characterization techniques to accomplish better outcomes (Fig. 1).

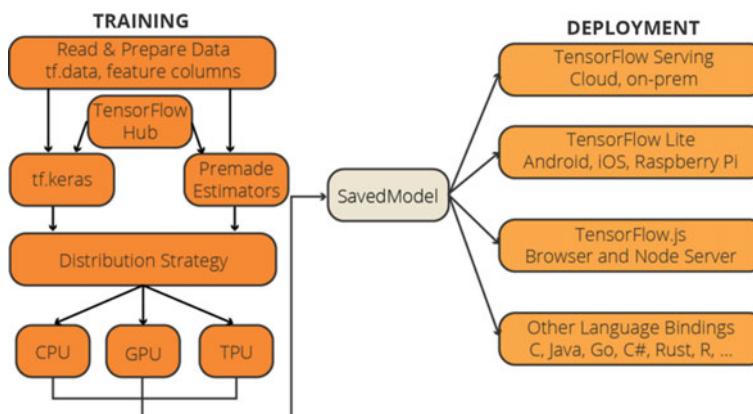


Fig. 1 Diagram of the TensorFlow platform

As you can see, TensorFlow is much more than a deep learning library for Python. It is an end-to-end platform that you can process your data, build and train machine learning models, serve the trained models across different devices with different programming languages.

4 Results and Discussion

The information comprises of 48×48 pixel grayscale pictures of countenances. The countenances have been consequently enlisted so the face is pretty much focused and possesses about a similar measure of room in each picture. Each picture relates to a look in one of seven classes (0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral). The datasets contains around 36K pictures. The first information comprised in exhibits with a grayscale an incentive for every pixel. Would you be able to figure which pictures are identified with which articulations?

The image articulations in our preparation dataset are really adjusted, with the exception of the ‘disgust’ classification. Deep learning models are prepared by being taken care of with clusters of information. Keras has an extremely valuable class to consequently take care of information from an index: ImageDataGenerator. It can likewise perform information expansion while getting the images (haphazardly pivoting the picture, zooming, and so on). This technique is frequently utilized as an approach to misleadingly get more information when the dataset has a little size.

We decided to utilize a convolutional neural network to handle this face acknowledgment issue. Without a doubt this sort of neural network (NN) is useful for extracting the components of pictures and is generally utilized for picture examination subjects like image characterization. Exemplary NNs are generally made out of a few completely associated layers. This implies that each neuron of one layer is associated with each neuron of the following layer. Convolutional neural networks additionally have convolutional layers that apply sliding capacities to gathering of pixels that are close to one another. In this manner those designs have a superior comprehension of examples that we can see in images.

We got outputs at each step of the training phase. We can use it to plot the evolution of the loss and accuracy on both the train and validation datasets. We have used epoch size as 10, 20, 50, and 100 and also check accuracy of the model.

4.1 *Epoch=10*

The validation accuracy begins to settle toward the finish of the 10 epochs somewhere in the range of 35–40% precision as displayed in Fig. 2. The validation loss is marginally higher than the validation loss for the primary epochs which can be amazing. To be sure we are more used to see higher validation loss than training losses in

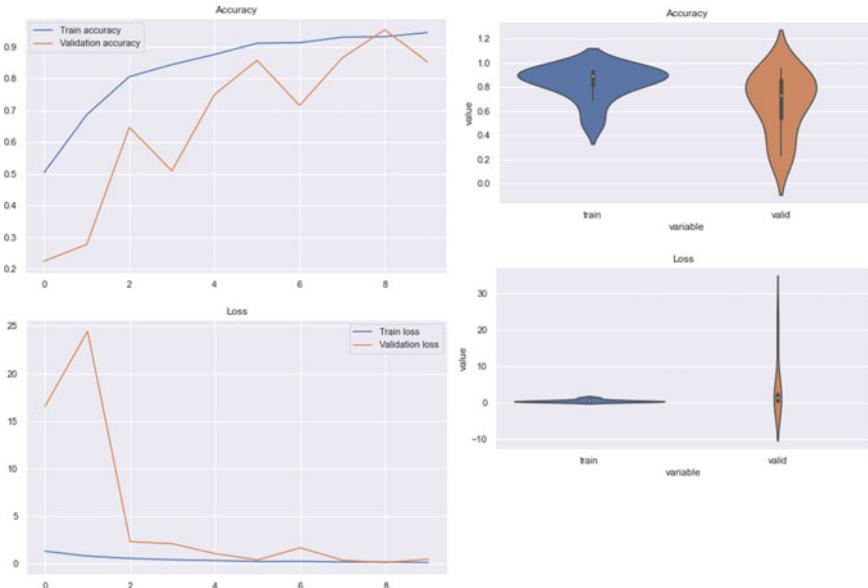


Fig. 2 Training versus validation w.r.t accuracy and loss when epoch=10

AI. Here this is just because of the presence of dropout, which is just applied during the preparation stage and not during the validation stage.

We can see that the training loss is turning out to be a lot more modest than the validation loss after the second epoch as displayed in Fig. 2. This implies that our model begins to once again accommodate our training datasets after an excess of emphases. That is the reason the validation loss doesn't diminish much after.

CNN Model Accuracy on test set: 0.9455 (10 epoch).

4.2 Epoch=20

The validation accuracy begins to settle toward the finish of the 20 epochs somewhere in the range of 60–65% precision as displayed in Fig. 3. The validation loss is marginally higher than the validation loss for the primary epochs which can be amazing. To be sure we are more used to see higher validation loss than training losses in AI. Here this is just because of the presence of dropout, which is just applied during the preparation stage and not during the validation stage.

We can see that the training loss is becoming much smaller than the validation loss after the 8th epochs as shown in Fig. 2. This means that our model starts to over-fit our training dataset after too much iterations. That is why the validation loss does not decrease a lot after.

CNN Model Accuracy on test set: 0.8747 (20 epoch).

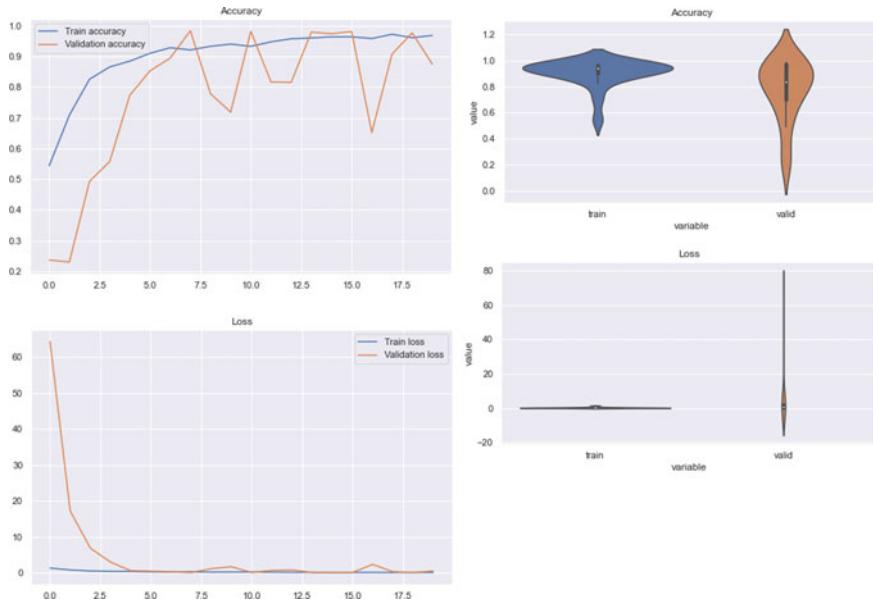


Fig. 3 Training versus validation w.r.t accuracy and loss when epoch=20

4.3 Epoch=50

The validation accuracy begins to settle toward the finish of the 50 epochs somewhere in the range of 60–65% precision as displayed in Fig. 4. The validation loss is marginally higher than the validation loss for the primary epochs which can be amazing. To be sure we are more used to see higher validation loss than training losses in AI. Here this is just because of the presence of dropout, which is just applied during the preparation stage and not during the validation stage.

We can see that the training loss is becoming much smaller than the validation loss after the 20th epoch as shown in Fig. 4. This means that our model starts to over-fit our training dataset after too much iterations. That is why the validation loss does not decrease a lot after.

CNN Model Accuracy on test set: 0.9637 (50 epoch).

4.4 Epoch=100

The validation accuracy begins to settle toward the finish of the 100 epochs somewhere in the range of 80–85% precision as displayed in Fig. 5. The validation loss is marginally higher than the validation loss for the primary epochs which can be amazing. To be sure we are more used to see higher validation loss than training losses in

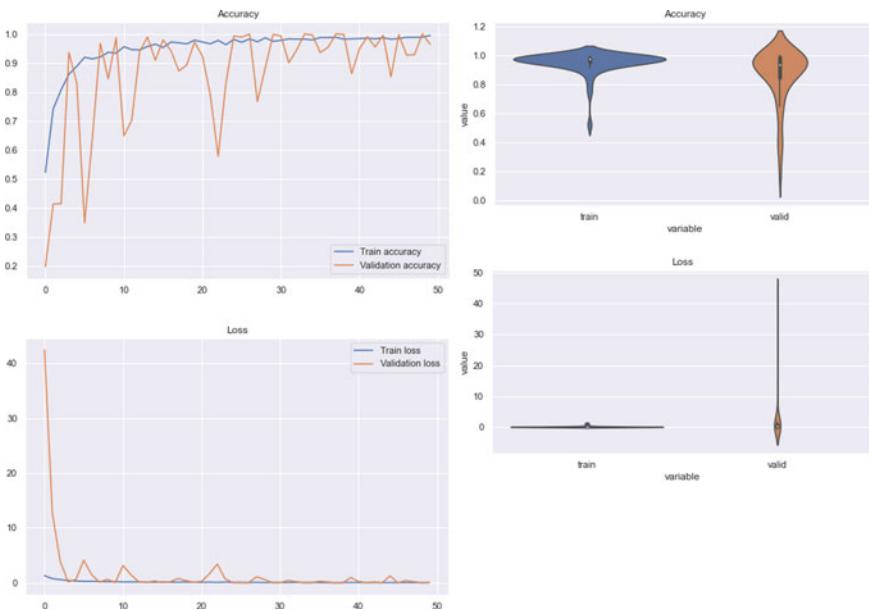


Fig. 4 Training versus validation w.r.t accuracy and loss when epoch=50

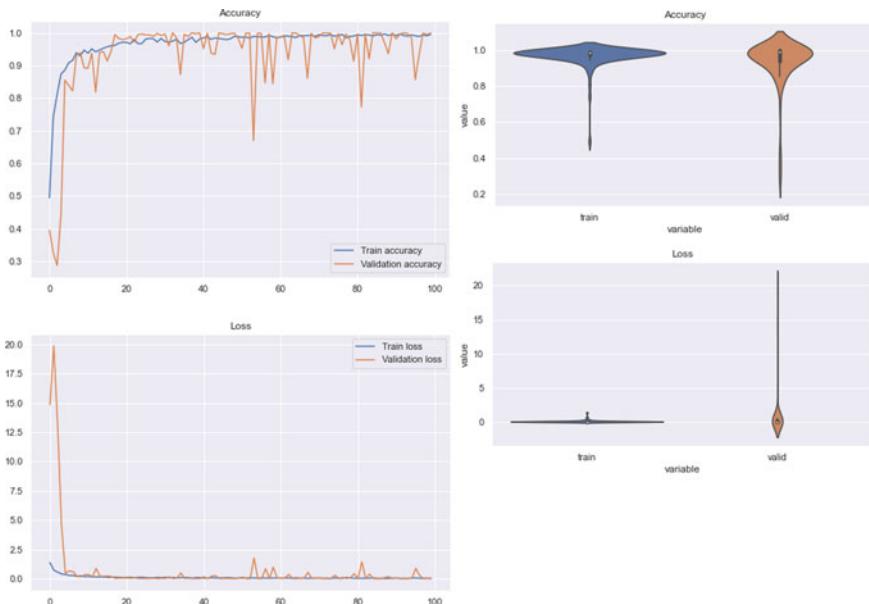


Fig. 5 Training versus validation w.r.t accuracy and loss when epoch=100

AI. Here this is just because of the presence of dropout, which is just applied during the preparation stage and not during the validation stage.

We can see that the training loss is becoming much smaller than the validation loss after the 55th epochs as shown in Fig. 5. This means that our model starts to over-fit our training dataset after too much iterations. That is why the validation loss does not decrease a lot after.

CNN Model Accuracy on test set: 0.9991 (100 epoch).

5 Conclusion

We have determined human behavior based on real life images supported by cognitive-based neural network. This work is mainly focused to identify the behavior of a human that could be anger, happy, sad, disgust, etc. The images are processed to make prediction followed by convolutional neural network (CNN). As we know that, a CNN model follows various layers to process any image. By this model, an image is filtered to generate better results as well as prediction. The real images are used to get the better in terms of accuracy and this gives us a better outcomes.

In the future, we will focus and analyze the characteristics of other behaviors such as walking, sleeping, weeping, crying, and fainting for our results. We are also planning to improve research work by new methods to improve the performance and ability of our proposed system. We are planning to work video-based behavior to identify behavior of human from CCTV cameras.

References

1. Abadi M et al (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv Preprint. [arXiv:1603.04467](https://arxiv.org/abs/1603.04467)
2. Almeida A, Azkune G (2018) Predicting human behaviour with recurrent neural networks. *Appl Sci* 8:305
3. Bhatia R (2013) Biometrics and face recognition techniques. *Int J Adv Res Comput Sci Softw Eng* 3
4. Chebbi I, Mellouli N, Farah IR, Lamolle M (2021) Big remote sensing image classification based on deep learning extraction features and distributed spark frameworks. *Big Data Cogn Comput* 5:21
5. Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. *IEEE Access* 2:514–525
6. Fatima I, Fahim M, Lee YK, Lee S (2013) A unified framework for activity recognition-based behavior analysis and action prediction in smart homes. *Sensors* 13:2682–2699
7. Fintz M, Osadchy M, Hertz U (2021) Using deep learning to predict human decisions and cognitive models to explain deep learning models. *bioRxiv*
8. Goldsborough P (2016) A tour of tensorflow. arXiv Preprint. [arXiv:1610.01178](https://arxiv.org/abs/1610.01178)
9. Haffner P (2016) What is machine learning—and why is it important. *Interactions* 7
10. Hartford JS (2016) Deep learning for predicting human strategic behavior. Ph.D. thesis. University of British Columbia

11. Krishna K, Jain D, Mehta SV, Choudhary S (2018) An LSTM based system for prediction of human activities with durations. *Proc ACM Interact Mob Wearable Ubiquit Technol* 1:1–31
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444
13. Riedl MO (2019) Human-centered artificial intelligence and machine learning. *Hum Behav Emerg Technol* 1:33–36
14. Sareen P (2014) Biometrics-introduction, characteristics, basic technique, its types and various performance measures. *Int J Emerg Res Manag Technol* 3:109–119
15. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
16. Shi Z et al (2019) Cognitive machine learning. *Int J Intell Sci* 9:111
17. Yao Y, Liang H, Li X, Zhang J, He J (2017) Sensing urban land-use patterns by integrating google tensorflow and scene-classification models. *arXiv Preprint. arXiv:1708.01580*
18. Zaccone G, Karim MR, Menshawy A (2017) Deep learning with TensorFlow. Packt Publishing Ltd

Single IC-Based Third-Order Sinusoidal Oscillator



Gurumurthy Komanapalli, Pandey Rajeshwari, and Pandey Neeta

Abstract This paper presents a new, third-order sinusoidal oscillator (TOSO) configuring a single current feedback operational amplifier (CFOA) and only six passive components. The proposed circuit bares independent control of the oscillation frequency through the capacitor. The nonideal analysis in the presence of CFOA parasitics is also evaluated for proposed TOSO. In addition, a detailed sensitivity analysis has also been carried out, and it exhibits low active and passive sensitivities. Workability, the effectiveness of the proposed formulation has been checked by the simulation results using PSPICE. Simulation results incorporate transient sinusoidal waveform and fast fourier transform (FFT) outputs yielded by employing the AD844 IC. Experimentation of the design has been incorporated, for which AD844 IC along with standard passive components have been utilized. Corroborating outcomes have been accomplished in both cases. Total harmonic distortion (THD) is obtained to be less than 2.6%.

Keywords Current feedback operational amplifier (CFOA) · Sinusoidal oscillator · Sensitivity analysis · Total harmonic distortion

1 Introduction

Sinusoidal oscillators (SOs) are ubiquitous in numerous electronic devices and play a vital role in measurement, power electronics, control, and other electronic systems and instrumentation [1]. Third-order oscillators are notable choice [2–4] over second-order oscillators, as they deliver good frequency response with low harmonic distortion which is useful in this kind of applications.

G. Komanapalli (✉)

School of Electronics Engineering, VIT-AP University, Inavolu, Beside AP Secretariat, Amaravati, Andhra Pradesh, India
e-mail: gurumurthy.k@vitap.ac.in

P. Rajeshwari · P. Neeta

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

Confining to the last three decades, our survey found that several third-order sinusoidal oscillators (TOSOs) are presented in literature [2–15] and references cited therein]. Commercially available ICs are used in [2–15] for experimental verification, while [16–19] and references cited therein] rely on pre/ post layout simulations. The emphasis of this paper is on TOSOs realized using commercially available ICs. Table 1 lists the distinctive features of these TOSOs. The comparative study of these TOSOs infers that existing oscillators use more than one IC for its hardware implementation. The TOSOs [7, 8] are not canonical in terms of capacitors, i.e., the capacitors are more than three in number. This paper aims at presenting single IC-based TOSO that use canonical number of capacitors and resistors. To the best of authors knowledge, this is the first single IC-based third-order SO. Section 2 describes the proposed oscillator topology. The nonideality calculations are shown in Sect. 3, and these calculations are carried out by considering parasitic resistances and capacitances at different nodes of CFOA. Section 4 gives sensitivity analysis and also discusses the implications arised on these calculations due to nonidealities of CFOA. The functionality of the proposed topology is confirmed through PSPICE simulations, and corresponding results are presented in Sect. 5. The experimental results, wherein the circuit is breadboarded using off-the-shelf ICs AD844, are demonstrated in Sect. 6. Finally, Sect. 7 concludes this paper.

2 Proposed Circuit Description

The CFOA is a four-terminal building block [20–23] which internally comprises positive second-generation current conveyor followed by voltage buffer. The symbolic representation of CFOA and its equivalent notation are depicted in Fig. 1a, b, respectively. The port characteristics are given by

$$\begin{bmatrix} V_x \\ I_y \\ I_z \\ V_w \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} I_x \\ V_y \\ V_z \\ I_w \end{bmatrix} \quad (1)$$

The proposed CFOA-based TOSO is depicted in Fig. 2a. It employs a single CFOA and six passive components. The characteristic equation (CE) for Fig. 2a is obtained by performing routine analysis assuming CFOA to be ideal and is given by

$$\begin{aligned} s^3 C_1 C_2 C_3 R_1 R_2 R_3 + s^2 \left[\begin{array}{l} C_1 C_2 R_1 R_2 + C_2 C_3 R_1 R_3 + \\ C_2 C_3 R_2 R_3 - C_1 C_3 R_1 R_2 \end{array} \right] \\ + s [C_2 R_1 + C_2 R_2 + C_3 R_3 - C_3 R_1 - C_3 R_2] + 1 = 0 \end{aligned} \quad (2)$$

Assuming $R_1 = R_2 = R$, $C_1 = C_2 = C_3 = C$, CE Eq. (2) reduces to:

Table 1 Comparison of earlier known TOSOs

Refs.	No./type of ABBs used	Hardware implementation/type of ICs used	No. of ICs used	No. of resistors used	No. of capacitors used
2	3 OTA	Y/MC14007	Fig. 7b. Eight	0	3
	4 OTA	Y/MC14007	Fig. 11. Seven	1	3
4	2DVCCTA	Y/AD844	Fig. 4a. Ten	1	3
		Y/AD844	Fig. 4b. Ten	2	3
5	2 OTRA	Y/AD844	Fig. 2d. Four	3	3
		Y/AD844	Fig. 3d. Four	3	3
6	3OPAMP	Y/LF351	Fig. 1. Three	5	3
		Y/LF351	Fig. 2. Three	3	5
7	3 CCII	Y/AD844	Fig. 1. Four	5	3
		Y/AD844	Fig. 2. Four	3	5
		Y/AD844	Fig. 3. Four	5	3
8	3 CCCII	Y/AD844	Nine	0	3
9	3 CFOA	N/AD844	Figure 3b. Three	4	3
10	4 DVCC	Y/AD844	Twelve	3	3
11	1MCCFTA	Y/AD844	Nine	0	3
12	2 OTRA	Y/AD844	Four	3	3
13	2MO-DVCCTA	Y/AD844, AD8130, MAX435, CA3080	Eight	2	3
14	1OTRA	Y/AD844	Two	3	3
15	1OTRA	Y/AD844	Two	3	3
Proposed	1 CFOA	Y/AD844	One	3	3

OPAMP operational amplifier, *CCII* second generation current conveyor, *CFOA* current feedback operational amplifier, *CCCII* current controlled conveyor, *DVCC* differential voltage current conveyor, *OTRA* operational transresistance amplifier, *DVCCTA* differential voltage current conveyor transconductance amplifier, *OTA* operational transconductance amplifier, *VDTA* voltage difference transconductance amplifier, *MCCFTA* Modified current controlled current follower transconductance amplifier

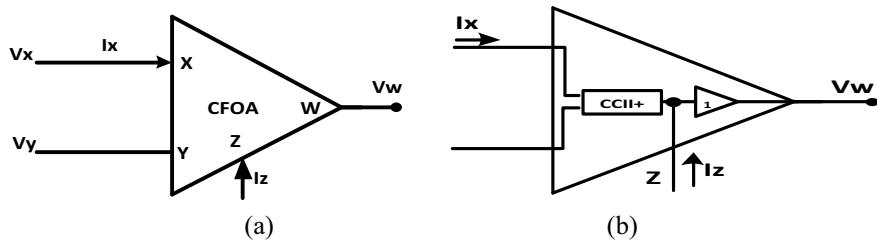


Fig. 1 **a** CFOA circuit symbol and **b** equivalent realization

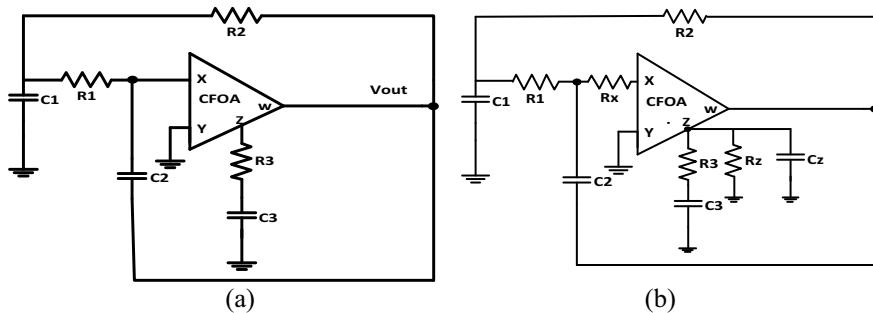


Fig. 2 **a** Proposed CFOA-based TOSO circuit, **b** proposed topology with nonidealities

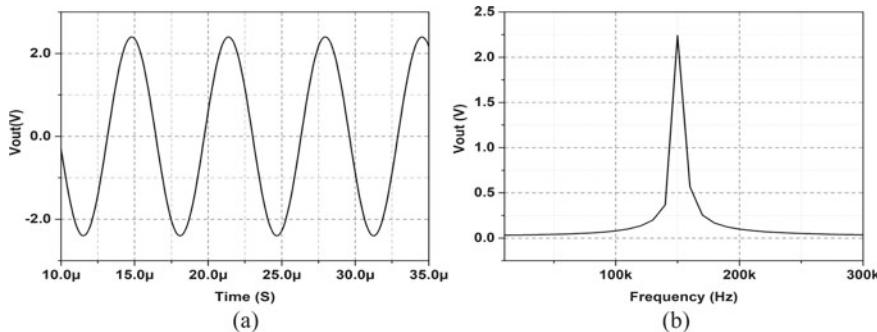


Fig. 3 **a** Steady-state output and **b** FFT response of proposed topology

$$s^3 C^3 R^2 R_3 + 2s^2 C^2 R R_3 + s C R_3 + 1 = 0 \quad (3)$$

Therefore, frequency of oscillation (FO) and condition of oscillation (CO) are given by

$$\text{FO : } f_o = \frac{1}{2\pi R C}; \text{ CO : } 2R_3 = R \quad (4)$$

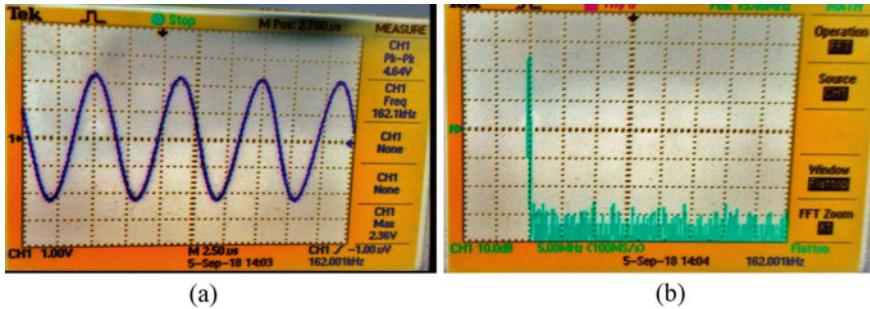


Fig. 4 Experimental, **a** steady-state waveform and its **b** FFT spectrum

It may be noted that from Eq. (4) CO can be adjusted using R_3 . Similarly, FO can be tuned using C without affecting CO. In practice, FO tuning via resistors is suitable choice compared to capacitor tuning. However, the capacitors may be replaced by a capacitor bank made by parallel connection of a capacitor and a series switch. The value of capacitance may then be adjusted by appropriate switch settings [15].

3 Consideration of CFOA Non-idealities

The FO and CO of the proposed SO may differ from the actual results because of the CFOA nonidealities. The major nonidealities of the CFOAs are a port- X parasitic input resistance R_x ($\cong 50 \Omega$), parasitic resistance R_y in parallel with $1/sC_y$ at the port- Y and the port- Z parasitic impedance consisting of a parasitic capacitance C_z ($\cong 5 \text{ pF}$) in parallel with parasitic resistance R_z ($\cong 3 \text{ M}\Omega$). As Y port is grounded in proposed TOSO the corresponding parasitics at this port can be easily precluded. Figure 2b shows the proposed TOSO with parasitics, and its analysis gives CE of

$$A_1 s^3 + A_2 s^2 + A_3 s + A_4 = 0 \quad (5)$$

where coefficients A_1 , A_2 , A_3 , and A_4 are given as

$$A_1 = C^3 R^2 R_3 R_z \quad (6)$$

$$A_2 = 2C^2 R R_3 R_z + C^2 R^2 R_z - \left\{ C^2 R R_3 \begin{pmatrix} (R + 2R_x) \\ -R_x(sCR + 1) \end{pmatrix} \right\} (1 + sC_z R_z) - C R R_z (R + 2R_x)(C + C_z) - C R R_z (1 + sCR)(C + C_z) \quad (7)$$

$$A_3 = C R_3 R_z - C R (R + 2R_x) - C R R_x (1 + sCR) - \{C R_3 (1 + sC_z R_z) - R_z (C + C_z)\} (R_x + 2R) + 2C R R_z \quad (8)$$

$$A_4 = R_z - (2R + R_x) \quad (9)$$

Consequently, the nonideal FO and CO expressions are given:

$$\text{FO: } \hat{f}_o = f_o \sqrt{\frac{CR(R + 3R_x - 2R_z) - CR_3R_z + (CR_3 + R_z(C + C_z))(R_x + 2R)}{C_zR_3R_z\left(\frac{3R_x}{R} + 1\right) - CR_3R_z + R_x(CR_3 + R_z(C + C_z))}} \quad (10)$$

$$\begin{aligned} \text{CO: } & \hat{f}_o^2 4\pi^2 C^3 R^2 R_3 R_z R_x C_z + \frac{(2R + R_x) - R_z}{\hat{f}_o^2 4\pi^2} \\ & = C^2 R R_3 (R + 3R_x) \\ & + \{C R R_z (C + C_z)\} (R + 3R_x) \\ & + C^2 R^2 R_x + C R_z R_3 C_z (2R \\ & + R_x) - C^2 R^2 R_z - 2C^2 R R_3 R_z \end{aligned} \quad (11)$$

A careful look at the Eqs. (10) and (11) divulges that due to nonidealities of CFOA the FO will slightly differ from its theoretical values.

Selecting $R > > R_x$ and frequency much below $1/C_z R_z$, the coefficients A_1 , A_2 , A_3 , and A_4 are approximated to

$$\begin{aligned} A_1 & \simeq C^3 R^2 R_3 \left(1 - \frac{R_x}{R_z} - \frac{R_x}{R_3}\right); A_2 \simeq 2C^2 R R_3 - C^2 R R_3 \left(\frac{R}{R_z} - \frac{R_x}{R_z}\right) - C^2 R R_x \\ A_3 & \simeq C R_3 - C R \left(\frac{R}{R_z} + \frac{R_x}{R_z} + 2\frac{R_3}{R_z}\right); A_4 \simeq \left(1 - 2\frac{R}{R_z}\right) \end{aligned} \quad (12)$$

As $1/R_z \rightarrow 0$ (very small), the FO and CO in Eq. (10) and (11) are approximated to Eq. (4). From Eqs. (6)–(9) and Eq. (12), it is noted that the discrepancies caused by the influence of parasitics of CFOA can be reduced by selecting all external resistor values larger than R_x but much smaller than R_z and R_y and external capacitors to be much larger than C_z and C_y .

4 Sensitivity Analysis

The sensitivity is an important performance criterion for any circuit, which enables analog IC designers to choose which elements should be carefully designed to maintain high circuit performance [24, 25]. The sensitivity of f_o with respect to a circuit component (say Y) is given by

$$S_Y^{f_o} = \frac{\partial f_o}{\partial Y} \cdot \frac{Y}{f_o} \quad (13)$$

Using Eq. (13), the sensitivity of FO (f_o) for the proposed TOSO with respect to C , R and R_3 turns out to be

$$\left| S_C^{f_o} \right| = \left| S_R^{f_o} \right| = 1; \left| S_{R_3}^{f_o} \right| = 0 \quad (14)$$

Considering Eq. (13), the sensitivity is computed as

$$\left| \hat{S}_C^{f_o} \right| = \frac{Ck_1}{2(Ck_1 + l_1)} - \frac{2m_1 + 3Cn_1}{2(Cn_1 + m_1)} \quad (15)$$

where

$$k_1 = R(R + 3R_x - 2R_z) - R_3R_z + (R_3 + R_z)(R_x + 2R); \quad l_1 = (R_x + 2R)C_z \\ m_1 = RR_3R_zC_z(3R_x + R) - R^2C_zR_3R_x; \quad n_1 = R^2(R_3R_x + R_xR_z - R_3R_z)$$

$$\left| \hat{S}_R^{f_o} \right| = \frac{2RC + k_2 + l_2}{2(RC + k_2 + l_2)} - \frac{Rm_2}{(n_2 + Rm_2)} \quad (16)$$

where

$$k_2 = 3R_xC + 2CR_3 - 2CR_z + 2R_z(C + C_z); \\ l_2 = CR_3(R_x - R_z) + R_xR_z(C + C_z) \\ m_2 = C^2(CR_3R_x + R_xR_z(C_z + C) + R_3R_z(C_z - C)); \\ n_2 = 3C^2C_zR_xR_3R_z$$

$$\left| \hat{S}_{R_x}^{f_o} \right| = \frac{R_xk_3}{2(R_xk_3 + l_3)} - \frac{R_xm_3}{2(R_xm_3 + n_3)} \quad (17)$$

where

$$k_3 = 3RC + CR_3 + R_z(C + C_z); \\ l_3 = 2R(CR_3 + R_z(C + C_z) + RC(R - 2R_z) - CR_3R_z \\ m_3 = 3C^2C_zR_xR_3R_z + C^2R^2(CR_3 + R_z(C_z + C)); \\ n_3 = C^2R^2R_3(R_zC_z - C)$$

$$\left| \hat{S}_{R_z}^{f_o} \right| = \frac{R_zk_4}{2(R_zk_4 + l_4)} - \frac{R_zm_4}{2(R_zm_4 + n_4)} \quad (18)$$

where

$$\begin{aligned}
k_4 &= (C + C_z)(R_x + 2R) - CR_3 - 2RCR_2; \\
l_4 &= RC(3R_x + R) + CR_3(R_x + 2R) \\
m_4 &= C^2C_zRR_3(3R_x + R) - C^3R_3R_zR^2 + C^2R^2R_x(C_z + C); \\
n_4 &= C^3R_3R_xR^2
\end{aligned}$$

$$\left| \hat{S}_{C_z}^{f_o} \right| = \frac{C_z k_5}{2(C_z k_5 + l_5)} - \frac{C_z m_5}{2(C_z m_5 + n_5)} \quad (19)$$

where

$$\begin{aligned}
k_5 &= R_z(R_x + 2R); \quad l_5 = RC(3R_x + R - 2R_z) - CR_3R_z + C(R_x + 2R)(R_3 + R_z) \\
m_5 &= C^2RR_3R_z(3R_x + R) + R_zC^2R^2R_x; \quad n_5 = C^3R^2R_3(R_x - R_z) + C^3R_zR^2R_x
\end{aligned}$$

Sensitivity equations from Eqs. (15) to (19) imply that its values can be maintained low by selecting all external capacitors larger than C_z , C_y , and the values of external resistors higher than R_x but much smaller than R_z , R_y .

5 Simulation Results

To validate the theoretical analysis, the proposed topology is simulated in PSPICE by using IC AD 844. Supply voltages taken are as ± 8.5 V. The proposed TOSO was simulated with component settings as $C_1 = C_2 = C_3 = C = 100$ pF, $R_3 = 5$ k Ω , $R_1 = R_2 = 10$ k Ω and oscillation frequency was observed to be 155 kHz against the theoretical value of 159.2 kHz.

The steady-state output waveform and its frequency spectrum are depicted in Fig. 3a, b, respectively. It may be observed that the theoretical and simulated values of FO are in close agreement. The %THD is found to be 2.6%.

6 Experimental Verification

The proposed CFOA TOSO is also verified experimentally by breadboarding AD844 IC's and considering same resistor and capacitor values with 5% tolerances as quoted in simulated driven output. The DC biasing \pm VDC is taken as ± 8.5 V.

The experimental-driven steady-state waveform, frequency response recorded by TDS 2012C Tektronix oscilloscope are depicted in Fig. 4a, b, respectively. The practical FO is obtained as 162.1 kHz. The % deviation in FO is found to be less than 2%.

7 Concluding Remarks

A new realization of third-order sinusoidal oscillator employing single CFOA and six passive components are proposed in this paper. The validity and functionality of the proposed configuration are verified through PSPICE. The experimental and simulation results are found to be in close agreement with theoretical propositions. The effect of parasitics present at various nodes of CFOA on output response has been analyzed. The sensitivity of oscillation frequency with respect to passive components is also inspected and values found out to be low. This topology is further verified experimentally using off-the-shelf CFOA IC AD844.

References

1. Sedra AS, Smith KC (2004) Microelectronic circuits, 4th edn. Oxford University Press
2. Prommee P, Dejhan K (2012) An integrable electronic-controlled quadrature sinusoidal oscillator using CMOS operational transconductance amplifier. *Int J Electron* 89:365–379. <https://doi.org/10.1080/713810385>
3. Horng JW (2009) Current-mode third-order quadrature oscillator using CDTAs. *Act Passive Electron Compon* 1–4. <https://doi.org/10.1155/2009/789171>
4. Pandey N, Pandey R (2014) Approach for third order quadrature oscillator realisation. *IET Circ Devices Syst* 9(3):1–11. <https://doi.org/10.1049/iet-cds.2014.0170>
5. Nagar BC, Paul SK (2016) Voltage mode third order quadrature oscillators using OTRAs. *Analog Integr Circ Sig Process* 88(3):517–530. <https://doi.org/10.1007/s10470-016-0781-6>
6. Horng J (2011) Quadrature oscillators using operational amplifiers. *Act Passive Electron Compon* 1–4. <https://doi.org/10.1155/2011/320367>
7. Horng JW, Hou CL, Chang CM, Chung WY, Tang HW, Wen YH (2005) Quadrature oscillators using CCIIs. *Int J Electron* 92(1):21–31. <https://doi.org/10.1080/00207210412331332899>
8. Maheshwari S (2010) Current-mode third-order quadrature oscillator. *IET Circ Devices Syst* 4(3):188. <https://doi.org/10.1049/iet-cds.2009.0259>
9. Soliman AM (2013) Generation of third-order quadrature oscillator circuits using Nam expansion. *J Circ Syst Comput* 22(07):1350060. <https://doi.org/10.1142/S0218126613500606>
10. Maheshwari S (2009) Analogue signal processing applications using a new circuit topology. *IET Circ Devices Syst* 3(3):106–115. <https://doi.org/10.1049/iet-cds>
11. Khaw-Ngam K, Kumngern M, Khateb F (2017) Mixed-mode third-order quadrature oscillator based on single MCCFTA. *Radioengineering* 26(2):522–535. <https://doi.org/10.13164/re.2017/05/22>
12. Nagar BC, Paul SK (2018) Realization of OTRA-based quadrature oscillator using third-order topology. In: *Advances in systems, control and automation*, pp 375–386. Springer, Singapore. https://doi.org/10.1007/978-981-10-4762-6_36
13. Chen HP, Hwang YS, Ku YT (2017) A systematic realization of third-order quadrature oscillator with controllable amplitude. *AEU-Int J Electron Commun* 79:64–73
14. Chien H (2016) Third-order sinusoidal oscillator using a single CMOS operational transresistance amplifier. *J Appl Sci Eng* 19(2):187–196. <https://doi.org/10.6180/jase>
15. Komanapalli G, Pandey N, Pandey R (2018) New realization of third order sinusoidal oscillator using single OTRA. *Int J Electron Commun (AEÜ)* 93(June):182–190. <https://doi.org/10.1016/j.aeue.2018>
16. Horng JW (2011) Current/voltage-mode third order quadrature oscillator employing two multiple outputs CCIIs and grounded capacitors. *Indian J Pure Appl Phys* 49(7):494–498

17. Pandey R, Pandey N, Paul SK (2012) MOS-C third order quadrature oscillator using OTRA. In: Third international conference on computer and communication technology, vol 1, pp 77–80. <https://doi.org/10.1109/ICCCCT.2012.24>
18. Channumsin O, Jantakun A (2014) Third-order sinusoidal oscillator using VDTAs and grounded capacitors with amplitude controllability. In: The 4th joint international conference on information and communication technology, electronic and electrical engineering (JICTEE), pp 1–4. IEEE
19. Komanapalli G, Pandey N, Pandey R (2017) OTRA based second and third order sinusoidal oscillators and their phase noise performance. AIP Conf Proc 1859:1–8. <https://doi.org/10.1063/1.4990170>
20. Bhaskar DR, Gupta SS, Senani R, Singh AK (2012) New CFOA-based sinusoidal oscillators retaining independent control of oscillation frequency even under the influence of parasitic impedances. Analog Integr Circ Sig Process 73(1):427–437. <https://doi.org/10.1007/s10470-012-9896-6>
21. Celma S, Martinez PA, Carlesena A (1994) Current feedback amplifier based sinusoidal oscillators. IEEE Trans Circ Syst I: Fundam Theory Appl 41(12):906–908. <https://doi.org/10.1109/81.340855>
22. Martinez PA, Celma S, Sabadell J (1996) Designing sinusoidal oscillators with current-feedback amplifiers. Int J Electron 80:637–646
23. Gandikota NC, Komanapalli G (2022) CFOA based second order low frequency sensitive sinusoidal oscillator. In: 2nd International conference on artificial intelligence and signal processing (AISP), pp 1–4. <https://doi.org/10.1109/AISP53593.2022.9760529>
24. Komanapalli G, Gupta P (2022) A new realization of third-order Inverse Bandpass filter and its application as an oscillator. J Circ Syst Comput 31(05):2250079
25. Komanapalli G, Pandey N, Pandey R (2022) New electronically tunable low-frequency quadrature oscillator using operational transresistance amplifier. IETE J Res 68(4):2571–2579. <https://doi.org/10.1080/03772063.2020.1721337>

Internet of Things (IoT)-Based Waste Dumping System



Birinderjit Singh Kalyan

Abstract Traditional system for the collection of garbage cost higher than the smart automated systems. Considering the facts, the hassle of a success waste control is one of the maximum essential troubles of our time, there is a most outrageous need to determine this issue. The system proposed in this paper is simply an improved solution with a superior intelligent waste level detection system, a higher-level design, which can immediately notify the officials of the status of the sorted city trash cans and monitor them in real time. It is time for remote control with Internet of Things (IoT). Imposing IoT improvements within side the modern-day city waste control surroundings is critical and permits bi-directional conversation such as the rules conveyed within side the town. The objective to achieve is a real-time monitoring system, a system that is centralized. The municipal and residents will take advantage of such optimized solution which ends up in primary financial monetary savings and plenty much less pollution that is metropolitan. A vital process of proper waste management is much needed for the sanitation society as an entire and the arena as an entire, also the automation of measures efficiently reduces the load on people.

Keywords Internet of Things (IoT) · Smart waste level detection system · Waste management · Pollution

1 Introduction

The Internet of Things is nothing more than applications running over the Internet [1]. State-of-the-art technology that stores all your data in the cloud with fast real-time data access and intelligence [1]. Those with Internet access while data is stored in the cloud will provide unparalleled access to those involved in the same task from anywhere in the world. Detectors and routers used to collect and transmit data over the Internet have similar improvements. This area may be utilized in all regions of ubiquitous computing and commercial enterprise intelligence. This paper

B. S. Kalyan (✉)

University Institute of Engineering, Chandigarh University, Mohali, India

e-mail: birinderjit@msn.com

acquaints you exactly how IoT can be used in these various places, where smart garbage detection using IoT can be an important aspect in converting cities into smart cities. In concerning mortal well-being and the landscape from the implied perils of deferred garbage removal and natural contamination a completely directed and controlled running of these squanders is must. The sort of squanders which establish natural contamination and what this work highlights is its adaptation to household waste from degradable food waste, leaves, dead animals, and non-degradable bones, similar to plastics, holders, nylons, clinical waste, and residential and commercial waste. The complexity of urban heavy waste management challenges necessitates the development and use of state-of-the-art devices capable of driving alternative mechanisms, numerical models, and inputs for expert judgment in multi-standard decision scenarios [2]. Waste control is an ever-developing hassle in worldwide and non-preferred contexts. Solid waste arises from human and animal conditioning and is normally discarded as vain or unwanted. As accessories made from natural and inorganic waste, created by a lively public, and lost value to the main masons. Garbage bins are placed in public areas at specific places on site/road to collect municipal waste. The most crucial and delicate challenge is the method of checking the rubbish can for rubbish collection. This is a common system that requires people to walk around and check the garbage collection point. This is an instead complex and time-consuming method. The waste that the current system represents is not as efficient as we would like it to be, given technological advances in the recent past. There is no guarantee of operation/disposal of garbage in all locations. To solve this problem, a new approach called IoT-based automatic waste disposal system has been proposed. Basically, it's a step that automatically makes garbage collection efficient. This is noticed via way of means of putting an ultrasonic detector at the bin on every occasion the bin is full, and it makes use of a Wi-Fi module to transmit it to the proper Garcon at a designated location in that area or location. The input signal indicates the status of the waste container in the monitoring and control system.

2 Literature Review

The idea of an intelligent waste detection system has been discussed for a long time [1]. The Internet of Things (IoT), the technology used to create this smart method, has also evolved. Each idea looks similar but has a slightly different personality, and the work it proposes is no exception. After the Internet of Things has settled into our lives, we plan to develop an intelligent scrap collection system that includes citizen participation and data analysis to make better timber decisions. The intelligent system is a waste container with ultrasonic detectors, microcontrollers, and Wi-Fi modules for data transmission. Cloud Vision Enables Global Internet of Things to Proliferate [3]. This work exploits crucial operations and technologies that are likely to drive IoT exploration with unborn possibilities. However, there is a solid foundation that explains the basics and how Arduino boards work. It's relatively intriguing as it implements a 'Get As You Throw' system conception as a way to

encourage recycling among citizens [4]. As we mentioned later, the civic engagement part of the arrangement is relatively dependent on their work.

3 Applications

- Detects the magnitude of garbage inside the trash can.
- Wireless transmission of information to involved officials.
- The system can be accessed anytime, anywhere.
- Send and access data in real time.
- Overfilling of trash can is prevented.
- This system will help city authorities or other private companies to address the municipal waste collection problem.
- This system is not for personal use and may be used by any city, state, or country.
- With this system, efficiency in garbage collection and reduction in shipping costs can be observed.

4 Hardwares

4.1 *Arduino Mega*

The Arduino is an open-source project that creates microcontroller accessories for creating digital displacement and interactive objects that can sense and control physical displacement [5]. A microcontroller design-based system from several vendors which uses different accessories. These systems provide a variety of digital and analog I/O branches that allow interoperability with a variety of extensions and other circuits as shown in Fig. 1. There are interfaces for cyclic communication, including some models of universal cyclic automatic machines for downloading programs on specific computers.

Fig. 1 Arduino Mega



Fig. 2 Ultrasonic sensor (HC-SR04)



4.2 Ultrasonic Sensor (HC-SR04)

An ultrasonic sensor is an electronic gadget that converts the reflected sound waves into an electrical signal by evaluating the distance of an objective item by emanating ultrasonic sound waves as shown in Fig. 2. Ultrasonic waves move quicker than the speed of perceptible sound. Ultrasonic sensors have two primary parts: the transmitter and the receiver. To ascertain the distance between the object and the sensor, the sensor estimates the outflow of the sound by the transmitter to its contact with the collector by calculating the time it takes between them [3]. For calculating distance, $D = 1/2 \times T \times C$ (where C is the speed of sound, T is the time, and D is the distance).

4.3 HC-05 (Bluetooth Module)

Bluetooth serial port protocol module in an HC05 module is an easy to apply Bluetooth module that provides clean wireless and serial communication as illustrated in Fig. 3. Serial port Bluetooth module is very good Bluetooth with 3Mbps modulation and full 2.4 GHz wireless modulation and transceiver. A CMOS technology and AFH, which is adaptive frequency hopping which is used on an external chip, a separate system. It has small feet measuring 12.7×27 mm, hopefully, this can simplify the entire design/development cycle.

Fig. 3 Hc-05 (Bluetooth module) [5]



Fig. 4 Esp8266 (Wi-Fi module) [5]



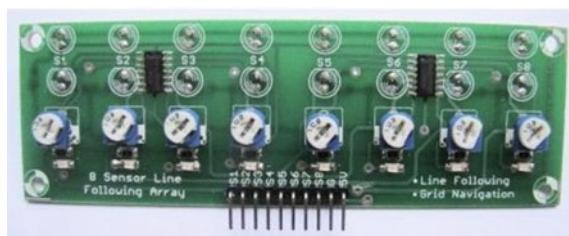
4.4 *ESP8266 (Wi-Fi Module)*

ESP8266 provides a complete described in Fig. 4 which is oneself included networking with a Wi-Fi, allowing it to either discharge all Wi-Fi networking functions from another procedure processor or host the application. When ESP8266 hosts the applying form, so when it's the application that is only within the device, having the ability to boot up straight from a flash that is outside. It has intertwined cache to meliorate the performance of the system in analogous operations and to minimize the memory conditions. Alternatively, serving as a Wi-Fi appliance, wireless access that's internet be included with any microcontroller-grounded design with easy connectivity through UART software or the CPU AHB ground screen.

ESP8266 provides a complete and oneself- included networking with a Wi-Fi, allowing it to either discharge all Wi-Fi networking functions from another procedure processor or host the application. When ESP8266 hosts the applying form, so when it's the application that is only within the device, having the ability to boot up straight from a flash that is outside. It has intertwined cache to meliorate the performance of the system in analogous operations, and to minimize the memory conditions. Alternatively, serving as a Wi-Fi appliance, wireless access that's internet be included with any microcontroller-grounded design with easy connectivity through UART software or the CPU AHB ground screen.

4.5 *IR Array*

IR detector array has six IR LEDs and six IR detectors. It can give two types of affairs. Analog output (direct analog data from IR detector) and digital output (using direct voltage comparator with potentiometer) as shown in Fig. 5.

Fig. 5 IR array [5]**Fig. 6** Servo motor [5]

4.6 Servomotor

Servo rotary linear actuators that provide precise control of linear or angular position, acceleration, and velocity as illustrated in Fig. 6. It actually consists of a suitable motor connected to a position feedback sensor. Servo machines in many cases are an assembly of mainly four things: position-sensor, a gearing set, a DC motor, and a control circuit. The career of servo motor can exactly be controlled more than those of standard DC engine, in addition, they usually have three cables (power, ground, and control).

The servo is controlled by sending a variable range electrical rate or pulse width modulation through a control line. There's a minimal palpitation, a maximum palpitation, and a reiteration rate. A mechanism of servo often just rotates 90° in either direction for a total of 180° movement.

4.7 L289N Motor Driver

L293D is just a Motor that is customary driver Motor Driver IC allowing DC motor to work a vehicle. L293D is an IC that is 16-pin that can handle a bunch of two DC engines all the while toward each path. In other words, one IC can control two DC motors. Binary ground H driver integrated circuit (IC). The L293D IC gets signals

through the microprocessor and transmits the sign that is relative to the motors. It has two voltage legs, one of that will be utilized to draw present for the working associated with the L293D, as well as the other is used to utilize a voltage towards the engines.

5 Software

The Arduino is designed with the Arduino Integrated Development Environment (IDE) written in the Java programming language for cross-platform operation. It all started with an IDE for processing and wiring. It is intended to provide an introduction to programming for artists and other non-software development beginners. It includes a program editor with similar functionality to syntax push, parentheses matching, automatic indentation, and provides a simple one-click medium to collect and upload programs to your Arduino board. Programs compiled in the full IDE for Arduino are called “sketches” [5]. The Arduino IDE supports both C++ and C languages with special conventions to structure the law. The Arduino IDE provides PC software that connects to a variety of common input and output routines. The Arduino IDE uses this technology to convert executable rules that are loaded onto the Arduino board using the board firmware’s bootloader system after compiling and linking using the GNU toolchain into a hexadecimal encoded text file and also for IDE deployments.

6 Methodology

The architecture of automatic waste dumping system as shown in Fig. 7, with this approach, the overall waste detection system is divided into four subsystems: the intelligent waste collection system, the vehicle system, the local base station, and the intelligent monitoring and controlling unit.

6.1 Smart Trash System (STS)

The IoT-based automatic waste unloading system is composed of four subsystems, and the main system that the rest operates is the Smart Trash System, and its functional unit is called the Smart Trash Bin. It consists of an ultrasonic detector, a Bluetooth module, and a Wi-Fi module. Detectors are used to locate trash in smart bins. Whenever the smart bin is full, the detector works, outputting a high voltage 25 V signal, which is transmitted via the Bluetooth module. This transmitted signal is entered by another Bluetooth module which is placed in the Vehicle System [5].

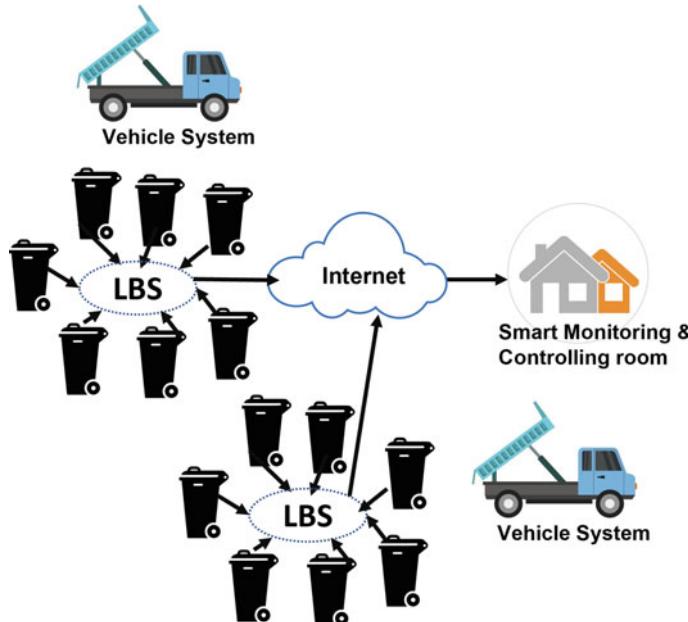


Fig. 7 Architecture of automatic waste dumping system

6.2 *Vehicle System (VS)*

The line follower vehicle system is a microcontroller control robot that detects and follows lines drawn on the floor. Paths consist of black lines with white edges (or vice versa). The control system used should detect the line and guide the robot to stay on course, and use a feedback environment to continuously correct erroneous movements to create a simple but effective infinite circle system. The robot is designed to walk along a very steep corner toward the bin, receive a signal via the Bluetooth module that the bin is full, and observe the situation to dispose of the trash from waste container via ultrasonic sensor.

6.3 *Local Base Station (LBS)*

The local base station is a base station for all the commerce match in which the transmission and receiving of data do between your smart trash system (STS) together with vehicle system (VS) which also inclusively deliver data to your Smart Monitoring and Controlling Hut (SMCH) for covering purpose. Local base section correspond of Hc-05 and Esp8266 [6].

6.4 *Smart Monitoring and Controlling Hut (SMCH)*

In the SMCH, the Wi-Fi module in the trash can receive the signal and then pass the signal over the internet to the cone controlling the cabin. Information and status are displayed in the monitoring and management site related to smart bins. Regarding Smart Monitoring and Hut interface control, important things such as status related to the entire smart bin are displayed.

7 Conclusion

This application of an intelligent garbage collection system using the Internet of Things ensures that the site is cleaned up soon when the scrap location reaches its maximum. However, if the tip is not culled within a certain period of time, the record is passed on to advanced organizations that can take necessary action against the public official. This arrangement also helps to hide fake reports, thus reducing damage to the entire operating system. This lessens the entire number of vehicle tour for scrap collection, reducing the overall cost associated with scrap collection. Helps keep the community clean over time. This is a design that is quite important in terms of creativity and ideas. We use IoT proposals to bring efficiency and unity of concept to this design. The layout specializes in the cleanliness of the trash can and the authentic introductory work it contains. As an extended operation, it covers the entire scrap collection system. We use ultrasonic detectors and other microcontrollers and processors similar to Arduino to analyze marital status and communicate information about it to directors and even host a scrap metal exchange. Another honestly essential issue of this layout is the internet gate designed in this type of manner that drivers and residents discover it convenient to hide unnecessary information in another location (as transmitted). From now on, software development plans based on the concept of the Internet of Things using electronic sensors are the backbone that will make the world incredibly supportive and unbelievable, all things considered. The intelligent system are evolved that include the human activities [7]. The driver alertness and other parameters will be involved in the future work. with the evolution of Quantum computer and Quantum electronics [8–10], the smart system will become smarter and free from any kind of interference from the hackers.

References

1. Dev A, Jasrotia M, Nadaf M, Shah R (2016) IoT based smart garbage detection system. *Int Res J Eng Technol* 3:12
2. Nithya L, Mahesh M (2016) A smart waste management and monitoring system using automatic unloading robot. *Int J Innov Res Comput Commun Eng* (An ISO 3297: 2007 Certified Organization) Website: www.ijircce.com, 4(12)

3. Ihara I (2008) Ultrasonic sensing: fundamentals and its applications to non-destructive evaluation (a draft). Nagaoka University of Technology, Jepang
4. Pan P, Lai J, Chen G, Li J, Zhou M, Ren H (2018) An intelligent garbage bin based on NB-IOT research mode. In: 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), IEEE, pp 113–117
5. Arduino SA (2015) Arduino. Arduino LLC, 372
6. Polianytsia A, Starkova O, Herasymenko K (2016) Survey of hardware IoT platforms. In: 2016 third international scientific-practical conference Problems of Infocommunications Science and Technology (PIC S&T), IEEE, pp 152–153
7. Paliwal S, Kalyan BS. Driver's activity detection system using Humanantenna
8. Kalyan BS, Kaur H, Pachori K, Singh B (2022) An efficient design of D flip flop in Quantum-Dot Cellular Automata (QCA) for sequential circuits. VLSI architecture for signal, speech, and image processing, p 253
9. Kalyan BS, Singh B (2021) Quantum-dot cellular automata-based encoder circuit using layered universal gates. In: Nanoelectronic devices for hardware and software security. CRC Press, pp 217–230
10. Kalyan BS, Kaur I, Singh B. Designing equivalent model of floating gate transistor for smart dust in rural areas. *Int J Comput Appl* 975:8887
11. Reis P, Pitarma R, Goncalves C, Caetano F (2014) Intelligent system for valorizing solid urban waste. In: 2014 9th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, pp 1–4

Constructing a Smart School Based on the Internet of Things Using RFID Technology



Soha Alhelaly

Abstract Concerns about the security and safety of students in schools promote the development of efficient systems, and smart schools need to be deployed. Controlling the movement of pupils between classrooms and labs as well as inside and outside a school is very difficult. This paper proposes a real-time pupil tracking and monitoring system for schools. The system uses Internet of Things (IoT) technology as well as cloud computing. Such a system will allow families, schools, and authorities to check student attendance, monitor, and track their movements in order to ensure their safety. Tags were detected when pupils carrying them and walking through the school and bus doors. However, the probabilities of detection when moving fast or running are 95% and 85%, respectively.

Keywords Smart school · Internet of Things · Cloud computing · RFID · Tracking system

1 Introduction

Child safety has always been a major concern, and a fertile area of research has gained much attention throughout the globe. The smart cities paradigm takes into consideration the need to provide a favorable environment for children's lives learning [2]. Having a smart city should take into account the need to provide safety, accuracy, and flexibility and enhance schools' effectiveness of the overall operation and service. Thus, countries try to improve and manage schools efficiently. Ensuring the safety and security of schools is a primary concern for society at large, and therefore, implementing smart schools is essential to smart cities. Student safety, the automatic recording of attendance, and tracking students are vital in a smart school. Therefore, several studies discuss the smart campus, which is an inevitable trend in the development of a digital campus. A smart campus will leverage IoT, cloud computing, and other technologies. The authors expound the concept of a smart campus, includ-

S. Alhelaly (✉)

College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
e-mail: s.alhelaly@seu.edu.sa

ing its overall architecture in [11, 14, 25]. However, monitoring students at school or between home and school is very difficult, especially if the school has several labs and classrooms. In addition, students sometimes move around the classrooms without any observing, which could be a significant problem when the students are children who may wander into a dangerous areas such as the swimming pool. This paper aims to provide a real-time tracking system application that will help ensure the safety of pupils.

The commute between home and school is a source of concern for parents: children may get on the wrong buses and alight at the wrong stops. In addition, bus drivers may not be able to identify all the children and may not realize whether a child is missing, or parents may send a driver to pick up their children, which poses a challenge in guaranteeing that the child has left the school. The risk of kidnapping and delays of school buses or vehicles due to road traffic has put additional stress on families. Moreover, anecdotal evidence indicates that the number of children who goes missing while commuting between home and school has drastically increased worldwide [13]. According to the FBI, there were 337,195 entries from the National Crime Information Center (NCIC) for children missing in 2021. However, in 2020, the total number of entries for missing children was 365,348 [16]. In previous years, there have been several accidents in Saudi Arabia where primary school students entered swimming pools without supervision and no one noticed that they drowned [17]. In October 2018 [10], a seven-year-old child accessed the swimming pool at a school in Saudi Arabia during classes and drowned. Since that tragic accident, the Ministry of Education ordered the closure of all school swimming pools until new safety rules are applied. Moreover, six people have been accused of negligence after a four-year-old Emirati boy drowned at a local school in Sharjah last year [26]. In another case, an eight-year-old boy was found dead in a school bus, which prompted an investigation by the education department. The school implemented a procedure, which was subsequently followed by other schools, requiring that two staff members check each school bus in the morning to determine whether any child has been left behind [17].

Today, several technologies have evolved to help people and communities. However, parents cannot monitor their children in school during school hours, and teachers cannot know students' exact locations at school. To the best of our knowledge, there is no study that ensures the safety of children while they commute between home and school and while at school. Our system solves this problem. Students carry a tag supported by RFID technology. Teachers and parents can track the movement of the children within the school and receive notifications if the child is late for any reason or enter a prohibited area, and also they can track the children between the school and home. A manager can also track students and receive a weekly or monthly report showing the percentage of attendance or absence for each student or for a specific class. In order to meet the requirement so flow cost and low-power consumption, our proposed system uses RFID technology. It requires minimal Internet connectivity, works in real time, and is built using basic off-the-shelf components. Currently, a comprehensive solution that is both cost and performance efficient continues to be a challenge. Most existing systems are limited to tracking children either inside buses

or tracking the buses and private vehicles. We present a solution to the problem by introducing a comprehensive workflow for tracking children inside schools and buses.

The remainder of this article is organized as follows: Sect. 2 presents some of related work, Sect. 3 proposes the system, and Sect. 4 presents the conclusion.

2 Related Work

Several studies have proposed solutions to track students, using IoT-based systems and a combination of the Global Positioning System (GPS), Raspberry Pi, Radio Frequency Identification (RFID), Global System for Mobile communication (GSM), General Packet Radio Service (GPRS), or by implementing different prediction modes to compute the arrival time of school buses [3, 8, 18, 24]. Other studies have proposed combining RFID technology with fingerprint identification algorithms that send notifications when students enter or exit school buses [1]. Yet other researchers have proposed IoT-based solutions for tracking student attendance, car parking availability, and for estimating waiting times at bus stops [9, 28]. There are many types of student monitoring attendance systems. For example, one of these systems is closed-circuit television (CCTV), which utilizes cameras in areas to ensure security, such as in stores and banks. The authors of [20] presented the design and implementation of an automated classroom surveillance and security system with CCTV camera and passive infrared (PIR) sensors to monitor students' behavior and observe teacher performance during classroom activities. However, this type of system requires employees to monitor in real time and view recordings. However, the system tracks movements but does not identify who is moving. Other studies propose face recognition for student authentication, for example, [15] proposes an attendance system in school that utilizes a face recognition technique that uses the Discrete Wavelet Transform (DWT) with the Discrete Cosine Transform (DCT) to extract student facial features. The study by [12] focuses on the development of automatic facial recognition systems for office door access control. This method could be used to monitor student attendance but not for tracking and monitoring. Other papers suggest using quick response (QR). In [4], a smart event attendance system is proposed for a university, using QR code and GPS technology to speed up the process of tracking student attendance. In [22], biometric fingerprint authentication is proposed for an automated attendance system. Other researchers have proposed a system that uses RFID technology. The use of RFID facilitates monitoring class attendance, where the system will automatically send student data to a database [27]. However, these studies do not track students with in or around the school when they move between classrooms, playgrounds, labs, or toilets.

Among the above-mentioned technologies, RFID systems are rapidly gaining ground, especially in the area of intelligent transportation systems [21]. The authors in [21] showed that RFID-based systems exhibit many advantages compared with other techniques. The versatile and advantageous RFID technology has proved to be

very useful in the field of intelligent transportation and particularly for enhancing driving safety. RFID is a low-cost wireless technology that connects a multitude of things, enabling consumers and businesses to engage, identify, locate, transact, and authenticate products [6]. The general RFID market has seen a considerable growth over the past years in terms of the number of RFID tags sold. RFID sensors are a new paradigm for the IoT. They have a limited cost and negligible maintenance, which make the map pealing for numerous applicative scenarios in manufacturing, logistics, health care, agriculture, and food [7]. A combination of RFID and computing technology is called an RFID system. An RFID system consists of five components: a tag (electronic label), an antenna (for tag reading) that has an identification code that can be transmitted to the reader (RFID antenna scan be used to collect information about any item), a reader (to read tag information), communication infrastructure (enables the RFID/reader to function through IT infrastructure), and application software (interface/application/user database). The readers sometimes referred to as interrogators or scanners, send and receive RF data to and from tags via antennas. An RFID tag is a small electronic device that is referred to as a transponder. The tag can be attached to any object and information is collected via the chip and can be transmitted wirelessly. An RFID tag can be active (battery powered), passive (not powered), or semi-passive (hybrid) [5].

3 The Proposed System

The proposed product is a real-time application that monitors and tracks students around the school and during the bus commute from home to school and from school to home. The idea is to provide passive tags in bracelets worn by students and to set up RFID scanners in all school and bus doors. The school's entrance door, bus doors, and all rooms within the school, such as toilets, labs, classrooms, and playgrounds should be equipped with scanners. Once students pass near a scanner, it will record them with the time and date in the system. Thus, the school will have a record of each student's movement history and attendance. In this real-time application, if some school areas, such as the swimming pool or labs, are prohibited at certain times, reader device will send an alarm or notification to a designated person at the school and to the parent, and appropriate action can be taken. This system monitors each student whether they are at school, riding the bus, entering the school, or leaving the school, ensuring child safety and providing records for each student's movement around the school and percentage of each student's attendance as weekly, monthly and/or annual reports sent to teachers and parents. All data is stored in an accessible database school.

Fig. 1 Block diagram of the proposed system

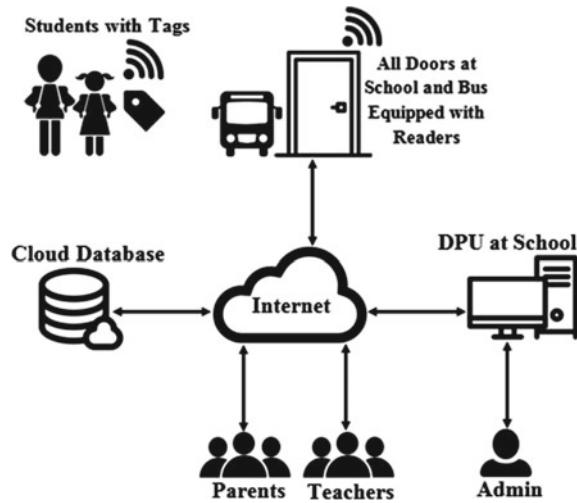
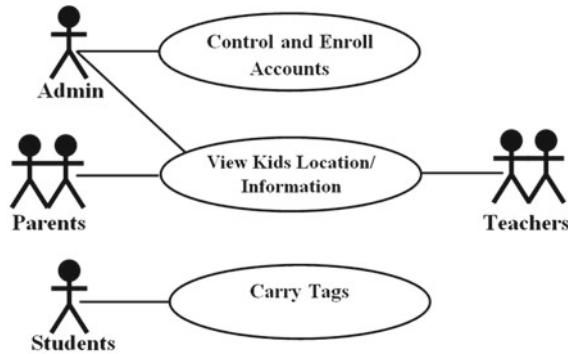


Fig. 2 Functional entities of the proposed system



3.1 The System's Block Diagram

As shown on Fig. 1, the proposed system consists of three components: all school and bus doors equipped with readers, tags attached to students, and a data processing unit (DPU) at the school. The system will acquire student data at each scanning point.

As shown on Fig. 2, there are four main entities that interact with each other, while the students move within the school and between the school and home: students with RFID tags, administrators, teachers, and parents.

3.2 *Implementation*

Each student is equipped with an RFID passive tag that is identified in the system by a unique ID (UID) and that stores the student's information. At each door, the RFID reader reads the student's UID when the door is opened. Thus, the tags are checked at each point. For every detected tag at the doors (entrance or exit), the RFID reader reads the student's UID, and the acquired information is sent from the scanner to the DPU at the school. The RFID scanner is connected to the DPU via a database that securely stores the student's information. The administrators, teachers, or parents can access all tracking information live from within the application.

3.2.1 **Detailed Design**

The system consists of two main parts: hardware and software. The hardware consists of the readers at each school and bus door, the passive tag attached to each student, and the DPU. For the implementation, we used an RFID reader that is commonly used in diverse industrial applications. The tags distributed to students work with the most common RFID readers and can be attached to practically anything imaginable. Each tag comes with a tag ID (TID) that cannot be changed, and it has sufficient memory for writing and reading additional information when needed. The DPU located at the school is a gateway computer able to process and transfer large amounts of data at high speeds. It hosts the main application as well as all the management operations and the database.

The software consists of two components: the front end and the back end. The front end component is the application interface for users (teachers and parents) to monitor the students. The application was developed using the Java programming language. There are two login modes: admin and user. The administrator can enroll students, parents, and teachers and can also link parents to children. This interface is dedicated to school authorities (or admins) only. Admins have complete access to the lists of enrolled students and their parents. The user login (parents and teachers) enables the monitoring of students' locations while commuting or at school. The back end component includes the developed application, the communication protocols, and the database. The EPC global Gen2 protocol is embedded within the application as a communication protocol for acquiring information from the tags and sending it to the database. For the database component, we chose MySQL as the database management system (DBMS) and deployed it over the RedHat OpenShift online cloud platform [23]. The database is stored and deployed on the Heroku cloud which is a cloud platform that offers platform as a service (PaaS), but a local copy is stored on the DPU for backup purposes [19].

Table 1 Effects of students' movements on tag detection

Scenario	Detected tags	
Walking	20 out of 20 tags	Were detected
Moving fast	19 out of 20 tags	Were detected
Running	17 out of 20 tags	Were detected

3.3 System Testing and Discussion

The system was initially tested in a laboratory environment. The testing started by adjusting the RFID reader to analyze the detected tags. The setup simulated a practical deployment scenario. We implemented a number of tests to analyze the effects of the tags with respect to the RFID reader. The RFID reader was configured to work efficiently once the students moved close to it and the students' information was sent to the DPU. Numerous tag reading trials were tested. This information is accessible by parents, teachers, and administrators, and action can be taken when required. The proposed system works on the Android platform; the application requires a connection to the Internet and an active phone. An actual experiment was performed at a school and on bus doors. Furthermore, we implemented a number of tests to analyze the different effects of students' movements on the ability of the readers to read the tags. The test was successful, collecting the students' information, and sending it to the DPU. The results are summarized in Table 1, which shows the probabilities of detecting the tags when students walk through a door. However, the probabilities of detection when moving fast or running are 95% and 85%, respectively.

4 Conclusion and Future Work

Controlling the movement of children between classrooms and labs as well as inside and outside a school is very essential. The proposed system monitors and tracks students within the school, when entering or exiting the bus, the school's main door, classrooms, toilets, etc. Furthermore, the system can send an alert to the teachers and parents when the student enters a prohibited area such as the swimming pool or a floor that is permitted to a specific grade. Thus, such a system will help teachers, parents, and school managers to track students entering or exiting classrooms. Each student will have a passive bracelet tag like a watch, which does not need to be recharged. The bracelet tag will be read by the RFID scanner and the information will be registered in the system by date and time. For each student, the proposed system provides weekly or monthly reports, such as the attendance and absence percentages. A set of test case scenarios were implemented to examine the efficiency of the system. Tags were detected when pupils carrying those tags walk through doors. However, the probabilities of detection when moving fast or running are 95% and 85%, respectively.

References

1. Ahmed A, Parvez MR, Hasan MH, Nur FN, Moon NN, Karim A, Azam S, Shanmugam B, Jonkman M (2019) An intelligent and secured tracking system for monitoring school bus. In: 2019 International conference on computer communication and informatics (ICCCI). IEEE, pp 1–5
2. Ahmed Z, Rawat A, Kumari P (2021) An analysis of IoT based smart cities. *Int J Eng Trends Appl (IJETA)* 8(4)
3. Anitha G, Ramesh S, Mohanavel V, Diwakaran S, Maheswaran U (2022) A wireless communication protocol enabled school bus tracking system using internet of things support. In: 2022 International conference on advances in computing, communication and applied informatics (ACCAI). IEEE, pp 1–6
4. Ayop Z, Lin CY, Anawar S, Hamid E, Azhar MS (2018) Location-aware event attendance system using QR code and GPS technology. *Int J Adv Comput Sci Appl* 9(9)
5. Chechi D, Kundu T, Kaur P (2012) The RFID technology and its applications: a review. *Int J Electron Commun Instrum Eng Res Dev (IJECIERD)* 2:109–120
6. Costa F, Genovesi S, Borgese M, Michel A, Dicandia FA, Manara G (2021) A review of RFID sensors, the new frontier of internet of things. *Sensors* 21(9):3138
7. Cui L, Zhang Z, Gao N, Meng Z, Li Z (2019) Radio frequency identification and sensing techniques and their applications—a review of the state-of-the-art. *Sensors* 19(18):4012
8. Dhanasekar N, Valavan C, Soundarya S (2019) IoT based intelligent bus monitoring system. *Int J Eng Res Technol (IJERT)* 7(11):1–4
9. Dong ZY, Zhang Y, Yip C, Swift S, Beswick K (2020) Smart campus: definition, framework, technologies, and services. *IET Smart Cities* 2(1):43–54
10. Gazette S (2018) Al-Issa orders schools to shut down unsafe swimming pools. <https://saudigazette.com.sa/article/544610/SAUDI-ARABIA/Al-Issa-orders-schools-to-shut-down-unsafe-swimming-pools>. (Online) Accessed 7 Mar 2021
11. Habibzadeh H, Kaptan C, Soyata T, Kantarci B, Boukerche A (2019) Smart city system design: a comprehensive study of the application and data planes. *ACM Comput Surv (CSUR)* 52(2):1–38
12. Ibrahim R, Zin ZM (2011) Study of automated face recognition system for office door access control application. In: 2011 IEEE 3rd International conference on communication software and networks. IEEE, pp 132–136
13. Jisha R, Jyothindranath A, Kumary LS (2017) IoT based school bus tracking and arrival time prediction. In: 2017 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 509–514
14. Liu YL, Zhang WH, Dong P (2014) Research on the construction of smart campus based on the internet of things and cloud computing. *Appl Mech Mater* 543:3213–3217. Trans Tech Publ
15. Lukas S, Mitra AR, Desanti RI, Krishnadi D (2016) Student attendance system in classroom using face recognition technique. In: 2016 International conference on information and communication technology convergence (ICTC). IEEE, pp 1032–1035
16. National Center (2021) National center for missing & exploited children—find missing children. <https://www.missingkids.org/HOME>. (Online). Accessed 31 Dec 2021
17. NEWS A (2021) Another child dies in school bus. <http://www.arabnews.com/featured/news/905781>. (Online) Accessed 1 Nov 2021
18. Ning Y, Zhong-Qin W, Malekian R, Ru-chuan W, Abdullah AH (2013) Design of accurate vehicle location system using RFID. *Elektronika ir Elektrotechnika* 19(8):105–110
19. Nokeri TC (2022) Deploying a web app on the cloud. Apress, Berkeley, CA, pp 215–221. https://doi.org/10.1007/978-1-4842-7783-6_12
20. Olamide OA, Asafe YN, Olawale OI, Akinleye AO (2017) Networking cctv cameras & passive infra-red sensors for e-classroom monitoring system: proactive approach to quality assurance in education system. *Int J Adv Netw Appl* 8(05):3213–3219
21. Pedraza C, Vega F, Manana G (2018) PCIV, an RFID-based platform for intelligent vehicle monitoring. *IEEE Intell Transp Syst Mag* 10(2):28–35

22. Rahman MM (2021) Study on introducing biometric fingerprint authentication in automated student attendance system, 4:121–131. BP International
23. RedHat (2021) Red hat openshift. <https://www.redhat.com/en>. (Online). Accessed 11 Apr 2021
24. Selvam M, Yadahalli AR, Dindi MM, Nithin B (2022) Iot enabled school bus monitoring and notification system. In: ICDSMLA 2020. Springer, pp 1205–1216
25. Su K, Li J, Fu H (2011) Smart city and the applications. In: 2011 international conference on electronics, communications and control (ICECC). IEEE, pp 1028–1031
26. Times K (2019) Six accused after 4-year-old boy drowns at school pool in UAE. <https://www.khaleejtimes.com/uae/six-accused-after-4-year-old-boy-drowns-at-school-pool-in-uae>. (Online). Accessed 7 Jan 2022
27. Ula M, Pratama A, Asbar Y, Fuadi W, Fajri R, Hardi R (2021) A new model of the student attendance monitoring system using RFID technology. *J Phys: Conf Ser* 1807:012026. IOP Publishing
28. Valks B, Arkesteijn MH, Koutamanis A, den Heijer AC (2021) Towards a smart campus: supporting campus decisions with internet of things applications. *Build Res Inf* 49(1):1–20

Effect of Confining Walls on Settling Permeable Rigid Isolated Semi-torus Particle Applying Immersed Boundary Method (IBM)



Pooja Yadav, Sudeshna Ghosh, Amit Sharma, and Rekha Panghal

Abstract The authors explored the confining wall's effect on sedimentation of permeable rigid isolated semi-torus particle submerged in incompressible, Newtonian, viscous fluid in a 2D rectangular fluid domain. The immersed boundary method (IBM) was used to conduct numerical studies of the related problem. It was noticed that by increasing the permeability of the structure immersed in the fluid, their corresponding settling(terminal) velocity also increases. Another important finding was that confining walls enforce an extra hindering effect on sedimenting particle, which concludes that terminal velocity of the particle decreases when the gap between the confining walls decreases.

Keywords Immersed boundary method · Sedimentation · Fluid-structure interaction · Semi-torus · Settling velocity

1 Introduction

The primary centre of attention of ongoing work is to study the effect of confining walls on settling of semi-torus shaped particle in between two impermeable walls. The approach that we have adapted to solve the present problem is the immersed boundary method (IBM). In this research, we investigate the bounding wall's effect and effect of permeabilities on the terminal velocity for semi-torus shaped particle.

In particular, we concentrate our attentiveness on interaction between fluid-structure where impermeable walls are present in the surrounding, effects the sedimenting particle are remarkably, e.g. a compressed section of a flow channel. In real-world problems, a lot of practical applications are present. Some of them are sand particles sedimentation in an hourglass, blood cells transport in constricted arteries, granular flow in a hopper, etc. In a nutshell, we can say that fluid-structure interaction has numerous example in industrial and real-life problems.

P. Yadav · S. Ghosh (✉) · A. Sharma · R. Panghal
Amity School of Applied Sciences, Amity University, Haryana, India
e-mail: sudeshnagh108@gmail.com

The driving factors for conducting this study are the immense applicability of biofilm and the lack of study done for settling of particle which is not of regular geometric shape. We are going to look at the semi-torus shape. The idea behind the chosen semi-torus shaped particle is because of its occurrence of semi-torus particle in many real-life applications. The current work is motivated by the detachment of biofilm in erosion mode. The individual particle, after settling due to gravity, can be of any arbitrary shape. For this reason, we conducted our study with this shape.

Sedimentation of impermeable and permeable irregular shaped particles is perceived in a lot of real life along with industrial processes. In [1], the research was carried on about colloidal aggregates, catalyst pellets and model macromolecules terminal velocity of permeable aggregates. In oceans, sedimentation of marine particles is highly permeable in nature. The sedimentation of permeable agglomerations is dominant for modelling particle transport in waste water plants treatment. The sedimentation of permeable structures also studied by many other researchers. In [2], researchers computed the terminal velocity of a permeable regular shaped structure. The mathematical simulations of the hydromechanical porosity of granular materials were conducted to study the effect dependence on the shape of structure [3]. A fascinating numerical approach to create granular permeable method was discussed in [4].

The current problem is solved by applying immersed boundary method (IBM). It is robust for simulating the interaction between structure and fluid problems involving intricate structures in an incompressible, viscous fluid. Many sedimentation problems are solved by using IBM. The sedimentation of impermeable single and multiple circular particles was studied by applying the IBM technique [5]. In [6], the sedimentation of impermeable semi-torus shape particle was researched by using IBM. This numerical approach has been applied by many researchers like Layton [7], Dillon and Fauci [8], Kim and Peskin [9], Stokie [10], Ghosh [11] for permeable structures. To the best knowledge of authors, no research has been done to investigate the effect of confining wall on settling of semi-torus structure make use of IBM.

2 Immersed Boundary Method (IBM)

Charles Peskin coined IBM in 1972 to research the heart valves interaction with blood. This approach is numerical definition as well as mathematical formulation with supposition fluid being incompressible, viscous and Newtonian. In this method, structural variables are handled by the Lagrangian approach and variables of fluid by the Eulerian approach. Here, we discussed governing equations in a compendious way, a detailed explanation described in [5].

2.1 Mathematical Equations

Mathematical formulation for the problem is discussed in this section. The model is justifiable for $\Delta\rho = \rho_s - \rho_f \ll \rho_f$, where ρ_s (density of structure) and ρ_f (density of fluid).

The variables and parameters applied are:

- Ω : Fluid domain,
- Γ : Structure immersed,
- $\delta(\mathbf{y}) = \delta(x) \cdot \delta(y)$: Two 1D dirac delta function Cartesian product,
- \mathbf{F}_{Ibm} [g/s²] : Force created by the particle immersed in the fluid,
- $\mathbf{Y}(\mathbf{s}, t)$ [cm] : Location of structure immersed,
- $\mathbf{y} = (x, y)$ [cm] : Ω coordinates,
- μ [g(cms)⁻¹] : Fluid viscosity,
- \mathbf{w}^* [cms⁻¹] : Velocity of fluid,
- $M(\mathbf{s})$: Additional Lagrangian mass density occurring because of structure.
- \hat{l} : Unit vector alongside in vertical direction.

The Navier–Stokes equation is used to research incompressible fluid flow:

$$\rho_f \frac{\partial \mathbf{w}^*}{\partial t} + \rho_f \mathbf{w}^* \cdot \nabla \mathbf{w}^* = \mu \nabla^2 \mathbf{w}^* - \nabla p + \mathbf{f}_{\text{Ibm}} + \mathbf{f}_{\text{gravity}}. \quad (1)$$

$$\nabla \cdot \mathbf{w}^* = 0 \quad (2)$$

Forces originated due to the structure and owing to gravitational force are expressed by \mathbf{f}_{Ibm} and $\mathbf{f}_{\text{gravity}}$, respectively, and the expressions are:

$$\mathbf{f}_{\text{Ibm}} = \int_{\Gamma} \mathbf{F}_{\text{Ibm}} \delta(\mathbf{y} - \mathbf{Y}) \, ds, \quad (3)$$

$$\mathbf{f}_{\text{gravity}} = -g \hat{l} \Delta \rho = -g \hat{l} \int_{\Gamma} M(\mathbf{s}) \delta(\mathbf{y} - \mathbf{Y}) \, ds, \quad (4)$$

Evolution of structure with time is shown as:

$$\frac{\partial \mathbf{Y}}{\partial t} = \int_{\Omega} \mathbf{w}^* \delta(\mathbf{y} - \mathbf{Y}) \, dy. \quad (5)$$

In [11], the relationship between the permeability (k) with porous slip velocity \mathbf{w}_p^* was discussed. The relationship is shown in Eq. (6).

$$\mathbf{w}^* p = \frac{k}{\mu} \nabla p' \quad (6)$$

where $\nabla p'$ —pressure gradient and k —permeability of structure. The structure with permeability accompanies the alike equations of governing as an impermeable excluding evolution structure equation demonstrated as:

$$\frac{\partial \mathbf{Y}}{\partial t} = -\mathbf{w}^* p + \int_{\Omega} \mathbf{w}^* \delta(\mathbf{y} - \mathbf{Y}) d\mathbf{y}, \quad (7)$$

2.2 Setup of the Problem

Figure 1 illustrated the setup of the studied problem. The sedimentation of particle to start is positioned along the centreline in a confined medium.

2.3 Analytical Results

In [11], the relationship between the settling velocity of impermeable and permeable structure was discussed (8).

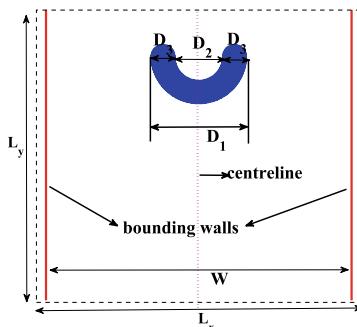
$$W_{s,p} = \sqrt{W_{s,0}^2 + Pk} \quad (8)$$

where

$$P: \frac{3\pi^2 g(\rho_s - \rho_f)}{16\mu} W_{s,0}.$$

$W_{s,0}$ and $W_{s,p}$ are analytical settling velocities of an impermeable and permeable structure, respectively.

Fig. 1 Sedimentation of an isolated semi-torus particle



From Eq. (8), we obtain:

$$W_{s,p}^2 = W_{s,0}^2 + Pk. \quad (9)$$

Equation (9) suggests that terminal velocity will increase with increasing the permeability of the structure.

In this study, the authors will compare the computed settling velocity (V_c) with ($W_{s,p}$) to study the accuracy of the results obtained.

3 Numerical Results

We now highlight the results obtained for the setup illustrated in Fig. 1. Variables considered for the numerical results are mentioned in Table 1.

3.1 *Sensitivity Study of Half Wall Gap for Fixed Density Difference Between Fluid and Immersed Structure ($\Delta\rho$) and Varying Permeability (k)*

For $\mu = 0.01$, $\Delta\rho = 0.01$ and permeability in the range $[10^{-8}, 5 \times 10^{-5}]$, effect of the permeability on the terminal velocity is demonstrated in Fig. 2. The bounded walls are separated by distance $\Delta W = 0.32$ [cm]. Here, we have noticed that the value of permeabilities and their correspondent terminal velocity is proportional to each other, which can be seen from Eq. (9). This reliability holds good for permeable regular structure also [11].

Table 1 Variables used for numerical results

ρ_f	1	[g/cm ³]
ρ_s	1.01	[g/cm ³]
D_1	0.28	[cm]
D_2	0.14	[cm]
D_3	0.07	[cm]
μ	0.01	[g/cm s]
$\Delta\rho$	0.01	[g/cm ³]

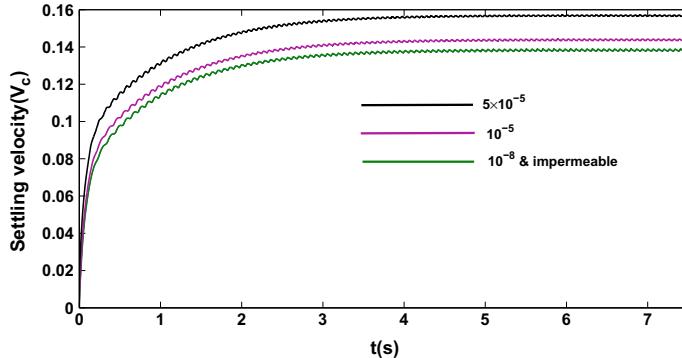


Fig. 2 Variation of terminal velocity varying as progresses time for different values of permeabilities

3.2 Sensitivity Study for Fixed Permeabilities and Viscosities and Variable Wall Gap (ΔW)

The gap between the bounded walls is varied in the range [0.32, 1.28] and the value of permeability, $k = 5 \times 10^{-5} [\text{cm}^2]$ fixed.

In literature [12], it has been documented that the bounding walls establish an additional inhibiting effect on the structure undergoing the process of sedimentation. We have noticed that when the gap between the confining walls decreases, their corresponding settling velocity decreases, and when the gap between the confining walls increases, their correspondent terminal velocity increases. The results obtained are shown both pictorial (Fig. 3) and in tabular form (Table 2).

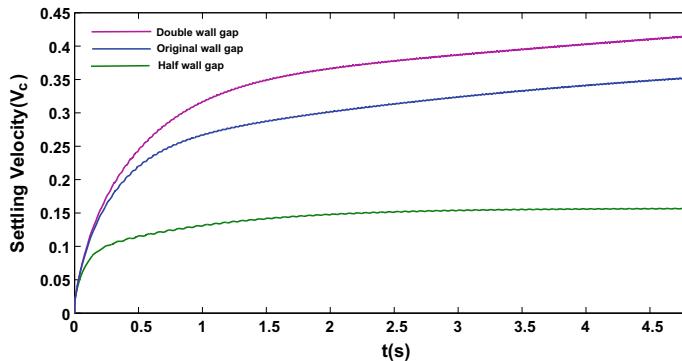


Fig. 3 Variation of terminal velocity varying as progresses time for changing the gap between the walls

Table 2 Comparison between numerically simulated (V_c) and analytically simulated ($W_{s,p}$) for different wall gap (W) and permeabilities (k)

ΔW	k	V_c	$W_{s,p}$
0.32	5×10^{-5}	0.1569	0.1778
	10^{-5}	0.3358	0.3448
	10^{-8}	0.3181	0.3260
0.64	5×10^{-5}	0.3635	0.4063
	10^{-5}	0.0086	0.0083
	10^{-8}	0.0043	0.00415
1.28	5×10^{-5}	0.4463	0.4896
	10^{-5}	0.4312	0.4401
	10^{-8}	0.4243	0.4280

4 Conclusion

The primary centre of attention of this work was to study the confining wall effect on permeable semi-torus shaped particle using IBM. The major considerations done are that for fixed wall gap, as the permeabilities increases, their corresponding settling velocity also increases. After that, we fixed permeability and viscosity and varying the gap between the walls. We have noticed that as gap between the confining walls decreases, corresponding settling velocity decreases because sedimenting particle faces retarding effect from the walls. The results derived from this research are physically justified.

In future, authors will expand the study for planktonic shaped particle. Also, will research the dynamics of two interacting permeable particles under the effect of confining walls.

References

1. Matsumoto K, Suganuma A (1977) Settling velocity of a permeable model floc. *Chem Eng Sci* 32(4):445–447. [https://doi.org/10.1016/0009-2509\(77\)85009-4](https://doi.org/10.1016/0009-2509(77)85009-4)
2. Masliyah JH, Polikar M (1980) Terminal velocity of porous spheres. *Can J Chem Eng* 58(3):299–302. <https://doi.org/10.1002/cjce.5450580303>
3. Garcia X, Akanji LT, Blunt MJ, Matthai SK, Latham JP (2009) Numerical study of the effects of particle shape and polydispersity on permeability. *Phys Rev E* 80(2):021304. <https://doi.org/10.1103/PhysRevE.80.021304>
4. Pilotti M (1998) Generation of realistic porous media by grains sedimentation. *Trans Porous Media* 33(3):257–278
5. Ghosh S, Stockie JM (2015) Numerical simulations of particle sedimentation using the immersed boundary method. *Commun. Comput. Phys.* 18(2):380–416. <https://doi.org/10.4208/cicp.061113.050115a>

6. Ghosh S, Yadav P (2021) Study of gravitational settling of single semi-torus shaped particle using Immersed boundary method. *Appl Math Comput.* <https://doi.org/10.1016/j.amc.2021.126643>
7. Layton A (2006) Modelling water transport across elastic boundaries using an explicit jump method. *SIAM J Sci Comput* 28(6):2189–2207. <https://doi.org/10.1137/050642198>
8. Dillon R, Fauci L (2000) A microscale model of bacterial and biofilm dynamics in porous media. *Biotechnol Bioeng* 68(5):536–537. [https://doi.org/10.1002/\(SICI\)1097-0290\(20000605\)68:5<536::AID-BIO68>3.0.CO;2-1](https://doi.org/10.1002/(SICI)1097-0290(20000605)68:5<536::AID-BIO68>3.0.CO;2-1)
9. Kim Y, Peskin CS (2006) 2-D parachute simulation by the immersed boundary method. *SIAM J Sci Comput* 28(6):2294–2312. <https://doi.org/10.1137/S1064827501389060>
10. Stockie JM (2009) Modelling and simulation of porous immersed boundaries. *Comput Struct* 87(11):701–709. <https://doi.org/10.1016/j.compstruc.2008.11.001>
11. Ghosh S (2020) Immersed boundary method for a permeable sedimenting circular particle between two parallel rigid walls. *Prog Comput Fluid Dyn Int J* 20(1):20–28. <https://dx.doi.org/10.1504/PCFD.2020.104708>
12. Chhabra RP, Agarwal S, Chaudhary K (2003) A note on wall effect on the terminal falling velocity of a sphere in quiescent Newtonian media in cylindrical tubes. *Powder Tech* 129:53–58. [https://doi.org/10.1016/S0032-5910\(02\)00164-X](https://doi.org/10.1016/S0032-5910(02)00164-X)
13. Peskin CS (2002) The immersed boundary method. *Acta Numer* 11:1–39. <https://doi.org/10.1017/S0962492902000077>
14. Fogelson AL (1984) A mathematical model and numerical method for studying platelet adhesion and aggregation during blood clotting. *J Comput Phys* 56(1):111–134. [https://doi.org/10.1016/0021-9991\(84\)90086-X](https://doi.org/10.1016/0021-9991(84)90086-X)
15. Wang J, Layton A (2009) Numerical simulations of fiber sedimentation in Navier-Stokes flow. *Commun Comput Phys* 5(1):61–83. <http://www.global-sci.com>
16. Persson PO, Strang G (2004) A simple mesh generator in MATLAB. *SIAM Review* 329–345. <https://doi.org/10.1137/S0036144503429121>
17. Peskin CS (1972) Flow patterns around heart valves: a numerical method. *J Comput Phys* 10(2):252–271. [https://doi.org/10.1016/0021-9991\(72\)90065-4](https://doi.org/10.1016/0021-9991(72)90065-4)
18. Ben Richou A, Ambari A, Lebey M, Naciri JK (2005) Drag force on a circular cylinder midway between two parallel plates at $Re < < 1$. Part 2: moving uniformly (numerical and experimental). *Chem Eng Sci* 60(10):2535–2543. <https://doi.org/10.1016/j.ces.2003.10.031>
19. Pianet G, Arquis E (2008) Simulation of particles in fluid: a two-dimensional benchmark for a cylinder settling in a wall-bounded box. *Euro J Mech B Fluids* 27:309–321. <https://doi.org/10.1016/j.euromechflu.2007.07.001>

Hybridized Shuffled Frog Leaping Algorithm for Solving Facility Location Problem for Maternal Healthcare



Ankit Chouksey, A. K. Agrawal, and Arkaprava Ray

Abstract In this paper, the shuffled frog leaping algorithm hybridized with a state-of-the-art solver is applied to solve the large-scale facility location-allocation problem for maternal healthcare. The problem considered is about to minimizing the total cost of establishing the various type of facilities and travelling costs incurred by mothers-to-be (MTBs) while providing them the required services within the maximum allowable distance using limited capacity of the facilities. The problem is formulated as a mixed-integer linear programming (MILP) mathematical model. The proposed model is NP-hard and combinatorial, and would thus require unmanageable computational effort in optimal planning of a real-world maternal healthcare network. To obtain good-quality solutions for large-sized problems in a reasonable amount of time, the shuffled frog leaping algorithm, and a population-based metaheuristic, is used in this paper. The binary version of this metaheuristics is used in this paper to determine the location of the facilities, and allocation is carried out using a state-of-the-art solver. To evaluate the efficiency and effectiveness of the proposed hybrid approach, extensive experiments are conducted on randomly generated problem instances. The computational results demonstrate that the proposed metaheuristic outperformed the solver for the large size problem instances. However, the objective function value for small size problem deviated by less than 10% from the optimal objective function value but required much lesser time. Since the real-world problems are large in size, the proposed hybridized approach is quite competitive both in terms of efficiency and efficacy.

Keywords Swarm intelligence · Combinatorial and numerical optimization · Frog leaping · Maternal healthcare network planning · Evolutionary approaches

A. Chouksey (✉) · A. K. Agrawal
Indian Institute of Technology (BHU) Varanasi, Uttar Pradesh, India
e-mail: ankitchouks.rs.mec17@itbhu.ac.in

A. Ray
Jadavpur University Kolkata, West Bengal, India

1 Introduction

According to the World Health Organization, 830 women die every day around the world from preventable causes that occur during pregnancy and childbirth [11]. Since this number is significant, the government has to plan meticulously for upbringing a right level of infrastructure for maternal healthcare. The maternal healthcare planning framework considered in this study, therefore, proposes the establishment of exclusive healthcare facilities for mothers-to-be (MTBs). Maternal healthcare facilities are planned independently from other healthcare facilities with enough capacity to ensure that all pregnant women are covered, and they do not have accessibility challenge, particularly at their stage. The problem is mathematically modeled as a mixed-integer linear programming (MILP) problem.

Hierarchical facility location models have been widely used for locating healthcare facilities. A little amount of research is available addressing facility location issues related to maternal care. For perinatal care, Galvão et al. [6] created a hierarchical location model for Rio de Janeiro, Brazil. Their model took the facilities to be uncapacitated. Galvão et al. [8] refined this model by considering the limited capacity of the facility for non-routine delivery and neonatal care. Later, Baray and Cliquet [1] suggested a capacitated p -median problem for placing maternal healthcare facilities in France. While comparing the results of theoretically located maternity hospitals with the actual distribution of facilities in the country, they observed a significant difference between the two. Jang and Lee [9] developed a model for locating neonatal critical care facilities in Korea. Jang and Lee [9] tweaked their model to account for multi-period planning of such facilities. It is observed that the p -median formulation has been adopted by a good majority of the investigators. As an alternative, the problem was treated as a fixed-charge facility location-allocation problem [5]. Chouksey et al. [3] developed a capacitated fixed-charge facility location-allocation for maternal care by considering no overburdening of the facilities. In another work, [4] allowed the overburdening of the facility but with a high penalty cost. Each modeling approach has its own advantage. The MILP model presented by Chouksey et al. [4] lacks in providing a solution in a reasonable time for problems involving a large number of locations.

In the present study, an effort has been made to solve the large-scale maternal healthcare facility location problem in a computationally efficient manner with the help of a metaheuristic. Traditional solution techniques, like the Branch and Bound, can be utilized to tackle small and medium-sized facility location problems (FLPs). For large size problems, these might not be able to find a solution in a reasonable amount of time. The complexity of MILP models for FLPs increases with problem size, and the problem become intractable sometimes. To address this issue, iterative local search [2], neighborhood search [8], and other metaheuristics are used to solve large size FLPs. Population-based metaheuristics inspired by nature have also drawn a lot of interest recently. These include simulated annealing, Tabu search, and others. Usually, evolution or swarm intelligence are the driving forces behind these

metaheuristics. Researchers have been particularly interested in swarm intelligence-based algorithms because of their capacity for exploration and exploitation [13]. A thorough examination of the literature has revealed that a number of swarm intelligence-based metaheuristics are very successful at resolving a variety of combinatorial optimization issues. Out of all these approaches, the shuffled frog leaping algorithm, a population-based metaheuristic, have been used in the present work to solve the considered maternal healthcare facility location-allocation problem. To compare the effectiveness and utility of these metaheuristics, extensive computational experiments have also been conducted.

2 The Maternal Healthcare Facility Location-Allocation Problem

MTBs require routine check-ups, blood tests ultrasound, etc., during pregnancy. MTBs with approaching delivery due date will require a standard or cesarean delivery services. MTBs, who have an unplanned cesarean and those who have complication in their pregnancy, will need neonatal care services. MTBs thus may require primary care (service type 1), routine and scheduled cesarean deliveries (service type 2), and complex deliveries with neonatal care (service type 3). These services are provided through a variety of hierarchical and sequentially inclusive maternity healthcare facilities, particularly in India. Service type 1 (primary care) is provided by facility type I (SC), and service types 1 and 2 (i.e., regular and scheduled cesarean deliveries besides primary care) are provided by facility type II (PHC). All three service types 1, 2, and 3 (unplanned cesarean deliveries with neonatal care in addition to primary care, regular and scheduled cesarean deliveries) are available at facility type III (CHC). The capacity for providing a type of service is the same at facilities of the same type but may be different from those at other types of facilities. Excess allocations at a facility beyond the capacity are not desirable, but the reality is otherwise. Therefore, the planning framework considered allows it but at a significant penalty cost (P) cost. Besides the direct allocation of MTBs to the facilities, the referral cases from lower-level facilities to higher-level facilities are also to be allocated.

To address the accessibility issue, maximum coverage distance is used to set a limit on the distance an MTB can travel. This will also be the case in the event of a referral. Establishing a facility type at a location will incur a fixed cost and will be more for an advanced facility type. The objective will be to plan the facilities at some of the identified locations such that the entire cost of establishing the facilities as well as the total cost incurred on visiting them by MTBs and the penalty cost associated with the excess allocations is the minimum. This problem is further referred to as Maternal Healthcare Facility Location-Allocation Problem (MHFLP). For the mathematical formulation of MHFLP, the following notations are being used.

Sets and Indices

- I : set of locations of MTBs, $I = \{1, 2, \dots, a\}$, indexed by i .
- J : set of locations of potential facilities, $J = I$, indexed by j, k and n .
- L : set of service types offered, $L = \{1, 2, 3\}$, indexed by p, l and m .
- T : set of types of facility, $T = \{I, II, III\}$, indexed by t and u .

Parameters

- M : a big number.
- P : penalty cost per additional MTB allocated beyond the capacity of any service type available at a facility.
- d_1 : a limit on the maximum distance to be covered by an MTB during a non-referral visit.
- d_2 : a limit on the maximum distance to be covered by an MTB during the referral visit.
- d_{ij} : distance between locations $i \in I$ and $j \in J$.
- d_{jk} : distance between facility locations $j \in J$ and $k \in J$.
- C_{ij} : travel cost incurred by an MTB for visiting the facility at a location $j \in J$ from its current location $i \in I$.
- C_{jk} : travel cost incurred by an MTB for referral visit from current facility location $j \in J$ to a referral facility at location $k \in J$.
- F_j^t : fixed cost on establishing a facility of type $t \in T$ at location $j \in J$.
- Q^{lt} : capacity of service type $l \in L$ available with facility type $t \in T$.
- W_i^l : number of MTBs at a location $i \in I$ requiring service type $l \in L$.
- θ^{lm} : proportion of referrals for service type $m \in L$ from service type $l \in L$, where $m > l$

$$\alpha_{ij} = \begin{cases} 1, & \begin{array}{l} \text{if a facility for non-referral visit at a location } j \in J \\ \text{is within the coverage distance of an MTB} \\ \text{at a location } i \in I \text{ (i.e., } d_{ij} \leq d_1\text{),} \end{array} \\ 0, & \text{otherwise} \end{cases}$$

$$\beta_{jk} = \begin{cases} 1, & \begin{array}{l} \text{if a referral facility at a location } k \in J \\ \text{is within the coverage distance of a lower level facility} \\ \text{at location } j \in J \text{ (i.e., } d_{jk} \leq d_2\text{),} \end{array} \\ 0, & \text{otherwise} \end{cases}$$

Decision Variables

- x_{ij}^l : number of mothers-to-be at location $i \in I$ being allocated to a facility at a location $j \in J$ to receive service type $l \in L$.
- x_{jk}^{lm} : number of mothers-to-be receiving service type $l \in L$ at facility $j \in J$ who require referral visit to a facility $k \in J$ for a higher service type $m \in L$.
- \underline{x}_j^l : number of mothers-to-be allocated to a facility at a location $j \in J$ to receive service of type $l \in L$ but beyond the available capacity of that service type

$$y_j^t = \begin{cases} 1, & \text{if a facility of type } t \in T \text{ is located at } j \in J, \\ 0, & \text{otherwise.} \end{cases}$$

The mathematical model of MHFLP:

$$\text{Minimize} \quad \left\{ \begin{array}{l} \sum_{j \in J} \sum_{t \in T} F_j^t y_j^t + \sum_{i \in I} \sum_{j \in J} \sum_{l \in L} c_{ij} x_{ij}^l + \\ \sum_{k \in J} \sum_{j \in J} \sum_{m \in L} \sum_{l \in L} c_{jk} x_{jk}^{lm} + \sum_{j \in J} \sum_{l \in L} P \bar{x}_j^l \end{array} \right\} \quad (1)$$

Subject to

$$\sum_{j \in J} x_{ij}^l = W_i^l, \quad \forall i \in I, l \in L \quad (2)$$

$$\sum_{k \in J} x_{jk}^{lm} = \theta^{lm} \left\{ \sum_{i \in I} x_{ij}^l + \sum_{p \in L} \sum_{n \in J} x_{nj}^{pl} \right\}, \quad (3)$$

$$\forall j \in J, l, m \in L, p < l < m$$

$$\sum_{i \in I} x_{ij}^l + \sum_{k \in J} \sum_{m \in L} x_{kj}^{lm} - \bar{x}_j^l \leq \sum_{t \in T} Q^{lt} y_j^t, \quad \forall j \in J, l \in L \quad (4)$$

$$\sum_{t \in T} y_j^t \leq 1, \quad \forall j \in J \quad (5)$$

$$x_{ij}^l \leq \sum_{t \in T} \alpha_{ij} y_j^t M^{lt}, \quad \forall i \in I, j \in J, l \in L \quad (6)$$

$$x_{jk}^{lm} \leq \sum_{t \in T} \beta_{jk} y_k^t M^{mt}, \quad \forall j, k \in J, l, m \in L, m > l \quad (7)$$

$$Q^{2I} = Q^{3I} = Q^{3II} = 0 \quad (8)$$

$$x_{ij}^l, x_{jk}^{lm}, \bar{x}_j^l \geq 0, \quad \forall i \in I, j \in J, l \in L, m \in L \quad (9)$$

$$y_j^t \in \{0, 1\}, \quad \forall j \in J, t \in T \quad (10)$$

The flow of MTBs to various types of facilities is depicted in Fig. 1. The four expressions in the objective function (1) indicate the fixed cost of creating the health-care facilities, the cost of traveling to the facilities, the referral cost, and the penalty cost, respectively. Constraint (2) ensures that the demand is completely met, whereas

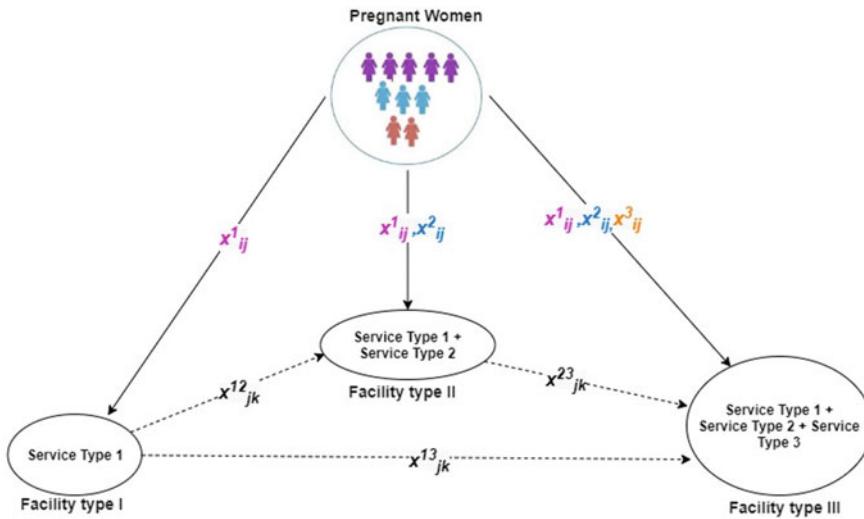


Fig. 1 Representation of flow of MTBs to different types of facilities

constraint (3) defines the referral of MTB from a lower-level facility to a higher one. Constraint (4) is a capacity constraint, and it also determines the excess number of MTBs beyond the capacity at a facility for each service type. According to the constraints (5), only one type of healthcare facility can be established at a given location (5). Constraint (6) indicates that an MTB can only be assigned to a facility type if it is located within the coverage distance, whereas constraint (7) states the same for referral situations. The true features of the facility types are captured by constraint (8), which shows service types 2 and 3 not being available at facility type I and service type 3 not being available at facility type II. The nature of the decision variables is defined by constraints (9) and (10).

3 Shuffled Frog Leaping Algorithm

Shuffled frog leaping algorithm (SFLA), a very well-known metaheuristic, is based on memetic evolution of a number of frogs searching the most food-rich area in wetlands. SFLA was proposed in 2003 by Eusuff and Lansey to combine the advantages of particle swarm optimization and the memetic algorithm. A collection of frogs (solutions), grouped memplexes, make up the population. Different memplexes are viewed as various frog cultures (memes), each of which conducts a local search. Songs, concepts, catchphrases, trends in clothing, and techniques for creating pots or constructing arches are a few examples of memes. Each memplex contains

frogs that go through a process known as memetic evolution. These frogs carry out local exploration of the problem space using certain methods that permit the transference of meme among local individuals. The merits of both algorithms are combined together to make SFLA an efficient and fast local optimizer. Solving facility location problems using SFLA was always an interest of various researchers. Due to the advantage of fast local search, SFLA found its application in many areas such as set covering problems (Broderick et al. 2015), optimizing bridge desk repairs (Emad 2006), optimizing large-scale water supply problems (Gunhui et al. 2009), closed-loop supply chain network (Santhosh et al. 2020), energy-efficient dynamic consolidation of virtual machines in cloud data centers [11], and opposition-based learning [13].

The considered maternal healthcare facility location-allocation problem addresses two main issues (i) determining the location of the different types of facilities, and (ii) the allocation of MTBs to these facilities. The MHFLP model is developed by considering both decisions. The MHFLP model was solved using Gurobi 9.0.2, which is the state-of-the-art mathematical programming problem solver. Numerical experiments carried out revealed that the solver could provide optimal solutions for small and some medium-sized problem instances within a reasonable time. However, with the increase in the problem size, the computational time started increasing exponentially. The solver could not even produce a feasible solution in reasonable amount of time for many large-sized instances. Therefore, obtaining an optimal or a good quality solution in a reasonable amount of time, a hybridized SFLA approach is being devised to solve the model efficiently and effectively.

Due to the special structure of the MHFLP model, the problem can be partitioned into two problems: (i) location and (ii) allocation problems. The location problem deals with binary decision variables, and the allocation problem deals with continuous decision variables. The literature and experimentation suggest that the main complexity of the model is governed by binary variables. To reduce the complexity of the model, the binary variables (location decisions) are handled by the SFLA meta-heuristic. After deciding the location of the facilities, the allocation of MTB is made by solving the allocation problem using the solver. Since the allocation problem is an LP problem, the solver does not take too much time comparatively to solve the problem. Solving allocation problems using a solver is advantageous because it provides optimal allocation decisions in the light of the location decisions arrived at. The steps of the proposed hybridized algorithm are shown in Fig. 2. The parameters used in the algorithm are decided based on pilot experiments. The number of frogs and number of memplexes are taken as 50 and 10, respectively. The local search is carried out for 10 iterations in each memplex. After the local search, the sigmoid function is used to convert the decimal values into binary values. The process is carried out for a maximum of 2 h of CPU in reporting the best solution obtained within this time limit.

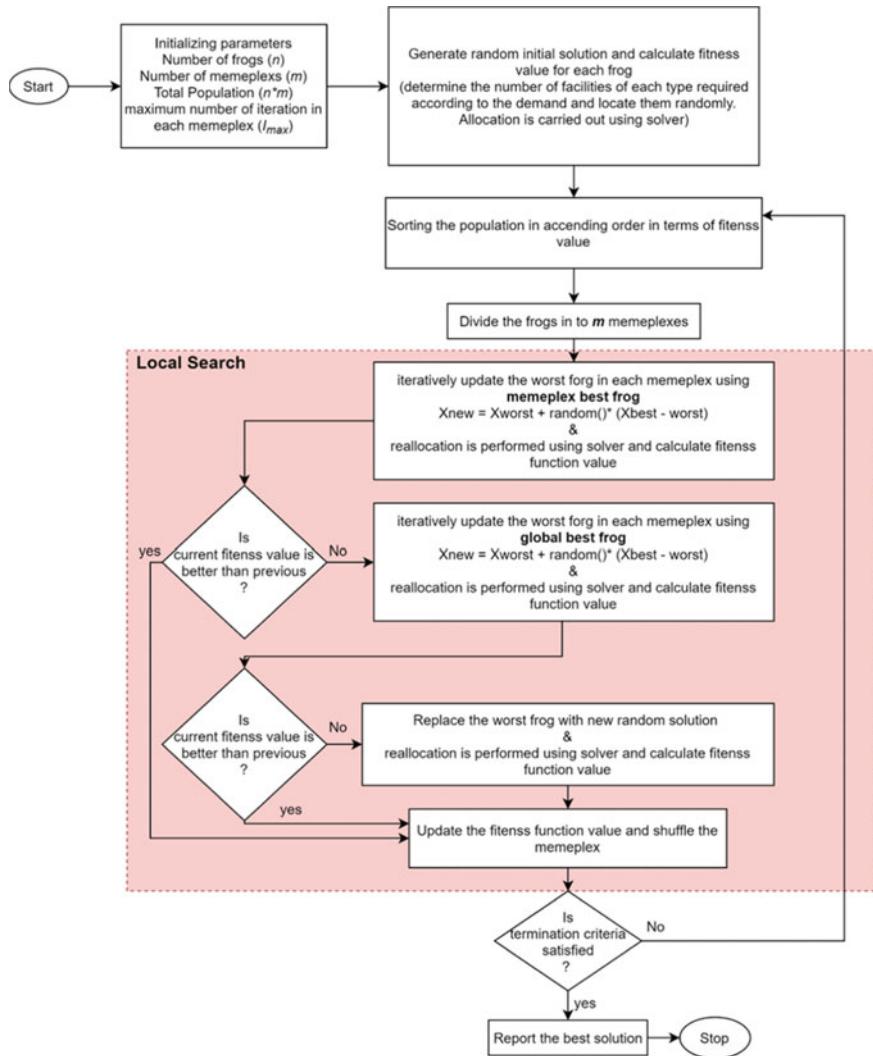


Fig. 2 Step involved in hybridized shuffled frog leaping algorithm

4 Computational Experiments

The comparative performance of the proposed solution approach is reported in this section. For solving the MHFLP mathematical model, the Gurobi optimization solver with Python 3.8 is used. Besides, the metaheuristic was coded in Python 3.8.

The computational experiments were carried out on the supercomputer “PARAM-SHIVAY” with a configuration of 2* Intel Xeon SKL G-6148, 4 core, 2.4 GHz, 192 GB DDR4 2666 MHz memory. The experimentation was carried out on 100, 200, 300, and 400 location problems. For each problem class, five test instances were generated. The details of the generation of data and other problem parameters can be found in Chouksey et al. [3].

Table 1 shows the experimental results obtained from the use of SFLA (discussed in Sect. 3) and Gurobi solver. The ‘Gap’ column reflects the difference between the upper bound and lower bound values obtained from Gurobi. Columns labelled as ‘percentage deviation’ represent the deviation of the objective value obtained from SFLA compared to the Upper Bound value yielded by Gurobi. The negative values of percentage deviation represent the cases where the proposed metaheuristics perform Gurobi. During the experimentation, SFLA was found to yield different solution and CPU time due to randomness involved in the approach. To be fair, average performance is reported for SFLA after solving each problem instance for five times. According to the findings in Table 1, the proposed hybrid SFLA’s reported solution quality is fairly comparable to that of the Gurobi solver. The variance for the smaller size problem is under 10%. However, hybrid SFLA reported superior UB for large and medium-sized problems. The MHFLP formulation, when applied directly (using the Gurobi solver), provides a better bound for small problems (100 and 200 nodes), but it is unable to handle most medium problems (300 nodes), as well as all large problems (400 nodes). In contrast, the proposed SFLA performed much better and provided good quality feasible solutions within the imposed time limit. The use of Gurobi alone could not converge even in 2 h. The hybridized approach could yield the solution in much lesser CPU time.

5 Conclusions

This paper is devoted to the development of a mathematical programming model for addressing a maternal healthcare facility location-allocation problem. The paper addresses the issue of the availability and accessibility of healthcare services. The objective is to minimize the total cost of establishment of healthcare facilities and travel costs of mothers-to-be. This paper proposes a shuffled frog leaping algorithm hybridised with a Gurobi solver to solve the large-scale facility location problem. The proposed hybrid SFLA approach was found to perform better than Gurobi in terms of solution quality. One may try to further improve the performance of the proposed metaheuristic by incorporating various search techniques such as local search, neighbourhood search, and Tabu search.

Table 1 Experimental results of proposed solution approaches

Problem size	Problem class	Problem number	Gurobi			SFLA	Deviation (%)
			Lower bound ($\times 10^7$)	Upper bound ($\times 10^7$)	Gap (%)	Avg. Obj ($\times 10^7$)	
Small	100	1	1.12	1.13	1.32	1.24	9.05
	100	2	1.1	1.12	2.14	1.23	9.04
	100	3	1.16	1.18	1.91	1.29	8.95
	100	4	1.08	1.1	1.74	1.19	7.9
	100	5	1.1	1.12	2.01	1.24	10.51
Small	200	6	2.16	2.19	1.2	2.41	9.73
	200	7	2.14	2.18	1.8	2.36	8.26
	200	8	2.16	2.18	1.15	2.39	9.27
	200	9	2.09	2.12	1.31	2.3	8
	200	10	2.14	2.18	1.66	2.36	8.23
Medium	300	11	3.17	3.49	9.13	3.49	-0.06
	300	12	3.21	3.25	1.43	3.52	7.95
	300	13	3.22	3.64	11.46	3.55	-2.72
	300	14	3.1	3.16	1.85	3.4	7.25
	300	15	3.21	3.61	11.04	3.52	-2.64
Large	400	16	4.21	5.73	26.55	4.6	-19.75
	400	17	4.24	5.09	16.7	4.64	-8.82
	400	18	4.25	4.94	13.97	4.65	-5.96
	400	19	4.15	4.4	5.8	4.58	3.91
	400	20	4.24	5.16	17.8	4.64	-10.08

References

1. Baray JÔ, Cliquet G (2013) Optimizing locations through a maximum covering/p-median hierarchical model: maternity hospitals in France. *J Bus Res* 66(1):127–132. <https://doi.org/10.1016/j.jbusres.2012.09.003>
2. Brito J, Ochi L, Montenegro F, Maculan N (2010) An iterative local search approach applied to the optimal stratification problem. *Int Trans Oper Res* 17(6):753–764. <https://doi.org/10.1111/j.1475-3995.2010.00773.x>
3. Chouksey A, Agrawal AK, Tanksale AN (2022) An optimization and simulation hybrid approach for maternal healthcare facility location-allocation in the Indian context. *Int J Oper Res* 1(1):1
4. Chouksey A, Agrawal AK, Tanksale AN (2022) A hierarchical capacitated facility location-allocation model for planning maternal healthcare facilities in India. *Comput Ind Eng* 167(February):107991
5. Daskin MS (2011) Network and discrete location: models, algorithms, and applications
6. Galvão RD, Espejo LGA, Boffey B (2002) A hierarchical model for the location of perinatal facilities in the municipality of Rio de Janeiro. *Eur J Oper Res* 138(3):495–517

7. Galvão RD, Espejo LGA, Boffey B, Yates D (2006) Load balancing and capacity constraints in a hierarchical location model. *Eur J Oper Res* 172(2):631–646. <https://doi.org/10.1016/j.ejor.2004.09.049>
8. Hansen P, Mladenović N, Moreno Pérez JA (2010) Variable neighbourhood search: methods and applications. *Ann Oper Res* 175(1):367–407. <https://doi.org/10.1007/s10479-009-0657-6>
9. Jang H, Lee JH (2019) A hierarchical location model for determining capacities of neonatal intensive care units in Korea. *Socioecon Plann Sci* 68(March):1–14. <https://doi.org/10.1016/j.seps.2019.03.001>
10. Luo JP, Li X, Chen MR (2014) Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers. *Expert Syst Appl* 41(13):5804–5816
11. Maternal health WHO (2021) Maternal health
12. Sharma TK, Pant M (2018) Opposition-based learning embedded shuffled frog-leaping algorithm. In: *Soft computing: theories and applications*. Springer, Singapore, pp 853–861
13. Sonamani Singh T, Yadava RDS (2018) Application of PSO clustering for selection of chemical interface materials for sensor array electronic nose. In: *Soft computing: theories and applications*. Springer, Singapore, pp 449–456

Key Observation to Prevent IP Spoofing in DDoS Attack on Cloud Environment



T. Sunitha, V. Vijayashanthi, M. Navaneethakrishan, T. A. Mohanaprakash, S. Ashwin, T. R. Harish, and Emmanuel A. Stanes

Abstract The utilization of Cloud to put together and execute Distributed Denial of Service (DDoS) assaults is well-known among programmers in the current scenario. In the existing framework, networks can use Border Gateway Protocol (BGP) declaration setups to change the internet courses and an assortment of other sources in a precise manner. In this paper investigates in detail how an organisation can regulate this data source and the connection where the traffic enters an organisation. In this proposed work, PKI-Public key validation techniques with Advanced Encryption Standard were utilized (AES) and technique works on the basis that the courses are influenced by network control, thus the organization getting the caricature traffic affects it when it gets traffic, rather than depending on switches that are not influenced at all. The proposed method which is does not correspond to the existing follow back approaches requires no progressions to send hardware or participation from different organizations. This procedure work best when the caricature traffic begins from not many sources, as is normal in intensification DDoS assaults and also the experimental findings demonstrated that the approach has a 99.3% accuracy rate for detecting internal and external low/high rate faked DDoS assaults and offers higher performance.

T. Sunitha

Department of Computer Science and Engineering, P. B. College of Engineering, Chennai, India

V. Vijayashanthi

Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College (Autonomous), Chennai, India

M. Navaneethakrishan

Department of Computer Science and Engineering, St. Joseph College of Engineering, Sriperumbudur, Chennai, India

T. A. Mohanaprakash (✉)

Department of Computer Science and Engineering, Panimalar Engineering College, Varadharajapuram, Poonamallee, Chennai, India

e-mail: tamohanaprakash@gmail.com

S. Ashwin · T. R. Harish · E. A. Stanes

Department of Computer Science and Engineering, Panimalar Institute of Technology, Varadharajapuram, Poonamallee, Chennai, India

Keywords IP spoofing · Network security · Cloud computing · Key observation

1 Introduction

IP spoofing is the most well-known approach to change the source address of Internet Protocol (IP) bundles to camouflage the transporter's personality, duplicate another PC structure, or both. It's a methodology that unethical performers constantly use to ship off DDoS assaults against an objective contraption or the including framework [1–5].

IP spoofing is a strategy consistently utilized by programmers to follow the servers with DDoS (Fig. 1) and man-in-the-center going after procedures in cloud climate. DDoS is a cloud-unequivocal attack in which the attack source is mostly more than one; distinct machines attack on a client data by sending packages with huge vertical data. Such pursues make the resources out of reach to the client by overwhelming the association with bothersome traffic [3, 6].

Cloud computing gives various advantages to both associations and people by taking advantage of the exceptionally proficient and accessible processing assets at a restricted expense. There are a few popularized Cloud administrations offering different arrangements, restricted by the previously mentioned downside in the supporting stage. Inertia is based on asset conflict between Virtualization Technology and the speed of the Internet service. As a result, such applications generate massive amounts of fluctuating data at an extremely fast rate, and due to the delay in getting to the Cloud stage for further investigation and decision, the chances of training the IoT device to make responsive decisions on a consistent basis may be too high [7, 8].

It is observed that the cybercriminals develop an engaging stage to the local area Cloud for executing an assortment of attacks. Programmers, in turn are utilizing the local area Cloud administrations to stop the intruders. Local area Cloud administrations give adaptable, on-request limit, and assets which can be laid out in only a couple of moments [4]. Besides, by increasing the transmission capacity from 1 to

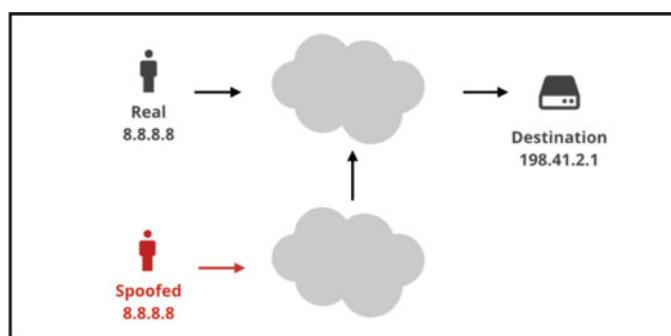


Fig. 1 IP spoofing in distributed denial of service (DDoS) attacks

10 Gbps, organizations are assaulted multiple times than what is possible with individual compromised gadgets, like home switches or IoT cameras as per the report. As indicated by [5], Cloud-based botnets were utilized in one-fourth of all DDoS attacks in Europe from July 2017 to June 2021. In February 2021, it was found that the most firms had employed DDoS constructed capacity out the gigantic Cloud [9].

The issue of ‘Misuse and vindictive utilization of Cloud administrations’ is termed as one of the ten worst cloud security worries by the Cloud Security Alliance (CSA) [11]. It further construes the issue right from the cybercriminals the local area Cloud. The biggest archived Distributed Denial of Service assault is depicted in Fig. 2 in which AWS is forestalled with 2.2 terabits of DDoS. The attacks in June 2020 might have delivered large number of clients’ administrations out of reach for a dubious timeframe. The cloud DDoS assault was likewise 44% greater than the information volume announced before by the Amazon organization, as indicated by AWS [12]. Other organizations too, have been affected by cloud DDoS assaults. Globally, many companies were exposed to a progression of DDoS assaults in August and September 2020. Kaspersky analyzed the quantity of DDoS attacks in Q3 and Q2 of 2020 to Q3 of 2019 and observed that all the assaults in Q3 were 1.5 times higher than in Q3 of 2019 [13].

According to Netscout, a shift in DDoS attack tactics occurred in Q1 and Q2 of 2020, with faster, shorter, and harder-hitting multi-vector attacks which were expected to persist. Attackers are now focusing their attention on online platforms and services that are essential in this digital age, such as education, financial services, e-commerce, and healthcare. The fraudulent money obtained from the attacks is really big money and hence such DDos attacks have increased enormously [14, 15]. Bulletproofs “2019 Annual Cyber Security Report”, DDoSattacks are very costly. If a small business is hit by a DDoS assault, they could lose anywhere from \$120,000 to

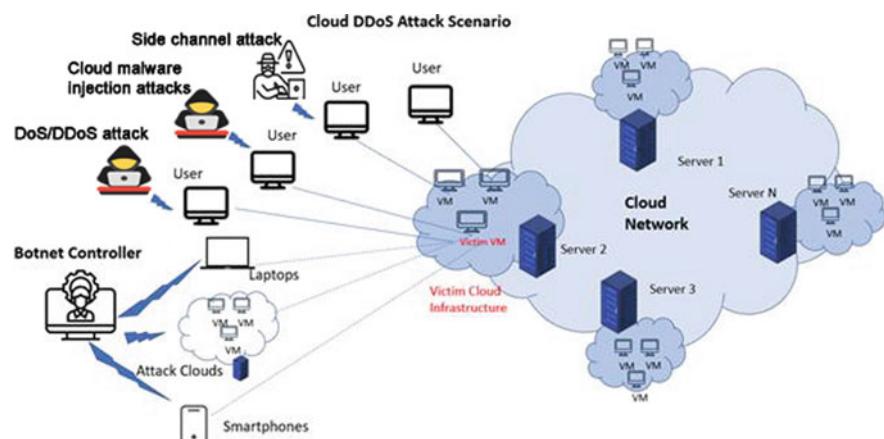


Fig. 2 DDoS architecture in a cloud environment [10]

\$2 million. Furthermore, according to the CISCO Visual Networking Index, global DDoS attacks are predicted to triple by 2023 [10].

The GitHub code hosting platform was struck by the greatest DDoS attack ever on 2021. Aggressors counterfeited GitHub's IP address and used it to send questions to memcached servers, which are regularly used to accelerate data set driven sites. The information from those GitHub inquiries was improved by a component of around 50 on the servers. Subsequently, the aggressor moved as much as 51 kB to the objective for every byte sent by him. GitHub received 1.35 TB/s of traffic, which forced the site to be down for 10 min [15].

The design and ubiquity of organization administrators' (i.e., ISPs') procedures for identifying and managing faked Internet traffic are notable and reported. Sadly, there is little data available about the province of IP mocking recognition and anticipation in the cloud infrastructure field. The motivation behind this ebb and flow research is to perceive the way that an organization could control this data source the looking connection where traffic enters an organization to all the sources and find faked traffic sources. The reason for this review with its fundamental concentration, is to fill that opening.

2 Related Literature and Significance of This Study

The quantity of examination papers on the issue of local area cloud security has coupled with the notoriety and universality of local area cloud stages and administrations over the last ten years. A portion of these papers, for example, [10, 12–16], have investigated an assortment of safety and protection issues regarding the local area cloud's organization, stockpiling, engineering, and programming. Other exploration has focused on more unambiguous subjects and issues, for example, distributed refusal of administration assaults [17–19] and IP satirizing [20–23] locally.

In the study, DDoS attacks in distributed computing [9, 11, 14], and the mitigation drawn and their headways were compared to DDoS assaults and IP satirizing in the cloud were accounted. The order, revelation, expectation, and countermeasure techniques generally incorporated a complete assessment [9, 11, 14]. Likewise, the distribution incorporated a point by point rundown of assets. To gage other elective arrangements, utilizing a scientific classification and an assortment of measurements were employed. According to the creator, the study was a new endeavour to highlight the requirement for DDoS relief techniques coordinating appropriate asset executives. During the assault, the asset executives and staggered data stream were significant [24].

The review portrays the headway of cloud innovation; however, it features various challenges that have emerged as security weaknesses, for example, IP parodying [6]. A broad study is taken up in this paper to analyze DDoS assault counteraction and location procedures, with a particular spotlight on the IP satirizing attack [23]. Another proficient method for the discovery and anticipation of disseminated

disavowal of-administration assaults in the cloud will be taken on because of this convincing finding [23].

Khan [21] this research inspected a scope of systems for identifying IP parodying and distinguished a particular philosophy for forestalling IP caricaturing that exhibits vindictive way of behaving [25]. It also uncovered an abnormal component of faked IP bundles in the cloud and introduced a host-based working framework fingerprinting approach as a hearty and compelling procedure to battle toward home bound parcels in working frameworks [26, 27]. In addition, the paper focused on the normal execution of tasks in forthcoming distributed computing [28].

An objective in scientific categorization, a flow dataset setup review, and existing IDS resources and limits were completely given by the analyst. To expand the effectiveness of IDS, this next rendition of IDS will be utilized to develop datasets and will likewise be utilized to reflect network with high precision in future datasets. Hanan-Hindy and partners [29]. The goal of [30] is to propose an adaptive and lightweight approach for accurately detecting low and high rate spoofed DDoS attack traffic. The method is used in a closed cloud environment, Agrawal N et al. [30].

Therefore, the ends have very restricted down to earth utility, best case scenario, and give knowledge into the genuine certifiable plausibility of Cloud-put together attacks that depend with respect to IP caricaturing. Since they are for the most part composed according to the viewpoint of the guard, they give close to nothing, if any, understanding into the exact objectives and strategies utilized by programmers while executing Cloud-based DDoS or potentially IP ridiculing campaigns [8, 30, 31]. Yet, while making viable true protections, a complete information on the 'why' and 'how' according to the programmer's viewpoint is useful, if not essential 100% of the time.

The work given in this paper attempts to address the weaknesses of existing exploration that have been noted previously. The conversation and consequences of this study will give reliable information on the issues addressed to both network safety analysts and professionals, subsequently aiding the advancement of a more secured Internet. For an effective anti DDoS assault, IP caricaturing is forestalled. There are six accessibles, including network observing, firewalls and organization assault blockers, bundle separating, local area key framework confirmation, and utilizing distant administration check techniques. In this proposed work, PKI-Public key validation techniques with Advanced Encryption Standard were utilized (AES) [32–34]. The Comparison table of IP spoofing and DDoS techniques and limitations shown in Table 1.

3 Proposed Method

Cryptography has been used to communicate secret messages between warring nations, users, and organizations, among other things; as a result, it has become a significant concern in national security and regulations. Cryptography is becoming a necessity for cloud-based applications as the demand for secured data exchanges

Table 1 Comparison table of IP spoofing and DDoS techniques and limitations

Existing work	Techniques used	Limitations of method
[6]	Border gateway protocol (BGP)—a network with several peering links can utilize to roughly pinpoint the Internet's sources of spoof traffic	Inability to detect packet loss, saturated transit services
[7]	Host-based OS fingerprinting	Complexity is high
[27]	Efficient spoofed mitigation scheme (ESMS), which combines the bloom filter trust model and the TCP probing method	Lacking of getting spoofed packets in the router
[28]	Revised hop count filtering (HCF) mechanism	Unable to identify HCF against replicated real-world attacks from real attack
[31]	The mixed rate IP spoofing DDoS assault in the IoT environment is detected and mitigated using an adaptable and lightweight approach	Vulnerable to numerous other attacks, including routing, sinkhole, and selective forwarding attacks

across computer networks grows for healthcare, economic, and other essential objectives. The laws and regulations governing cryptography are changing dramatically all across the world. Import and export limitations on cryptographic products are being debated and amended.

The proposed security mechanism is employed at both ends of the cloud-based network, namely the source or sender and the destination or receiver. At the source end, records or information that should be imparted are scrambled utilizing the Advanced Encryption Standard (AES) and afterwards confirmed utilizing a mystery key. ‘View all’ of the IP addresses from the LAN are put away in parallel records, most of which are numeric information documents with the expansion ‘arp’ design. IP extraction is the most common way of separating all of the IP addresses from the bat documents. This is done line by line and put in a cluster object to avoid the commotion in the preprocessing system. When this interaction is finished, the source picks an organization beneficiary objective by hiding the IP address and port number while sending the encoded information to the objective end. The cushion IP address with the recipient end framework arrangement is checked utilizing the disguised IP address and the objective end’s port number at the objective end. The confirmation check is finished by providing the specific mystery key in the organization once the recipient IP and cradle IP have been coordinated. Figure 3 portrays the IP parodying detail. This proposed worldview is separated into two stages: End-of-Sender-Phase and End-of-Receiver-Phase.

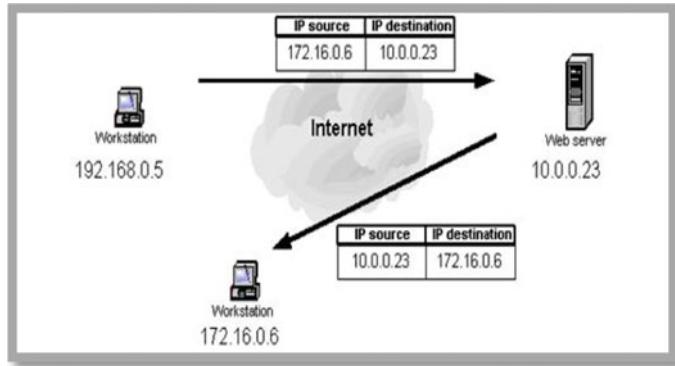


Fig. 3 IP spoofing specification

3.1 *End-of-Sender-Phase*

In this stage, the information to be transferred is scrambled utilizing the Advanced Encryption Standard (AES) technique, which is done to keep unapproved people from getting to the setting of the information, and just the approved individual can see utilizing the mystery key given by the relating shipper, as displayed in Fig. 4. All IP addresses in this intra-network have a novel succession of whole numbers isolated by full stops that are utilized to recognize every PC that utilizes IP to associate over an organization, and this is saved in double document design (.arp). Accordingly, the IP extraction fix is done in line with all the IP addresses in the LAN organization and afterwards put in an exhibit object to take out commotion utilizing the preprocessing approach. A source associates with the intranet from their nearby PC, which has an IP address of 192.168.0.5, and picks the recipient IP address 172.16.0.6 in the organization; this collector IP address and port number are camouflaged, and the shipper conveys scrambled information to the supported IP address 172.16.0.6.

3.2 *End-of-Receiver-Phase*

The source information is gotten at the objective end during this stage. The IP address of the recipient is 192.168.0.6, which is utilized to check the cushion IP address. The objective end is portrayed in Fig. 5 contrasting the cushion IP address with the recipient end's framework arrangement utilizing the hid IP address and the objective end's port number. If the collector IP and support IP match, the confirmation is checked by giving the organization's indistinguishable mystery key. To get the information, the particular collector should supply the IP address and port number to affirm whether the objective beneficiary is a verified person to get the information.

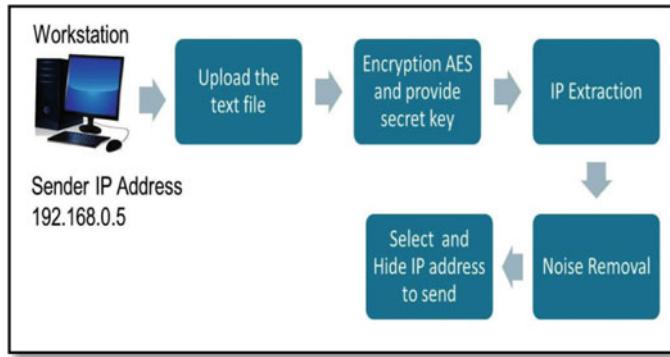


Fig. 4 End-of-sender-phase

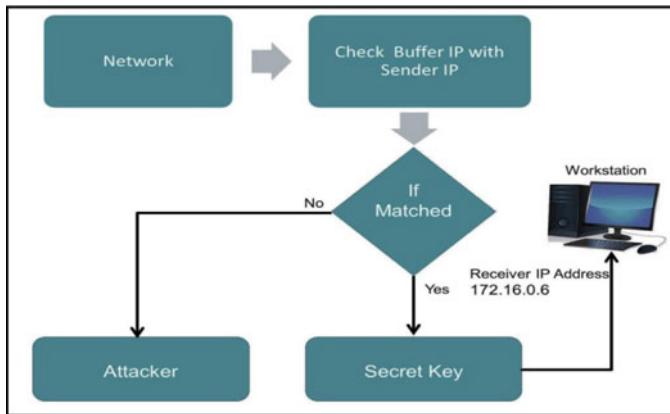


Fig. 5 End-of-receiver-phases

Algorithm 1—Sender (Signing)

- Step 1: Uploading text file \rightarrow t_{file}
 Step 2: call *encryption function* \rightarrow AES(t_{file})
Secret key Generated using PKI function
 Step 3: *Ip extraction*
 Step 4: *Noice removal*
 Step 5: *Hide IP address and send to receiver*
-

Algorithm 2—Receiver (Verification)

Step 1: call *Decryption function* → AES(tfile)
 Step 2: Checking *Ip address with server*
if (ip address matched)
Secret key extracted using PKI function
Authitication process
Allow user to access data
then
msg → *attaker found*
end if

Advanced Encryption Standard (AES) Function
AES(tfile)

Inputs:
 File in of 4MegaBytes // input plain text
 File out of 4MegaBytes // output plain text
 File w of 4*(Nr+1)* Megabytes // expanded key

Pseudo code:

```

login()           // login by the cloud user
{
  up\oad()
  {
    if(size>2 && size <= 4)
    {
      Select_128bit;      // performs 128bit key
    }
    elseif(size == 2)
    {
      Select_192bit;      // performs 192bit key
    }
    elseif(size < 2 && size > 0)
    {
      Select_256bit ;     //performs 256bit key
    }
  }
  Store ();        // store the data in the cloud
  download()
  {
    Select_download;   //downloads the file from cloud
  }
}
  
```

Moreover, a two-way correspondence was laid out to evaluate our strategies. Situation I: During the aloof phase of laying out an association, the p0f was used to assess the SYN parcel. During the dynamic stage, it found Windows 7 as the connecting machine's working framework, while Windows 7 was distinguished as the remote working framework. The departure association followed a comparable example, with Nmap observing a Linux OS with the 2.6.32 portion adaptation as the examined source address and p0f identifying a Linux OS with the 2.6 part as the associated host. Situation 2: The Spoofed IP was sent off. For the DDoS assault, a traffic generator was utilized, as well as a legitimate association for the two sources were employed. Utilizing the showing up bundle's source address, the genuine parcel's source during the functioning mode is affirmed. During the confirmation stage, the

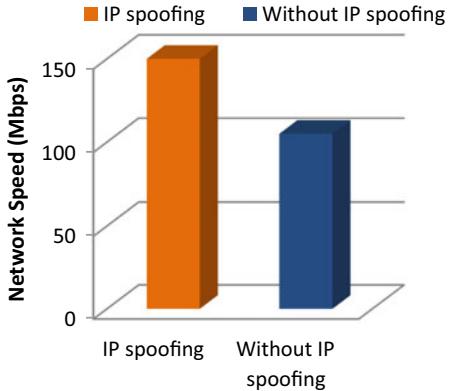
legitimacy of the association was affirmed by connecting the OS during dynamic and aloof fingerprinting to recognize the far off OS as Windows 7. During the dynamic and inactive stages, a dubious association with a faked IP address and an OS crisscross was found.

4 Performance Analysis

The information transmission speed is analyzed with and without IP ridiculing in this review. The IP mocking goes at a quicker rate than information sent without it. Figure 6 shows a correlation diagram for data transfer capacity levels with and without IP spoofing. Scenario one the ip packet extraction was used to analyze the SYN packet at the time of connection establishment during the passive stage. Windows 7 was recognized as the operating system (OS) running on the bridging machine, and during the active stage, Windows 7 was also identified as the remote OS. The egress connection experienced a similar process, with Nmap identifying a Linux OS with the 2.6.32 kernel version as the probed source address.

The Spoofed IP was released in scenario two. This was accomplished by using a valid connection for both sources and a traffic generator for the DDoS attack. Using the source address of the incoming packets, we checked the source of the real packet during the working mode. During the verification procedure, the validity of the connection was confirmed by comparing the OS during active and passive fingerprinting to identify the distant OS as Windows 7.

Fig. 6 Correlation chart for data transfer capacity levels with and without IP spoofing



5 Conclusion

IP address spoofing is a typical practice in PC network security that includes making Internet Protocol (IP) parcels with a produced or fake source IP address, which can prompt huge assaults on cloud communities. This study gives an inside out take at parodying assaults as a feature of DDoS assaults, as well as the security methodology that is right now available in the distributed computing climate. The proposed IP spoofing procedure centers around shielding information while it is being sent starting with one PC framework then onto the next in cloud climate. To take a look at the approved client and uncover the document sets on the server, IP address mocking is utilized. The unapproved client will endeavor to get information utilizing the objective end's IP address, checking the support IP address with the recipient end's framework settings utilizing the secret IP address and the objective end's port number. When the beneficiary and cushion IP addresses are coordinated, the confirmation is checked by giving the specific mystery key in the cloud organization; in any case, the interloper IP address is shipped off at the source end as a blunder message. The transmission capacity level with IP spoofing has a better exhibition assessment, showing that it has predominant strength and secured information in a cloud network.

References

1. Dhanapal A, Nithyanandam P (2019) The slow HTTP DDoS attacks: detection, mitigation and prevention in the cloud environment. *Scalable Comput Pract Exp* 20:669–685. <https://doi.org/10.12694/scep.v20i4.1569>
2. Osanaiye O, Choo K-KR, Dlodlo M (2016) Distributed denial of service (DDoS) resilience in cloud: review and conceptual cloud DDoS mitigation framework. *J Netw Comput Appl* 67:147–165. ISSN 1084-8045
3. Veeraghavan P, Hanna D, Pardede E (2020) NAT++: an efficient micro-NAT architecture for solving IP-spoofing attacks in a corporate network. *Electronics* 9:1510
4. Masdari M, Jalali M (2016) A survey and taxonomy of DoS attacks in cloud computing: DoS attacks in cloud computing. *Secur Commun Netw* 9. <https://doi.org/10.1002/sec.1539>
5. Vlajic N, Chowdhury M, Litoiu M (2019) IP spoofing in and out of the public cloud: from policy to practice. *Computers* 8(4):81
6. Fonseca O et al (2021) Identifying networks vulnerable to IP spoofing. *IEEE Trans Netw Serv Manage* 18(3):3170–3183. <https://doi.org/10.1109/TNSM.2021.3061486>
7. Agoni AE, Dlodlo M (2018) IP spoofing detection for preventing DDoS attack in fog computing. In: 2018 global wireless summit (GWS), pp 43–46. <https://doi.org/10.1109/GWS.2018.8686626>
8. Dawle Y, Naik M, Vande S, Zarkar N (2017) Database security using intrusion detection system. *IJLERA* 02(03):01–06. ISSN: 2455-7137
9. ITPRO (2019) Public cloud used to power supercharged DDoS attacks, Sept 2018. Available online: <https://www.itpro.co.uk/public-cloud/31884/public-cloud-used-to-power-supercharged-ddos-attacks#gref>. Accessed on 28 Oct 2019
10. Chinnasamy V (2020) Understanding cloud DDoS attacks and cloud-based DDoS protection, Nov 2020. <https://www.indusface.com/blog/understanding-cloud-ddos-attacks-and-cloud-based-ddos-protection/>

11. Pohle T (2018) Public cloud services increasingly exploited to supercharge DDoS attacks: new link11 research, Sept 2018
12. Cloud Security Alliance (CSA) (2016) The treacherous 12: cloud computing top threats in 2016, Feb 2016. Available online: https://downloads.cloudsecurityalliance.org/assets/research/top-threats/Treacherous-12_Cloud-Computing_Top-Threats.pdf. Accessed on 28 Oct 2019
13. Lutkevich B, Rosenrance L (2021) IP spoofing, Oct 2021. <https://www.techtarget.com/searchsecurity/definition/IP-spoofing>
14. IP Address spoofing in DDoS attacks IP Spoofing, Mar 2020. <https://www.imperva.com/learn/ddos/ip-spoofing>
15. Hummel R, Hildebrand C (2021) Crossing the 10 million mark: DDoS attacks in 2020, Jan 2021. <https://www.netscout.com/blog/asert/crossing-10-million-mark-ddos-attacks-2020>
16. Cimpany C (2018) Operator of eight DDoS-for-hire services pleads guilty, Feb 2018. Available online: <https://www.zdnet.com/article/operator-of-eight-ddos-for-hire-services-pleads-guilty/>. Accessed on 28 Oct 2019
17. Somani G, Gaur MS, Sanghi D, Conti M, Buyya R (2017) DDoS attacks in cloud computing: Issues taxonomy, and future directions. *Comput Commun* J 107:30–48
18. Osanaiye OA (2015) Short paper: IP spoofing detection for preventing DDoS attack in cloud computing. In: Proceedings of the IEEE international conference on intelligence in next generation networks (IEEE ICIN), Paris, France, 17–19 Feb 2015
19. Deepali, Bhushan K (2017) DDoS attack mitigation and resource provisioning in cloud using fog computing. In: International conference on smart technologies for smart nation (SmartTechCon), Bangalore, India
20. Saranya RG, Kousalya A (2017) A comparative analysis of security algorithms using cryptographic techniques in cloud computing. *Int J Comput Sci Inf Technol* 8(2):306–310
21. Khan MT (2017) Review: network security mechanisms and cryptography. *IJCSMC* 6(7):138–146
22. Dongare AS, Alvi AS, Tarbani NM (2017) An efficient technique for image encryption and decryption for secured multimedia application. *IRJET* 04(04)
23. Koch R et al (2015) Behavior-based intrusion detection in encrypted environments. *IEEE Commun Mag Netw Serv Manag Ser* 11(4)
24. Somani G, Gaur M, Sanghi D, Conti M, Buyya R (2015) DDoS attacks in cloud computing: issues, taxonomy, and future directions. *Comput Commun* 107. <https://doi.org/10.1016/j.comcom.2017.03.010>
25. Bhardwaj A, Mangat V, Renu V, Halder S, Conti M (2021) Distributed denial of service attacks in cloud: State-of-the-art of scientific and commercial solutions. *Comput Sci Rev* 39:100332. <https://doi.org/10.1016/j.cosrev.2020.100332>
26. Shaw S, Choudhury P (2015) A new local area network attack through IP and MAC address spoofing. In: 2015 international conference on advances in computer engineering and applications, 2015, pp 347–350. <https://doi.org/10.1109/ICACEA.2015.7164728>
27. Kavisankar L, Chellappan C, Venkatesan S, Sivasankar P (2017) Efficient SYN spoofing detection and mitigation scheme for DDoS attack. In: 2017 second international conference on recent trends and challenges in computational models (ICRTCCM), 2017, pp 269–274. <https://doi.org/10.1109/ICRTCCM.2017.55>
28. Sultana S, Nasrin S, Lipi FK, Hossain MA, Sultana Z, Jannat F (2019) Detecting and preventing IP spoofing and local area network denial (LAND) attack for cloud computing with the modification of hop count filtering (HCF) mechanism. In: 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2), pp 1–6. <https://doi.org/10.1109/IC4ME247184.2019.9036507>
29. Hindy H et al (2018) A taxonomy and survey of intrusion detection system design techniques, network threats and datasets, vol 1(1). Association for Computing Machinery
30. Agrawal N, Tapaswi S (2017) A lightweight approach to detect the low/high rate IP spoofed cloud DDoS attacks. In: 2017 IEEE 7th international symposium on cloud and service computing (SC2), pp 118–123. <https://doi.org/10.1109/SC2.2017.25>

31. Bhale P, Biswas S, Nandi S (2018) An adaptive and lightweight solution to detect mixed rate IP spoofed DDoS attack in IoT ecosystem. In: 2018 15th IEEE India council international conference (INDICON), pp 1–6. <https://doi.org/10.1109/INDICON45594.2018.8987008>
32. Zibideh WY, Matalgah MM (2015) Modified data encryption standard encryption algorithm with improved error performance and enhanced security in wireless fading channels. *Secur Commun Netw* 8(4):565–573
33. Kaur M, Kaur N, Singh B (2017) Comparative study of different cryptoghraphic algorithms. *Int J Adv Res Comput Sci* 8(4)
34. Verma SVK (2016) A recent study of various encryption and decryption techniques. *Int Res J Adv Eng Sci* 1(3):92–94

Performance Enhancement of Magnetic Levitation System Using GWO-ABC Tuned High-Dimensional Robust Controller



Shirish Adam and Prashant Gaidhane

Abstract This paper describes a simplistic parameter tuning methodology using Grey Wolf Optimizer-Artificial Bee Colony (GWO-ABC) algorithm for large-variable Fractional order Fuzzy Proportional-Integral-Derivative (FO-Fuzzy-PID) controller applied to Magnetic Levitation System (MAGLEV). As MAGLEV is a complex nonlinear system, robust controller scheme and a substantial tuning methodology is needed. A systematic strategy to tune FO-Fuzzy-PID controller and shapes of fuzzy sets is proposed here. Optimal right structures of fuzzy sets are obtained using constrained optimization method. An optimization of different parameters is obtained through GWO-ABC algorithm. Performance of the FO-Fuzzy-PID controller is substantiated by comparison of results with conventional controllers. Suitable performance criteria is defined and robustness analysis and disturbance rejection are also studied.

Keywords MAGLEV · Fuzzy controller · GWO-ABC algorithm · Robustness analysis

1 Introduction

Magnetic Levitation (MAGLEV) technology is widely used in various real world application for its features like non-contact friction-less working and low maintenance [1]. Diverse applications, like, MAGLEV trains, mobile charging, magnetic levitation vehicles, satellite launching and precision Engineering, non-contact actuators, magnetic bearings, contact-less melting, etc., are seen around us [2]. This technology have scope in non-contact actuators and structures, precision engineering, launching of satellites, transportation techniques, etc. [3] and will be extensively used in future.

S. Adam · P. Gaidhane (✉)
Government College of Engineering, Jalgaon, India
e-mail: pjgaidhane@gmail.com

S. Adam
e-mail: shirish.adam@gcoej.ac.in

The complex systems are accompanied by different types of uncertainties and variable loads and disturbances. The controller designed to handle such systems must be robust to these deviations. For this, we have to minimize the objective functions based on difference performance criteria [4]. Generally, weighted sum of objective functions is assessed as a optimization criteria [5]. In literature, various nature inspired optimization techniques were used for tuning the controller parameters [6]. In recent years, researchers worked on strategies to optimize the structural shapes of the fuzzy sets to get the best of footprint of uncertainty (FOU) provided by them [7].

In this work, we are optimizing the shape of fuzzy sets along with controller parameters, hence, the optimization problem becomes high-dimensional constrained problem. With several parameters to tune, the optimization problem becomes complex and needs a substantial method to solve it. Hereby, we are proposing systematic strategy to tune 2-DOF-Fuzzy-PID controller along with optimizing the shapes of fuzzy sets.

Next, Sect. 2, introduces mathematical model of MAGLEV. Section 3 describes some methodologies used in this work, followed by, explanation of proposed controller and strategy in Sect. 4. Next, results and discussion are reported in Sect. 5 and conclusions from the simulations are drawn at last in Sect. 6.

2 Structure and Mathematical Modeling of MAGLEV

As depicted in Fig. 1, strong magnetic field is generated by electromagnets. This levitates the steel ball in air by balancing the gravitational force mg with applied magnetic force. In this system, the position of ball is detected to generate feedback signal and provided for stability of closed loop control. Usually, optical sensors are applied to recognize the ball position. Later, the equivalent modulated current is produced as a controller response.

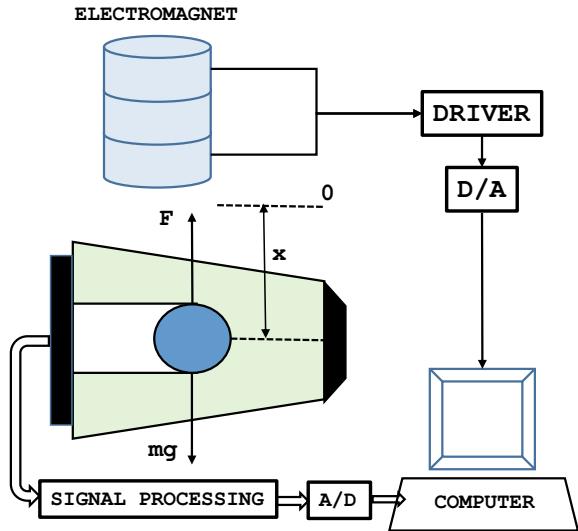
Mathematically, there is a balance between the gravitational force mg acting on the ball and applied magnetic force F . Hence, the mathematical model can be expressed as modeled in papers [4, 5].

$$m \frac{d^2x}{dt^2} = F(i, x) + mg \quad (1)$$

where, x is the optical sensor output defining the distance of steel ball with respect to magnetic reference, g is acceleration of gravity, m is mass of steel ball, and the current i is driving the electromagnet. Further, Electromagnetic force is generated by the applied control current, can be estimated as

$$F(i, x) = k \left(\frac{i}{x} \right)^2 \quad (2)$$

Fig. 1 Magnetic levitation system (MAGLEV)



where k is proportionality constant defined based on the physical parameters. As given earlier, there is nonlinear relationship between the current i and air gap x . Therefore, linearization of the system model at an equilibrium point (i_0, x_0) is required first. Thus, Eq. (2) is further modified using Taylor's expansion. The higher order terms are ignored in this evaluation. We obtained,

$$F(i, x) = F(i_0, x_0) + F_i(i_0, x_0)(i - i_0) + F_x(i_0, x_0)(x - x_0) \quad (3)$$

In which $F(i_0, x_0)$ represents the magnetic force required to balance the ball gravitational force, when the current is i_0 and air gap is x_0 .

$$F(i_0, x_0) = mg \quad (4)$$

$$K_i = F_i(i_0, x_0) = \frac{\delta F(i, x)}{\delta i} \bigg|_{(i_0, x=x_0)} = \frac{2K_{i_0}}{x_0^2} \quad (5)$$

$$K_x = F_x(i_0, x_0) = \frac{\delta F(i, x)}{\delta x} \bigg|_{(i_0, x=x_0)} = -\frac{2K_{i_0}^2}{x_0^3} \quad (6)$$

here, k_x is the coefficient of stiffness at the equilibrium point due to air gap, and k_i is the coefficient of the stiffness due to magnetic force to current, overall, whole system equation can be expressed as

$$F(i, x) = K_i i + K_x x + F(i_0, x_0) \quad (7)$$

$$m \frac{d^2 x}{dt^2} = K_i(i - i_0) + K_x(x - x_0) \quad (8)$$

From the schematic, with the static inductance (L), applying KVL across the loop, we can obtain the voltage equation as follows,

$$U(t) = Ri(t) + L\left(\frac{di}{dt}\right) \quad (9)$$

From [3], other MAGLEV system parameters are noted as (1) $g = 9.8 \text{ (m/s}^2)$, (2) $K_s = 458.7204$, (3) $K_a = 5.8929$, (4) $m = 0.22(\text{kg})$, (5) $r = 0.0125 \text{ (m)}$ is the radius of ball. Also, it is found at equilibrium point that $x_0 = 0.03(\text{m})$ and $i_0 = 0.6105 \text{ (A)}$. Henceforth, we obtained

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 980.0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 2499.1 \end{bmatrix} u_{\text{in}}$$

$$y = x_1$$

3 Optimizing Algorithm and Other Methodologies

This section briefly describes the methodologies used in the proposed work. The optimization algorithm, FLC controller, and fractional order approximation are discussed.

3.1 GWO-ABC Algorithm

Several swarm intelligent based optimization algorithms were proposed by researchers in last few decades [8]. Few of them are widely appreciated for their ability to get globally optimized solution in available high dimensional search space. Then also, the conventional versions of these algorithms have certain limitations, such as, poor exploration abilities and immature locking to local optima [9]. Henceforth, with the aim to improve their performance, modified or hybrid versions are proposed by inculcating good methodologies of different algorithms. Novel GWO-ABC is such a newly proposed algorithm which inherit the advantages of conventional GWO [8] and ABC [10]. It is evident from the comparative analysis that GWO-ABC performs much better and have faster convergence.

The GWO-ABC succeeds in 3 phases, known as, Population initialization phase, GWO phase, and ABC information sharing phase. The major features of the algorithm can be deciphered from [9]. By hybridizing major features of two algorithm, each solution element shares information with other element and overall global search

technique is ameliorated to elude inappropriate stagnation. In controller tuning problems, with multiple-dimensional search space, such extensive techniques provide exhaustive coverage and attain jump-offs from any locally sub-optimal regions to deduce optimal controller parameters.

3.2 FO Approximation

Fractional-order operators (μ and λ), suggested by Oustaloup's recursive approximation method [11] is also implemented in proposed FO-Fuzzy-PID controller. Mathematically, the operators involves a recursive distribution of zeros and poles, and given as

$$s^\lambda = K_f \prod_{k=-N}^{k=N} \frac{s + \omega_{zr}}{s + \omega_{pr}} \quad (10)$$

The order of the approximation is approximated to $2N + 1$. and λ is assigned as a order of fractional integro-differentiator. The other parameters poles (ω_{pr}), zeros (ω_{zr}), and gain of filter (K_f) are expressed as,

$$\omega_{pr} = \omega_b \left(\frac{\omega_h}{\omega_b} \right)^{\frac{K+N+\frac{1}{2}(1+\lambda)}{2N+1}}, \omega_{zr} = \omega_b \left(\frac{\omega_h}{\omega_b} \right)^{\frac{K+N+\frac{1}{2}(1-\lambda)}{2N+1}}, K_f = \omega_h^\lambda \quad (11)$$

where, $k \in [-N, N]$ and ω_b, ω_h denotes the frequency ranges, here, $\omega = (10^{-3}, 10^3)$ rad/s. The reason behind selection of the Oustaloup's recursive approximation method, among others, is that it can be realized in terms of high-order analog and digital filters [12]. For the fractional-order operator design used in this work the order of Oustaloup's approximation is 5 and $N = 2$.

3.3 Fuzzy Logic Controller

The FLC is assimilated to adopt the designer's knowledge in inference mechanism development [12–14]. As shown in Fig. 3, two inputs, error $e(t)$ and its fractional derivative $\Delta e(t)$, are applied to FLC. Generally, triangular fuzzy sets are preferred for their ease of implementation in hardware. As illustrated in 2a, linguistic labels 'Negative (N)', 'Zero (Z)', and 'Positive (P)' are used to define three triangular fuzzy sets. On the other hand, the output consequent fuzzy sets are represented in five crisp singletons, and labeled as: 'Negative Big (NB = -1)', 'Negative Medium (NM = -0.8)', 'Zero(Z = 0)', 'Positive Medium (PM = 0.8)', and 'Positive Big (PB = 1)', as depicted in Fig. 2b. A 3×3 rule base is used for the prime functioning of FLC inference mechanism. The rules are defined in accordance with nature of the

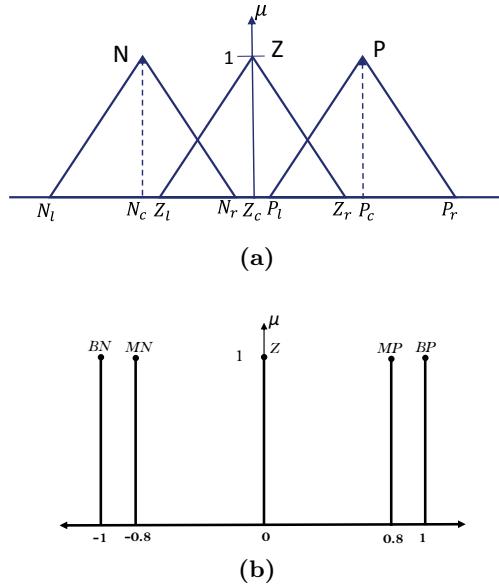


Fig. 2 a and b General structures of input and output fuzzy sets, respectively

response, process dynamics, and the knowledge of expertise. Their non-linear surface plot is depicted in Fig. 4a. MATLAB is used and simulations are performed in Fuzzy Logic Toolbox.

4 Proposed Controller and Optimization Strategy

4.1 FO-Fuzzy-PID Controller

Here, the fractional order parameters, which gives extra freedom to designer [12], are combined with fuzzy logic PID controller. Thus, the schematic of FO-Fuzzy-PID controller applied to MAGLEV for robust control is presented in Fig. 3. We can observe that two performance indices ISA and ITAE are used in the model. FLC is applied for inculcating human expertise in inference mechanism. For comparison other controllers like PID, FOPID and Fuzzy-PID are used. Their schematics can be obtained by some modifications in Fig. 3.

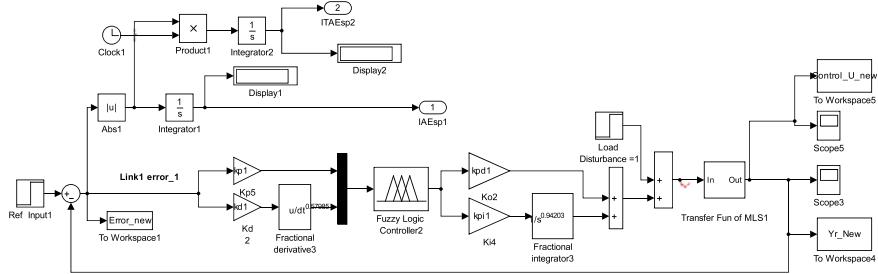


Fig. 3 FO-Fuzzy-PID controller applied to MAGLEV system

4.2 Problem Definition

The optimization is applied to tune the parameters with the aim to minimize the objective function. In this work, sum of two performance indices, (1) integral time absolute error (*ITAE*) and (2) integral absolute error (*IAE*), is opted as the objective function, as shown in Fig. 3.

$$J_1 = \text{ITAE} = \int t|e(t)|dt, \quad \text{and} \\ J_2 = \text{IAE} = \int |e(t)|dt \quad (12)$$

$$\text{Objective function : } J_0 = w_1 \times J_1 + w_2 \times J_2 \quad (13)$$

here, $e(t)$ is the errors between desired and actual output.

4.3 Optimization Strategy

The optimization strategy of FO-Fuzzy-PID controller is presented in this section. As depicted in Fig. 3, six controller parameters, labeled as, k_{p1} , k_{d1} , λ , μ , k_{pd1} , and k_{pi1} are tuned to generate desired output. Along with this, the shapes of fuzzy sets of the fuzzy controller are tuned. The three fuzzy sets for each input to fuzzy are labeled as shown in Fig. 2c. Nine parameters N_L , N_C , N_R , Z_L , Z_C , Z_R , P_L , P_C , P_R are also tuned with considering different constraints. In the design strategy of fuzzy sets optimization, constraints, given in Fig. 4b, have to be fulfilled by FLC while optimizing the shape of the antecedent fuzzy sets. Overall, the optimization algorithm is executed to optimize 15 parameters and final results are presented in next section.

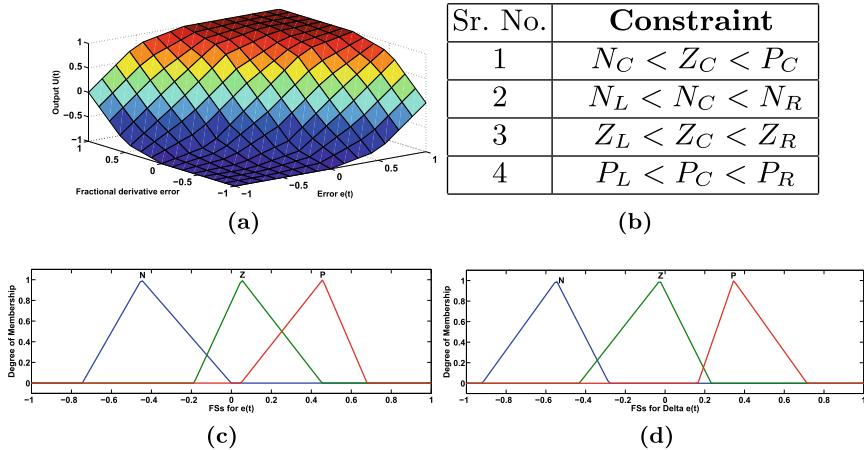


Fig. 4 **a** Rule surface chart of FLC, **b** constraints considered for fuzzy sets optimization, **c** and **d** optimized antecedent fuzzy sets for $e(t)$ and $\Delta e(t)$, respectively

5 Results and Discussion

To substantiate the performance of proposed tuned controller and effectiveness of algorithm various results are discussed in this section. The PID, FOPID, Fuzzy PID, and FO-Fuzzy-PID controllers are tuned using GWO-ABC algorithm for given performance criteria J_0 for 100 iterations and 100 element size for 20 runs. The min–max values for all variables are kept equal for all controllers, as shown in Table 2, and, the simulations are executed in MATLAB version R2014b. The Best-of-the-run values are noted here and responses are plotted for them.

5.1 Step Response

The optimized values of k_{p1} , k_{d1} , λ , μ , k_{pd1} , and k_{pi1} , for respective controllers, are recorded in Table 2. Along with this, 9 parameters of the optimized fuzzy sets of the fuzzy controller can be seen in Fig. 4c and d. The convergence speed is very significant in optimization algorithms, therefore, the convergence curves of GWO-ABC is compared with that of GWO and ABC algorithms in Fig. 5d, to prove its efficacy. We can observe that hybrid GWO-ABC is much faster in terms of convergence to optimal values. As discussed in Sect. 2, the step input of $r = 0.2$ amplitude is considered for all simulations and the step response is demonstrated in Fig. 5a.

The sum of two performance indices IAE and ITAE are used for optimization and results are noted in Table 1. The optimum values of proposed controller are depicted in Table 1 in bold fonts. The step response is depicted in Fig. 5a, the zoomed plot

Table 1 IAE, ITAE, and objective function values of controllers

Index	PID	FOPID	Fuzzy-PID	FO-Fuzzy-PID
IAE	1.21×10^{-1}	3.32×10^{-2}	5.14×10^{-2}	1.37×10^{-4}
ITAE	1.51×10^{-1}	5.11×10^{-2}	4.31×10^{-2}	1.08×10^{-4}
Objective function J_0	2.72×10^{-1}	8.43×10^{-2}	9.45×10^{-2}	2.45×10^{-4}

Table 2 Values of optimized controller parameters

Parameters	PID	FOPID	Fuzzy-PID	FO-Fuzzy-PID	Min	Max
k_{p1}	95.4571	15.1240	17.4251	19.7748	0	100
k_{d1}	89.4618	96.3589	86.3251	59.8694	0	100
k_{pd1}	0.4568	0.4527	0.6521	0.4571	0	1
k_{pi1}	0.1278	0.8911	0.4578	0.8844	0	1
μ	–	0.2304	1	0.3522	0	1
λ	–	0.7796	1	0.7581	0	1

shows that the proposed controller have less oscillatory response and settle fast to steady state.

5.2 Robustness Analysis

Robustness of controller is checked for external disturbance rejection and variable input response. A step signal of amplitude $d = 0.1$ is applied at $t = 0$ s to all the control and the disturbance behaviour is plotted in Fig. 5b, which shows that proposed controller settle efficiently than other.

Further, to evident the behaviour of the controller output for tracking the variable input, the response of system is presented in Fig. 5c. The expanded plot shows lower oscillatory behavior and faster settling than other controllers. Overall, we can also comment that the optimized fuzzy sets of FO-Fuzzy-PID controller enhance the execution of FLC and work better for added uncertainties.

6 Conclusion

This paper presents optimization of high dimensional robust FO-Fuzzy-PID controller applied to Magnetic Levitation System. A simple design scheme and parameter tuning approach using Grey Wolf Optimizer-Artificial Bee Colony (GWO-ABC)

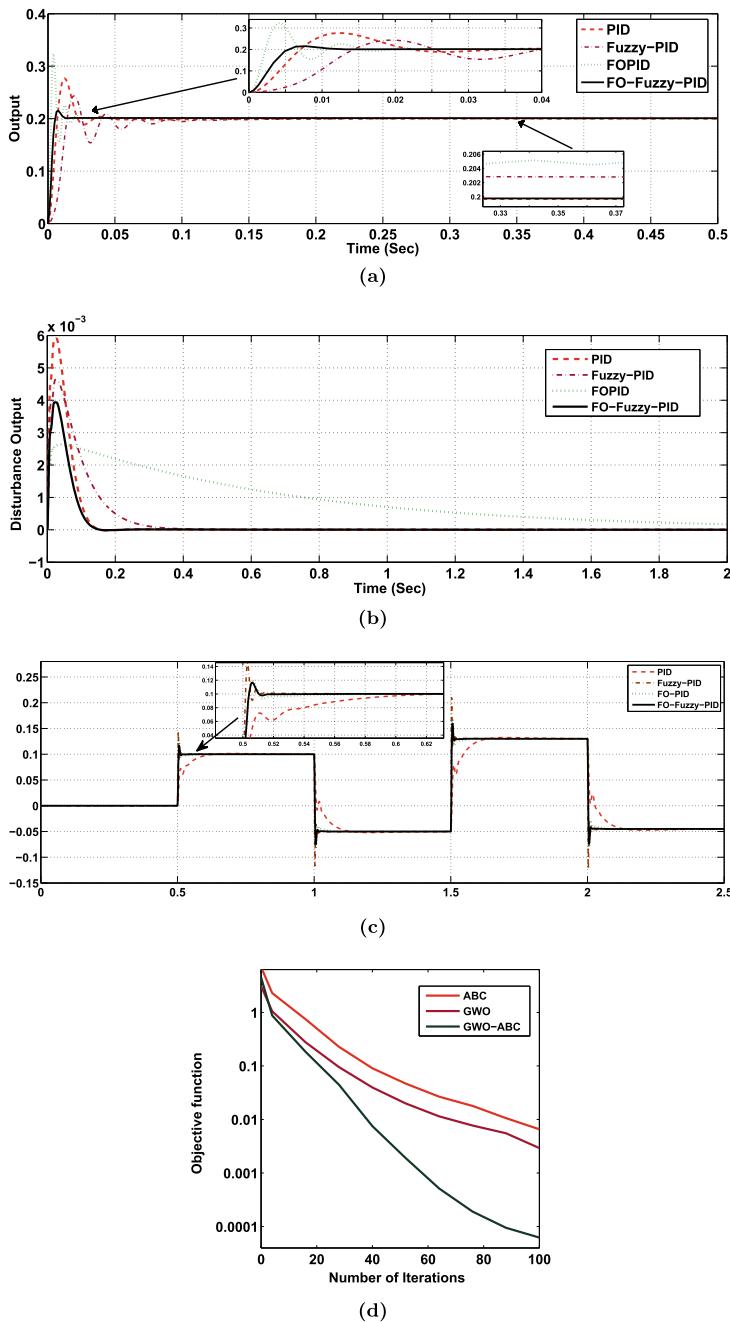


Fig. 5 **a** Step response and **b** disturbance rejection response, **c** robustness analysis for variable input, **d** convergence curve for GWO, ABC, and GWO-ABC algorithms

algorithm is presented and detail results and discussion is reported. It is observed that the incorporation of FO operators with advanced FLC, with conventional PID controller, enhanced the overall robustness. Extra flexibility is provided to get required path tracking with rejection of any external disturbances. After defining suitable performance criterion, IAE and ITAE, the performance of the FO-Fuzzy-PID controller is substantiated by comparison of results with conventional controllers. All graphs and values of IAE and ITAE prove that the FO-Fuzzy-PID controllers outperforms other controller and provide robustness and better disturbance rejection. Further, the tuned shapes of FLC fuzzy sets enhance the overall inference by capturing best of footprint of uncertainty. The ability of ABC-GWO for better convergence, without sticking to local sub-optimal result, is also manifested. Overall, it can be concluded from the results that the proposed FO-Fuzzy-PID controller is efficient counterpart when elimination of steady state error is required. Further experimentation of this proposed controller over other complex nonlinear systems can manifest its efficiency and applicability.

References

1. Ghosh A, Rakesh Krishnan T, Tejaswy P, Mandal A, Pradhan JK, Ranasingh S (2014) Design and implementation of a 2-DOF PID compensation for magnetic levitation systems. *ISA Trans* 53(4):1216–1222
2. Lin CM, Lin MH, Chen CW (2011) SoPC-based adaptive PID control system design for magnetic levitation system. *IEEE Syst J* 5(2):278–87
3. Google Technology Ltd. (2007) GML series magnetic levitation system, user manual and experimental manual
4. Kumar A, Panda MK, Kundu S, Kumar V (2012) Designing of an interval type-2 fuzzy logic controller for magnetic levitation system with reduced rule base. *Comput Commun Netw Technol* 1–8
5. Kumar A, Kumar V (2015) Performance analysis of interval type-2 Fuzzy controller applied to a magnetic levitation system. In: 2015 International conference on soft computing techniques and implementations, pp 107–112
6. Gaidhane PJ, Nigam MJ (2017) Multi-objective robust design and performance analysis of two-DOF-FOPID controller for magnetic levitation system. In: The proceedings of 14th India council international conference (INDICON), Dec 2017, pp 1–6
7. Gaidhane PJ, Nigam MJ, Kumar A, Pradhan PM (2018) Design of interval type-2 fuzzy pre-compensated PID controller applied to two-DOF robotic manipulator with variable payload. *ISA Trans*. Elsevier
8. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey Wolf optimizer. *Adv Eng Softw* 69:46–61
9. Gaidhane PJ, Nigam MJ (2018) A hybrid grey wolf optimizer and artificial bee colony algorithm for enhancing the performance of complex systems. *J Comput Sci* 27:284–302. <https://doi.org/10.1016/j.jocs.2018.06.008>
10. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J Glob Optim* 39(3):459–471
11. Oustaloup A, Levron F, Mathieu B, Nanot FM (2000) Frequency-band complex noninteger differentiator: characterization and synthesis. *IEEE Trans Circuits Syst I Fundam Theory Appl* 47:25–39
12. Kumar A, Gaidhane PJ, Kumar V (2017) A nonlinear fractional order PID controller applied to redundant robot manipulator. In: Proceedings of computer application in electrical engineering—recent advances (CERA), Oct 2017, pp 527–532

13. Kumar A, Kumar V, Gaidhane PJ (2018) Optimal design of fuzzy fractional order $PI^\lambda D^\mu$ controller for redundant robot. *Procedia Comput Sci* 125:442–448
14. Gaidhane PJ, Kumar A, Nigam M (2017) Tuning of two-DOF-FOPID controller for magnetic levitation system: a multi-objective optimization approach. In: 6th IEEE international conference on computer applications in electrical engineering—recent advances, pp 497–502

Approximation of Function Belonging to $\text{Lip}(\xi(t), p)$ Class by Using Borel's Mean



Smita Sonker and Rozy Jindal

Abstract Various researchers worked on trigonometric and Fourier approximation which is having great practical significance. By working on different classes, many researchers provided very useful results. As Chandra (J Math Anal Appl 275:13–26, 2002) [1] approximated a function $f \in \text{Lip}(\alpha, p)$ ($0 < \alpha \leq 1, p \geq 1$) by using trigonometric polynomials. Further, his result was improved by Leindler (Math Anal Appl 302:129–136, 2005) [5] by weakening the condition of monotonicity. In this paper, we use Borel exponential means for a function $h \in \text{Lip}(\xi(t), p)$ for $p \geq 1$ under L_p -norm. The established result generalizes and improves various existing results.

Keywords Borel mean · Holder's inequality · L_p -norm · Euler mean

Subject Classification MSC 2010 41A25 · 42B05 · 42B08

1 Introduction

Let $h \in L_p[0, 2\pi]$, ($p \geq 1$) be 2π -periodic function. The partial sum of the Fourier series of h is

$$S_n(h; x) = a_0/2 + \sum_{z=1}^n (a_k \cos zx + b_k \sin zx).$$

Let

$$\omega_p(\delta; h) = \sup_{0 < |t| \leq \delta} \left\{ \frac{1}{2\pi} \int_0^{2\pi} |h(x+t) - h(x)|^p dx \right\}^{1/p}; \quad t > 0,$$

S. Sonker (✉) · R. Jindal

Department of Mathematics, National Institute of Technology Kurukshetra, Thanesar, Haryana 136119, India

e-mail: smitafma@nitkkr.ac.in

be the integral modulus of continuity of $g \in L_p$. We say $h \in \text{Lip}(\alpha, p)$ ($\alpha > 0$, $p \geq 1$), whenever $h \in L_p$, if

$$\omega_p(\delta; h) = \mathcal{O}(\delta^\alpha); \quad 0 < \alpha \leq 1.$$

Definition 1 For h , the L_∞ -norm is given by

$$\|h\|_\infty = \sup\{|h(x)| : x \in \mathbb{R}\},$$

and L_p -norm is given by

$$\|h\|_p = \left\{ \frac{1}{2\pi} \int_0^{2\pi} |h(x)|^p dx \right\}^{1/p}, \quad p \geq 1.$$

Definition 2 For $h \in \text{Lip} \alpha$, $\alpha \in [0, 1]$ if

$$|h(x+t) - h(x)| = \mathcal{O}(|t|^\alpha), \quad t > 0,$$

and $h \in \text{Lip}(\alpha, p)$ if

$$\left\{ \int_0^{2\pi} |h(x+t) - h(x)|^p dx \right\}^{1/p} = O(|t|^\alpha).$$

For +ive increasing function $\xi(t)$, $h \in \text{Lip}(\xi(t), p)$ if

$$\left\{ \int_0^{2\pi} |h(x+t) - h(x)|^p dx \right\}^{1/p} = O(\xi(t)).$$

Notations:

$$\Psi_x(t) \equiv \frac{1}{2} \{h(x+t) - 2f(x) + h(x-t)\},$$

and

$$J(\Psi, t, r) \equiv \|\Psi_x(t) - \Psi_x(t+g_r)\|_p,$$

where $g_r = \pi/r$. Let \mathcal{A}_n^β ($\beta \geq 0$) denotes the binomial co-efficients in the expansion of $(1-z)^{-\beta-1}$ for $|z| < 1$:

$$(1-z)^{-\beta-1} = \sum_{n=0}^{\infty} \mathcal{A}_n^\beta z^n.$$

Initially to originate the approximation theory, Weierstrass theorem was used and further this study was carried out using trigonometric polynomials. The method of trigonometric approximation of signals was developed by Zygmund and Stegun [12] for the periodic series. Though it is very helpful in solving many problems related to PDE and different branches of engineering, many work on this research area has been done. Some helpful results in the respective direction is mentioned here. Mittal et al. [6] studied conjugate series using almost $\text{Lip}(\xi(t), p)$ class. Sonker [10] worked on function's approximation. Verma and Saxena [11] gave result on double product summability. Sonker and Jindal [7] and Sonker et al. [8] worked on triple product summability means and absolute matrix summability. Recently, Sonker [9] worked on triple product summability. Also, many more extended and new results can be drive in this direction. Therefore, in the present paper, we study approximation of h , $h \in L_p(\xi(t), p)$ by Borel's exponential transform given by (see p. 182 [3])

$$B_r(h; x) = \exp(-r) \sum_{n=0}^{\infty} (r^n / n!) s_n(h; x) \quad (r > 1).$$

2 Known Results

In 1981, Holland, Mohapatra, and Sahney (see p. 373, [4]), among other results, obtained some results as:

Theorem 1 *Let $h \in \text{Lip}(\alpha, p)$ for $p > 1$ and $\alpha \in (0, 1]$. Then,*

$$\|h - B_r(h)\|_p = \mathcal{O}(r^{-\alpha\beta}) \quad \left(0 < \beta < \frac{1}{2}\right). \quad (1)$$

Theorem 2 *Let $h \in \text{Lip}(\alpha, p)$ where $\alpha \in (0, 1)$ and $p > 1$ and let*

$$\int_{g_r}^{g_r^\beta} t^{-1} J(\Psi, t, r) \exp\left(-\frac{1}{2}rt^2\right) dt = \mathcal{O}(r^{-\alpha}), \quad (2)$$

where $\alpha/(1 + \alpha) \leq \beta < \frac{1}{2}$. Then,

$$\|h - B_r(h)\|_p = \mathcal{O}(r^{-\alpha}). \quad (3)$$

The order estimate $\mathcal{O}(r^{-\alpha})$ in above equation is called Jackson order.

In 1987, Chandra [2] obtained some results for the space L_p ($p \geq 1$).

Theorem 3 *If $h \in \text{Lip}(\alpha, p)$, where $p > 1$ and $\alpha \in (0, 1]$. Then,*

$$\|h - B_r(h)\|_p = \mathcal{O}(r^{-\alpha} \log r). \quad (4)$$

Theorem 4 *Let $h \in \text{Lip}(\alpha, p)$, $\alpha \in (0, 1)$ and $p > 1$ and let*

$$\int_{g_r}^{h_r} t^{-1} J(\Psi, t, r) \exp\left(-\frac{1}{3} r t^2\right) dt = \mathcal{O}(r^{-\alpha}), \quad (5)$$

where $h_r = r^{-\frac{1}{2}} \log r$. Then,

$$\|h - B_r(h)\|_p = \mathcal{O}(r^{-\alpha}). \quad (6)$$

Here, we prove the approximation of $h \in \text{Lip}(\xi(t), p)$ using Borel's mean. The result can be stated as:

3 Main Result

Theorem 5 *If $h \in \text{Lip}(\xi(t), p)$. Then,*

$$\|h - B_r(h; x)\|_p = \mathcal{O}\left[(l+1)^{\frac{1}{p}} \xi \left\{ \frac{1}{l+1} \right\}\right] \quad (7)$$

provided $\xi(t)$ satisfies

$$\left(\int_0^{\frac{1}{l+1}} \left(\frac{t|\Psi_x(t)|}{\xi(t)} \right)^p dt \right)^{1/p} = \mathcal{O}\left(\frac{1}{l+1}\right) \quad (8)$$

and

$$\left(\int_{\frac{1}{l+1}}^{\pi} \left(\frac{t^{-\delta}|\Psi_x(t)|}{\xi(t)} \right)^p dt \right)^{1/p} = \mathcal{O}\left((l+1)^\delta\right) \quad (9)$$

where $q(1-\delta)-1 > 0$, $\frac{1}{p} + \frac{1}{q} = 1$, $1 \leq p < \infty$.

Proof Now, let us consider the Borel mean $B_r(h; x)$ of $h \in \text{Lip}(\xi(t), p)$ class. If at a fixed point x

$$\|g - s_l(h; x)\|_p = \frac{1}{2\pi} \int_0^\pi \Psi_x(t) \frac{\sin(l + \frac{1}{2})t}{\sin \frac{t}{2}}.$$

The Borel mean transform $s_n(h; x)$ is:

$$\begin{aligned} \|g - B_r(h; x)\|_p &= \frac{e^{-r}}{2\pi} \int_0^\pi \Psi_x(t) \sum_{k=0}^{\infty} \frac{r^k}{k!} \sin\left(k + \frac{1}{2}\right) t \, dt \\ &= \int_0^\pi |\Psi_x(t)| |K_r(t)| \, dt \end{aligned} \quad (10)$$

where

$$\begin{aligned} K_r(t) &:= e^{-r} \sum_{k=0}^{\infty} \frac{r^k}{k!} D_k(t) \\ &= e^{-r} \sum_{k=0}^{\infty} \frac{r^k}{k!} \frac{\sin\left(k + \frac{1}{2}\right) t}{2 \sin \frac{t}{2}} \\ &= e^{-2r \sin^2 \frac{t}{2}} \frac{r^k}{k!} \frac{\sin\left(r \sin t + \frac{1}{2}t\right)}{2 \sin \frac{t}{2}}. \end{aligned}$$

It is simple to see, for every $r > 0$

$$|K_r(t)| \leq r + \frac{1}{2}; \quad (-\infty < t < \infty),$$

$$|K_r(t)| \leq \frac{\pi}{2\delta} e^{-2r(\delta/\pi)^2}; \quad (0 < \delta \leq t \leq \pi),$$

and

$$\left| \int_x^\delta K_r(t) \, dt \right| \leq \frac{2\pi}{r x}; \quad (0 < x \leq \delta \leq \pi).$$

Now, Eq. (10) can be written as

$$\begin{aligned} \|h - B_r(h; x)\|_p &= \left[\int_0^{1/(l+1)} + \int_{1/(l+1)}^\pi \right] |\Phi_x(t)| |K_r(t)| \, dt \\ &= I_1 + I_2 \text{ (say).} \end{aligned}$$

Now,

$$I_1 = \int_0^{1/(l+1)} |\Psi_x(t)| |K_r(t)| dt$$

equation (8) and $\Psi_x(t) \in \text{Lip}(\xi(t), p)$, we have

$$\begin{aligned} &\leq \left[\int_0^{1/(l+1)} \left(\frac{t|\Psi_x(t)|}{\xi(t)} \right)^p dt \right]^{1/p} + \left[\int_0^{1/(l+1)} \left(\frac{\xi(t)|K_r(t)|}{t} \right)^q dt \right]^{1/q} \\ &= \mathcal{O}\left(\frac{1}{l+1}\right) \left[\int_0^{1/(l+1)} \left(\frac{\xi(t)|K_r(t)|}{t} \right)^q dt \right]^{1/q} \\ &= \mathcal{O}\left(\frac{1}{l+1}\right) \left[\int_0^{1/(l+1)} \left(\frac{\xi(t)}{t^2} \right)^q dt \right]^{1/q} \end{aligned}$$

since, $\xi(t)$ is a +ively increasing, we get

$$\begin{aligned} &= \mathcal{O}\left(\frac{1}{l+1}\xi\left(\frac{1}{l+1}\right)\right) \left[\int_{\epsilon}^{1/(l+1)} \frac{dt}{t^{2q}} \right]^{1/q} \text{ for } 0 \leq \epsilon < \frac{1}{l+1} \\ &= \mathcal{O}\left(\frac{1}{l+1}\xi\left(\frac{1}{l+1}\right)\right) \left[\left\{ \frac{t^{-2q+1}}{-2q+1} \right\}_{\epsilon}^{1/(l+1)} \right]^{1/q} \\ &= \mathcal{O}\left(\frac{1}{l+1}\xi\left(\frac{1}{l+1}\right)\right) \left[(l+1)^{2-\frac{1}{q}} \right] \\ &= \mathcal{O}\left((l+1)^{1/p}\xi\left(\frac{1}{l+1}\right)\right) \left(\because \frac{1}{p} + \frac{1}{q} = 1 \right) \end{aligned} \tag{11}$$

and

$$I_2 = \int_{1/(l+1)}^{\pi} |\Psi_x(t)| |K_r(t)| dt$$

using equation (9), we have

$$\begin{aligned} &\leq \left[\int_{1/(l+1)}^{\pi} \left(\frac{|\Psi_x(t)|}{\xi(t)t^{\delta}} \right)^p dt \right]^{1/p} + \left[\int_{1/(l+1)}^{\pi} \left(\xi(t)t^{\delta}|K_r(t)| \right)^q dt \right]^{1/q} \\ &= \mathcal{O}(l+1)^{\delta} \left[\int_{1/(l+1)}^{\pi} \left(\xi(t)t^{\delta-1} \right)^q dt \right]^{1/q}. \end{aligned}$$

Putt = $1/z$,

$$\begin{aligned}
&= \mathcal{O}(l+1)^\delta \left[\int_{1/\pi}^{l+1} \left(\left(\xi \left(\frac{1}{z} \right) \right) z^{-\delta+1} \right)^q \frac{dz}{z^2} \right]^{1/q} \\
&= \mathcal{O} \left[(l+1)^{1-\frac{1}{q}} \xi \left(\frac{1}{l+1} \right) \right] \\
&= \mathcal{O} \left[(l+1)^{\frac{1}{p}} \xi \left(\frac{1}{l+1} \right) \right]. \tag{12}
\end{aligned}$$

Collecting (11) and (12), our result has been proved.

4 Applications and Conclusion

The application part of the approximation field is very vast. Here, we are approximating functions using Fourier series, which in turns having many applications in engineering and real life. Fourier series is used in the solution of partial differential equations. The equations are used in electromagnetics and wave propagation and in many engineering problems. There are uses in signal processing and control systems. The Fourier series expresses any signal in terms of sinusoids (cosines and sines). This allows us to view signals in the “Frequency domain.” It is very useful in EE, especially in communications and having some applications which go beyond present technology.

This result focuses on approximation of signal which belongs to class $\text{Lip}(\xi(t), p)$, ($p \geq 1$) by using Borel’s mean. To summarize all, the purpose of introducing higher class is to minimize the error of approximation. As much as we reduce the error, the result becomes stronger. In doing this, the concept of product summability is very helpful. Since, by using double, triple, or higher product means, the error of approximation decreases. So, we can develop new results with the help of various kinds of product summability by using appropriate conditions.

Acknowledgements The authors offer their true thanks to the funding agency Science and Engineering Research Board through Project No.: EEQ/2018/000393.

References

1. Chandra P (2002) Trigonometric approximation of function in L_p -norm. *J Math Anal Appl* 275:13–26
2. Chandra P (1987) Functions belonging to $\text{Lip}\alpha$ and $\text{Lip}(\alpha, p)$ spaces and their approximation. *Soochow J Math* 13:9–22
3. Hardy GH (1949) Divergent series, Oxford
4. Holland ASB (1981) A survey of degree of approximation of continuous functions. *SIAM Rev* 23(3):344–379

5. Leindler L (2005) Trigonometric approximation in L_p -norm. *J Math Anal Appl* 302:129–136
6. Mittal ML, Rhoades BE, Mishra VN, Priti S, Mittal SS (2005) Approximations of functions belonging to $\text{Lip}(\xi(t), p)$, ($p \geq 1$) class by means of conjugate Fourier series using linear operators. *Indian J Math* 47:217–229
7. Sonker S, Jindal R (2022) Approximation of signals by the triple product summability means of the Fourier series. *Soft Comput Theor Appl* 425:169–179
8. Sonker S, Jena BB, Jindal R, Paikray SK (2022) A generalized theorem on double absolute factorable matrix summability. *Appl Math Inf Sci* 16(2):315–322
9. Sonker S (2021) On generalized matrix summability $\phi - |U, \gamma; \delta|$ of orthogonal series. *Soft Comput Theor Appl* 519–528 (2021)
10. Sonker S (2014) Approximation of functions by means of its Fourier-Laguerre series. *Proc ICMS* 1(1):125–128
11. Verma S, Saxena K (2017) A study on $(H, 1)(E, q)$ product summability of Fourier series and its conjugate series. *Math Theory Model* 7(5)
12. Zygmund A, Stegun IA (1959) (EDS): trigonometric series, 2nd edn. Cambridge University Press, Cambridge

Combining Genetic Algorithm and Support Vector Machine for Classification of Cancer on Microarray Data



Tanja Plagemann, Rolf Dornberger, and Thomas Hanne 

Abstract This paper discusses the classification of microarray data for breast cancer gene expressions using a Genetic Algorithm. The available CuMiDa dataset is investigated regarding its suitability for Machine Learning (ML) applications as well as presenting the benchmark scores of a collection of selected ML algorithms. The methodology and use of a Genetic Algorithm (GA) both as a classifier and for feature pre-selection is explored and compared with hybrid or fusion architectures. Finally, an ensemble setup of a GA with a Support Vector Machine (SVM) is implemented with a subset of features. It is compared to a simple SVM on the whole feature set with the result that it is able to match it in performance across all applied metrics, although just a relatively small number (10–20) from the total number of features (36,000) is used.

Keywords Cancer classification · Genetic algorithm · Microarray · Gene expressions · CuMiDa · *F*beta-score · Support vector machine · Ensemble method

1 Introduction

Cancer is a focus topic in medical research: The earlier cancer is detected the better treatment options can be applied. One possibility to analyze a cell sample is to create a deoxyribonucleic acid (DNA) microarray. Its purpose is to identify a large number of gene expressions from a single cell and to measure their level of activity. When compared with a non-cancerous sample this activity can provide insight for diagnosis and disease progression prediction [1]. Additionally, searching for gene expression patterns predictive of the cancer and their associated pathways might result in a new drug target.

As a basis for accurate and reliable predictions a large amount of scientifically validated data is needed. The Curated Microarray Database (CuMiDa) [2] provides a range of microarray datasets which are specifically built and curated for Machine

T. Plagemann · R. Dornberger · T. Hanne (✉)

University of Applied Sciences and Arts Northwestern Switzerland, Basel, Olten, Switzerland

e-mail: thomas.hanne@fhnw.ch

Learning (ML) applications, thus the computationally supported prediction of cancer. The data are collected from publications, preprocessed, and validated scientifically by checking the methods used for obtaining the results. The resulting datasets are balanced in terms of class distribution (tumor and normal) and are numerous enough for the use with ML algorithms. The main challenge for many algorithms arises from the high dimensionality of the data which can exceed several thousands in terms of features, i.e., gene expressions in case of microarray data.

2 Description of the Optimization Problem

2.1 Problem Model

Microarray data present a classification problem with comparably few data samples, but many features. The goal is to classify whether the data record describes a cancerous cell or a normal (i.e., non-cancerous) cell. The secondary task is to search for a pattern of gene expressions which is predictive for the disease.

For this work, a dataset regarding breast cancer has been selected from CuMiDa. It consists of 289 samples of which 146 are classified as normal and 143 as breast cancer of type adenocarcinoma (usually formed in milk-producing glands or milk ducts). For each sample up to 36,000 gene expressions are available to be used as features.

The success of a classification model can be measured in accuracy (1) i.e., correctly predicted class label or true positive (TP) and true negative (TN), but this is not sufficient in such a case. The calculated *F1*-score (4) combining precision (2) and recall (3) can give further insight about the performance of the algorithm. These measures take false positives (FP) and false negatives (FN) into account.

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{total sample size} \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$F1\text{-score} = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}) \quad (4)$$

It is important for the algorithm to be able to determine all cancerous samples, thereby avoiding false negatives, i.e., undetected cancer which could harm the patient in the long term. That means to maximize the recall (4) a score of 1.0 denotes no false negatives. However, false positives should also be minimized to avoid unnecessary burden on a patient. This is captured by precision (3). The two measures, recall and precision, conflict with each other. Maximizing one is often at the cost of the

other, depending on the threshold for classification. The $F1$ -score assumes equal importance for both precision and recall. Since this is not the case in the described model, the $F\beta$ -score (5) can be used instead.

$$F\beta\text{-score} = \frac{2 \times (1 + \beta^2) \times \text{precision} \times \text{recall}}{(\beta^2 \times \text{precision} + \text{recall})} \quad (5)$$

The $F1$ -score is a specific form of the $F\beta$ -score with the weight $\beta = 1$. Decreasing β to less than 1 puts more weight on precision, while increasing β favors recall. The latter is more beneficial in the cancer model. The $F2$ -score and the $F0.5$ -score are derived similarly with the β weights indicated in their name.

2.2 Optimization Methods

For the chosen dataset, benchmark results for some of the most common ML algorithms were obtained by using the program Weka. The available methods include the Support Vector Machine (SVM), Decision Tree (DT), K -nearest neighbor (KNN) and the baseline measurement ZeroR (always predicting the majority class). The accuracy ranges from 0.8 (DT) to 0.93 (SVM) [2]. However, as the parameters for the ML algorithms are not provided in detail, the reproduction of the same numerical experiment resulting in the same scores in Weka is not possible. Therefore, in our numerical experiments, executions of the mentioned ML algorithms with own preset parameters are used as a benchmark. Comparing literature and our results to the field experiments, the tested algorithms still misclassify about 10–20% of the samples. Thus, the motivation arises that further algorithms have to be considered to improve on the scores for this particular microarray dataset from CuMiDa.

Applying a Genetic Algorithm (GA) to microarray data, which achieves a comparable accuracy to other ML algorithms like SVM, has already been researched [3]. When using high-dimensional datasets (i.e., with a wide range of features), feature pre-selection can improve the performance of ML algorithms by focusing on a smaller number of features. The GA does not necessarily require pre-selection of features to perform well, which makes it particularly interesting for high-dimensional data.

The structure of the GA is as follows [3]:

- (1) Create an initial population (the final solution may vary significantly based on the selection of features for the first population). The population consists of chromosomes which are a representation of selected features. Features can optionally be encoded prior to this step.
- (2) Evaluate the fitness of each solution (i.e., a subset of features) with a fitness function. This can be for example the accuracy of a classification algorithm.
- (3) Check whether a termination condition (e.g., target accuracy or the number of generations to be executed) is reached.

- (4) If the termination condition is not fulfilled, then generate a new population by applying genetic operators.
 - (a) Reproduction: take over selected good solutions unchanged for the next iteration to avoid losing a possible best solution (known as elitism).
 - (b) Crossover: generate a new solution (offspring) by combining parts of two solutions (parents).
 - (c) Mutation: random features are changed of randomly selected solutions to maintain diversity.

Continue with step 2 until the condition in 3 is satisfied.

The GA will try to find a combination of features (i.e., genes) that are predictive of the disease. This makes the GA well-suited to answer the second research question about specific patterns of gene expression with predictive qualities.

The Support Vector Machine (SVM) is a widely used simple algorithm which can be applied to either regression or classification problems. It tries to separate the data points into distinct groups by a hyperplane so that the distance of all points toward the hyperplane is maximized. The dimension of the n -dimensional space is determined by the number of features. For example, in a two-dimensional space this could be equal to drawing a line between the points. For the SVM the parameter C can be adjusted for regularization. It is used to influence the weight placed on correctly classifying the points versus optimizing the margin toward the hyperplane. A low value of C accepts more misclassifications to achieve a better margin. Additionally, it is possible to add a kernel function which allows, for instance, to classify non-linear data in a higher dimension.

2.3 *Implementation and Comparison of Different Optimization Methods*

In the literature three main approaches are described to implement the GA: filter, wrapper, or a hybrid approach. In the filter approach, the GA relies on statistical properties of the dataset to determine relevant subsets of features. It assumes the features are independent of each other and does not use a classification algorithm. The wrapper approach consists of subset generation and evaluation, while for the latter a ML algorithm is applied. For the hybrid approach, a combination of the two is used [4].

Singh et al. [5] use the Spearman's Correlation Coefficient (SCC) for feature selection, then applying the GA with the accuracy of a neural network classifier as fitness function. The resulting feature subset is finally evaluated for classification accuracy with several ML algorithms. Alomari et al. [4] follow a similar approach, pre-filtering the data, feeding a GA which uses a SVM and two other algorithms for classification. The results are combined by a majority voting scheme. Ahmed et al. [6] propose feature pre-selection with KNN to improve the classification, executed with a fusion of multiple SVMs.

The first step in the classification process is the preparation of the data: validation of the values and the labels as well as transforming the dataset into a machine-readable format (e.g., a table or applying binary encoding for the class label [5]). To apply a dataset to an ML algorithm, balanced classes are beneficial. If one class is overrepresented, an algorithm might revert to always predicting the majority class. This would achieve high accuracy but is not useful to detect the minority cases. In case imbalance is unavoidable, the minority class could be oversampled (i.e., used multiple times) to add weight to these cases. The other approach is to undersample the majority class, though this reduces the size of the overall dataset. If the dataset is small to begin with, this approach is often not feasible.

For the chosen dataset from CuMiDa, the data preparation step has already been performed, and the classes are almost equally represented. Adding a feature pre-selection step at this point is optional, though often applied. The feature subset is then fed into a classification algorithm. Both steps can be taken over by the GA. If there were multiple feature sets, the prediction results could be combined into one using, e.g., the majority voting scheme [4]. With this method, the most frequently represented class label would be chosen as the final label. Instead of a majority vote, an additional ML classifier can be added, resulting in an ensemble setup.

For this work such an ensemble setup based on the GA and the SVM is chosen (Fig. 1) and compared with the performance of the baseline scores. The workflow denoted in Fig. 1 is as follows:

- Import data, separate target column from feature columns, perform training/validation/test split
- Define parameters for the Support Vector Machine (SVM) classifier and the Genetic Algorithm
- Run classifiers with accuracy and *F*1-score as baseline measures
- Define fitness function using classification accuracy of the SVM with cross validation as measure
- Run Genetic Algorithm (class instance) multiple times and output *k* best solutions (feature sets)

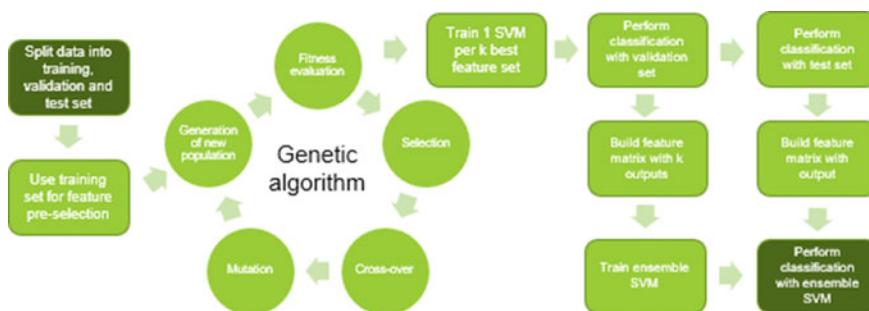


Fig. 1 The ensemble setup

- Train k SVM base classifiers with selected feature sets and classify on validation dataset
- Build feature matrix with all outputs and train ensemble SVM classifier
- Perform classification with base classifiers and then ensemble classifier with test set.

The experiment is implemented in Python (version 3.10.4) on Jupyter Notebook. It uses the following libraries: PyGAD (version 2.16.3) [7], pandas, numpy, Scikit-learn (version 0.18.1) [8], and matplotlib. The setup consists of:

- (1) Dataset preparation
- (2) Several runs of the GA
- (3) Training and validating intermediate SVMs
- (4) Training an ensemble SVM
- (5) Testing the ensemble SVM.

For the first step, the complete dataset from CuMiDa is loaded and the feature columns (i.e., gene expressions) are separated from the class label (i.e., normal or cancer). The data is then divided into three separate datasets for training, validation, and testing. The ensemble SVM needs sufficient samples based on the validation dataset for training. Thus, the usual size of 60% for the training set was reduced in favor of the validation set. This results in a training set consisting of 50% of the data, while 30% are allocated to the validation set, leaving 20% for the testing set. At this point, the 3 classifiers (SVM, KNN, and DT) are trained and tested with similar parameters as presumably used for Weka [2]. Table 1 presents the results for accuracy, recall, precision, and the $F1$ -score. The SVM provides—at least in 4 of 5 parameters—the best performance and will be used as a benchmark for the new setup.

Step two consists of several iterations of a standard GA algorithm in order to optimize the set of features to be used later on. There are many parameters (mutation rate, etc.) to steer the function of the GA. The most important part of the GA is the chosen fitness function: based on each solution (i.e., a selection of gene expressions) a SVM is trained using the training set. The classifier is evaluated on a metric, e.g., accuracy based on the predictions made with the validation set. The goal is to misclassify as few samples as possible, while keeping precision high. Thus, the $F\beta$ -score (5) has to be applied with $\beta > 1$. Several values are experimentally tested for β . The fitness determines which solutions are chosen to build the population for

Table 1 Benchmark results for SVM, KNN, and DT

	SVM	KNN	DT
Accuracy	0.9138	0.7931	0.8276
Recall	0.8621	0.6552	0.7586
Precision	0.8615	0.9048	0.88
$F1$ -score	0.9091	0.76	0.8148
$F2$ -score	0.9002	0.6934	0.7801

the next generation. Some test runs have shown that in most cases 100 generations are sufficient to reach a stable solution. For further generations the fitness values only rarely change. Each iteration consists of evaluating the fitness of all the candidate solutions in the population, choosing the best solutions, applying crossover and mutation operations, and generating the next population.

The mutation operation can be influenced with two settings. PyGAD allows a fixed mutation probability or two probabilities with the adaptive method [7]. The adaptive method means that it will calculate the average fitness of a population. Then it applies the high mutation rate to the solutions with a fitness less than the average fitness, while using the low mutation rate for the other (better) solutions.

The size of the population denotes the number of solutions called chromosomes. Each chromosome consists of a fixed number of genes, in this case the number of gene expressions or features. Since the secondary goal is to find predictive sets of gene expressions, the number of genes was set only to 15. The low number also avoids having most of the genes covered by the different solutions and iterations. The size of the population was kept similarly small at 10 as it influences the matrix size and the number of selected genes. This means the GA could represent a maximum of 150 gene expressions from 36,000 available features in the solutions of one generation. The GA class of PyGAD has a setting to choose whether duplicate genes are allowed within a solution and within the population. It was observed that depending on the other settings for the GA the algorithm might loop through all the available gene expressions and run out of options. Thus, for this experiment, the setting was maintained as true. This also allows the GA to put more weight on specific features by selecting them multiple times.

The number of parents is chosen based on the number of solutions in the population. If the setting is too close to the population size, the selection mechanism (elitism-like direct copying of a fixed number of solutions to the next generation) does not produce good results anymore, since almost every solution survives to the next generation. To avoid this problem and at the same time not to lose good solutions too quickly, the number of parents was set to half the population size—five in this case.

Reproducibility is generally helpful so that results of different settings can be compared. The GA, if not provided a preset initial population, generates the starting population randomly. Furthermore, due to the randomness of the selection, mutation, and crossover operators, each execution of the GA produces different results, making comparison futile. To make the GA process reproducible, the initial population for each run of the GA was selected randomly based on a seed. For this operation the random choice function of library numpy was applied. To have a different seed available for each run, the list of seeds itself was randomly generated with the same seed used for the training-validation-test split of the dataset.

To build the dataset for the ensemble SVM, the GA was executed for several runs. To avoid overfitting and still obtain a reasonable width for the matrix, the number of runs was set to 20. Each run of the GA with 100 generations produced one best solution. For each solution, the training set was filtered as per the selected feature columns. This filtered dataset was used to train a separate SVM and perform the

classification on the (equally filtered) validation set (step 3). The predictions of the SVMs were collected and assembled into a matrix (length: number of samples in validation set; width: number of SVMs trained based on GA runs). There are only two possible values for each data point in the matrix: either 1 for cancer or 0 for normal. This makes the next step a binary problem. With this matrix as input the ensemble SVM was trained (step 4).

To test the ensemble SVM, a matrix like the one used for training in the previous step is needed. The testing matrix consist of the predictions made by the 20 SVM classifiers of step 3 based on the previously unseen test data. Finally, the ensemble SVM tries to classify the test samples (step 5).

3 Results

The complexity of the setup and of the two classifiers results in a huge list of parameters. Within the limited scope of this work, a lot of the parameters have to be chosen with a few experimental test runs or based on standard settings covered by literature. The investigation of the impact of each parameter and of all the combinations on the evaluation measures has to be done in future research. The parameters used in this research are summarized in the following list:

- Runs (of the GA): 20
- Generations (in one run of the GA): 100
- Population size: 10
- Genes per solution: 15
- Parents: 5
- Duplicate genes: allowed
- Mutation type: adaptive (2 mutation rates)
- Mutation probabilities: 0.35, 0.15.

In the first prototype implementation of this optimization problem, the GA fitness function was using the accuracy metric to evaluate the fitness of the SVM trained on the solution. The focus was more on testing the overall setup versus creating an implementation fitting the data model. Each execution of the GA took close to one minute to finish (on a standard laptop). When plotting the fitness of the best solution of each generation and overlaying the 20 iterations, the improvement in fitness can be clearly observed (Fig. 2). As stated earlier and visualized in the plot, changes in fitness rarely occur after 60–70 generations.

The prototype produced surprisingly good results that were identical to the baseline scores with the above-mentioned settings. The results changed noticeably when the fitness function was adjusted to respect the increased weight on recall, thus trying to avoid false negatives. First, accuracy was replaced with recall. This resulted in overall decreased performance. Then, instead of accuracy, the metric *F*beta-score was tested. According to its definition, a beta larger than 1 will increase the weight of recall. Table 2 shows the results of the ensemble SVM algorithm for fitness measures

Fig. 2 Fitness of the solutions of 20 runs

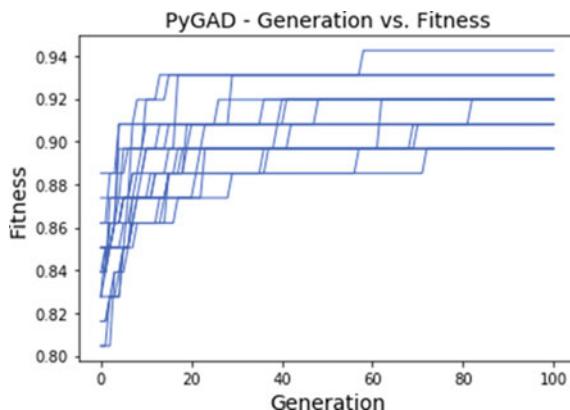


Table 2 Results of the ensemble algorithm with different fitness functions

	Baseline	Trained on accuracy	Trained on recall	Trained on <i>F</i> 1-score	Trained on <i>F</i> 2-score	Trained on <i>F</i> 0.5-score
Accuracy	0.9138	0.9138	0.8621	0.9138	0.8793	0.8621
Recall	0.8621	0.8621	0.7586	0.8621	0.7931	0.8966
Precision	0.9615	0.9615	0.9565	0.9615	0.9583	0.8387
<i>F</i> 1-score	0.9091	0.9091	0.8462	0.9091	0.8679	0.8667
<i>F</i> 2-score	0.8803	0.8803	0.7914	0.8803	0.8214	0.8844

with different beta values and compared to the baseline scores, recall, and accuracy of the previous executions (rounded to four decimals).

The ensemble SVM was able to match the baseline results with the GA fitness functions based on either accuracy or the *F*1-score (columns shown in bold). As soon as beta was adjusted to give more weight to recall, the performance decreased to about the same level as using only recall, though precision stayed at a very high level. In another run beta was decreased below 1, which would give precision precedence over recall. The *F*0.5-score showed similar low results as trained on recall. Interestingly, the recall value itself stayed on a higher level, though at the cost of precision. This also resulted in a higher *F*2-score. The two values are underlined in Table 2.

For each of the executions, the selected features (gene expressions) were collected. There were no significant findings concerning recurrently chosen gene expressions. Only a few features occurred more than once and could not be connected to higher predictive quality.

4 Conclusions

A GA-based classifier achieves accuracy comparable to SVM and KNN, though hybrid architectures for the GA often outperform such simple setups [4, 5]. The settings for the GA like population size or crossover and mutation rate are mostly determined by trial and error [5] which is also the case in the presented experiment. A positive trend in accuracy is observed in [6] for an SVM fusion setup when increasing the number of subsets. However, [4–6] have less than 150 samples and between 2000 and 12,000 features, which makes them smaller in size and width compared to the selected dataset from CuMiDa with 289 samples and up to 36,000 features used in this research.

The chosen SVM ensemble setup with the GA for feature selection has proven as effective as using only an SVM. The positive aspect is that it successfully uses only a relatively small subset of the 36,000 features for classification. Larger datasets, especially in terms of samples, and with other diseases should be tested to validate the setup further.

It is also noted that the SVM reacts unexpectedly when putting the focus of the fitness function on recall. Even though the avoidance of false positives was favored, the algorithm has difficulties in separating the data points into distinct groups. This may be due to the binary nature of the matrix that is the input for the ensemble SVM. Other algorithms should be considered that might perform better on binary data. This could result in a mixed ensemble setup where the fitness of the GA still uses the SVM (which performed well on the large dataset) and applies a different classifier on the binary matrix for the final classification step.

Due to the limited scope of the work only few parameters of the setup are considered. For future work, a grid search of the parameters or other approaches for parameter tuning could be applied to optimize the whole setup rather than only a fraction of it. Since there are many combinations of parameters, this process could be computationally expensive but would possibly allow for a better balancing of exploration and exploitation strengths of the considered GA [9].

Another area of improvement is the only briefly examined secondary goal of the optimization. Finding a set of gene expressions in the data that is especially predictive of the disease is an important study area. Every gene expression is linked to some pathways that trigger them. If there are a few genes that are overexpressed and have predictive significance, they could be a future drug target.

References

1. Daoud M, Mayo M (2019) A survey of neural network-based cancer prediction models from microarray data. *Artif Intell Med* 97:204–214. <https://doi.org/10.1016/j.artmed.2019.01.006>
2. Feltes BC, Chandelier EB, Grisci BI, Dorn M (2019) CuMiDa: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *J Comput Biol* 26(4):376–386. <https://doi.org/10.1089/cmb.2018.0238>

3. Hengpraprophm S, Mukviboonchai S, Thammasang R, Chongstivtavata P (2010) A GA-based classifier for microarray data classification. In: 2010 International conference on intelligent computing and cognitive informatics. IEEE, pp 199–202. <https://doi.org/10.1109/ICI-CCI.2010.62>
4. Alomari OA, Khader AT, Al-Betar MA, Alkareem Alyasseri ZA (2018) A hybrid filter-wrapper gene selection method for cancer classification. In: 2018 2nd international conference on biosignal analysis, processing and systems (ICBAPS). IEEE, pp 113–118. <https://doi.org/10.1109/ICBAPS.2018.8527392>
5. Singh P, Shukla A, Vardhan M (2017) Hybrid approach for gene selection and classification using filter and genetic algorithm. In: 2017 international conference on inventive computing and informatics (ICICI). IEEE, pp 832–837. <https://doi.org/10.1109/ICICI.2017.8365253>
6. Ahmed E, El-Gayar N, El-Azab IA (2010) Support vector machine ensembles using features distribution among subsets for enhancing microarray data classification. In: 2010 10th international conference on intelligent systems design and applications. IEEE, pp 1242–1246. <https://doi.org/10.1109/ISDA.2010.5687078>
7. Gad AF (2021) PyGAD: an intuitive genetic algorithm python library. <https://doi.org/10.48550/arxiv.2106.06158>
8. Pedregosa F et al (2012) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:28252830. <https://doi.org/10.48550/arxiv.1201.0490>
9. Yang XS, Deb S, Hanne T, He X (2019) Attraction and diffusion in nature-inspired optimization algorithms. *Neural Comput Appl* 31(7):1987–1994

Design and Simulation of a Novel Digital Technology for Assembly Operations: A Case Study of Railcar Bogie Application



Ilesanmi Daniyan , Khumbulani Mpofu , Lanre Daniyan, Felix Ale, and Nokulunga Zamahlubi Dlamini 

Abstract The railcar bogie, which consists of many interrelated components, can only meet the required service condition if the components are assembled correctly in line with the standard and specifications. To minimize assembly error and for high productivity and efficiency of operation, this work proposes the design of an intelligent system for the assembly operations of a railcar bogie. The intelligent system consists of the integration of a CCD camera, array of smart sensors, part component box, LED light source, control system, conveyor as well as a robot for material handling and assembly. The component parts of the railcar bogie are classified and stored according to sizes and parts families into standard and non-standard parts. The system which is capable of generating and extracting assembly features, has visual detection and inspection modules, hence, it can plan, execute, monitor and control assembly operations with the aid of a flexible algorithm. A 3D iterative process of design, modelling and simulation of the assembly process was carried in the Simlab software. The results obtained indicated that the system could perform the intelligent coordination of the bogie assembly process with significant improvement in the assembly precision as well as the cycle time during the assembly process. Hence, this work provides an insight into the assembly process automation and intelligent manufacturing of the rail car.

Keywords Assembly process · Bogie · Modelling and simulation · Intelligent system · Railcar

I. Daniyan  · K. Mpofu · N. Z. Dlamini

Department of Industrial Engineering, Tshwane University of Technology, Pretoria 0001, South Africa

e-mail: afolabiilesanmi@yahoo.com

L. Daniyan

Department of Instrumentation, University of Nigeria, Nsukka, Nigeria

F. Ale

Department of Engineering and Space Systems, National Space Research & Development Agency, Abuja, Nigeria

1 Introduction

The railcar bogie is an important component of the railcar bearing and transmission system which supports the railcar body and provides stability on both straight and curved track [1, 2]. It consists of many interrelated components and can only meet the required service condition if the components are assembled correctly in line with the standard and specifications. A well-designed and assembled railcar bogie can also promote safety, significant ride comfort through vibration absorption, reduction of impacts due to centrifugal forces most especially when the train runs on curves at high speed. The quest for improved safety, productivity, reduction in assembly error as well as timely delivery of assembled railcar bogie necessitate the need for the automation of the assembly process. Assembly error could affect the performance of the railcar bogie through rail abrasion, thus, affecting the stability of the railcar most especially along curved path or irregular track profiles. An automated assembly line may allow for more compactness of the assembled components with reduction in dimensional inaccuracies or assembly error because of less human involvement. Besides, the manual assembly process is time consuming, laborious and may be unsuitable for handling complex assembly operations. On the other hand, the automation of the assembly process boasts of significant cost savings over time, and safety of the system [3]. Automation of the assembly process can also promote the reduction in time required for loading and unloading components, or the time required for set up, part identification or changing of production tools [4, 5]. Assembly or machining errors can trigger the development of residual stresses in the bogie components which can affect the structural integrity, performance and useful service life component if not carefully controlled [6, 7].

2 Literature Review

There is an increasing interest in the automation of rolling stock in the quest for improved quality and productivity, efficiency and repeatability [8]. Many authors have reported on the automation of rolling stock for operations such as inspection and monitoring. For instance, Vithanage et al. [9] proposed an automated system that demonstrates capability for conducting inspection operation of railway coupler electric heads using industrial robots. Sasikala and Kishore [3] presented a model for the automation of rail rolling stock with the aid of the computer vision algorithms. The work featured the use of multi-object, multi-template algorithm with a template that can be updated. The results obtained indicated about 91% accuracy of the proposed algorithm.

Daniyan et al. [10] proposed a Flexible Manufacturing System (FMS) for the assembly operations of railcar subassemblies. The simulation of the proposed FMS demonstrated that the conveyor system could promote effective movement of the component part according to the designed sequence of assembly with minimal human

intervention or interruptions. The use of a dual-arm arm, 16-axis robot with intelligent system for railcar manufacturing operations has also been proposed [11]. The simulation of the proposed dual-arm robot show capability for manufacturing operations such as machining, assembly and handling operations during railcar manufacturing. Skosana et al. [12] developed a framework for virtual assembly training system. The work provides a practical guided approach for virtual assembly operation. The use of Virtual Reality (VR) system can enhance the assembly process and improve personnel's skill in a cost-effective manner [13–15]. The application of VR for the simulation of assembly process has gain momentum in this era. It could enhance the training and upskilling of personnel with reduction in assembly errors. Due to the increasing industry evolution, competitiveness, as well as the need for effective responsiveness, it has become essential to integrate VR into the design, training, simulation and commissioning of the industry assembly process. Towey et al. [16] states that VR is the creation of computer-generated 3D interactive worlds, with the intention of giving the user the illusion of being immersed in a different reality. Currently, VR system can be categorized in three classifications based on the immersion level, the type of interfaces or components used in the system. The classifications are the fully immersive, semi-immersive and non-immersive VR system [17]. DeFanti et al. [18] mention that in a semi-immersive system, the user is mostly surrounded by the projected high-resolution screens or walls with images and this surrounding is defined as a cave automatic virtual environment. In non-immersive or desktop VR system, the user observe virtual world in high-resolution monitor and user remains visually conscious of the real world. Fully immersive VR system is administered with Head Mounted Display (HMD), this wearable device consists of small screens in front of both eyes. These screens display the 3D images, magnifies, fill a wide field of view, respond based on the head direction, position of the user and they include earphones by the side of the ears while interaction is achieved by the hand controllers [19]. Assembly operation is an interactive process involving the operator and the handled objects, hence, simulation environments must be able to react according to the user's actions in real time. According to Seth et al. [20] virtual assembly is defined as the ability to assemble virtual representations of physical models through simulating realistic environment behaviour and part interaction. This is to reduce the need for physical assembly prototyping resulting in the ability to make more encompassing design or assembly decisions in an immersive computer-generated environment. In the manufacturing industry arena, Virtual Environment (VE) technology provides a useful method to interactively evaluate assembly-related engineering decisions, and to factor the human elements and considerations into finished products very early in the development cycle [21]. Jayaram et al. [22] explain that although virtual assembly is defined as a technology, it is a combination of several technologies such as advanced visualization, simulation, decision theory, assembly and manufacturing procedures and assembly or manufacturing equipment development.

The existing literature demonstrate various techniques for the automation of the operations and processes of the railcar system. However, there are emerging technologies such as the VR technology for the demonstration of the automation of the assembly process of the railcar bogie. This study employs the VR technology to

simulate the assembly process. The novelty of this study lies in the fact the design and simulation of railcar bogie assembly with the aid of the VR technology has not been sufficiently highlighted by the existing literature.

3 Methodology

To minimize assembly error and for high productivity and efficiency of operation, this work proposes an intelligent system for the assembly operations of a railcar bogie. The intelligent system consists of the integration of a CCD camera, array of smart sensors, part component box, LED light source, control system, conveyor as well as a robot for material handling and assembly. The component parts of the railcar bogie are classified and stored according to sizes and parts families into standard and non-standard parts. The system which is capable of generating and extracting assembly features, has visual detection and inspection modules, hence, it can plan, execute, monitor and control assembly operations with the aid of a flexible algorithm. A 3D iterative process of design, modelling and simulation of the assembly process was carried in the Simlab soft virtual environment. Simlab software is a complete wide-ranging set of creative tools for game development, design, training, visualization, rendering, to create fully interactive VR and real time applications.

3.1 Assembly Process of the Railcar Bogie and Its Enabling Technology

The following are the railcar bogie components that are to be assembled in a single functional bogie system: (1) Bogie frame (a chassis that carries a wheelset). It is a modular subassembly of the wheels and axles. (2) Wheelset comprising of the wheel at each end. (3) A rotatable axle box comprising of springs (rubber airbags) to permit up and down movements and rolling bearings. The rotatable axle is usually attached to the wheelset. (4) Suspension system: usually the primary suspension system, which connects the wheelset to the bogie frame, and the secondary suspension, which connects the bogie frame to the railcar bogie. (5) Link arm. (6) Adapter assembly. (7) Bolster: a cross member connected to the bogie frame via the secondary suspension. (8) Braking system: comprising of brake shoe, brake beam, brake discs and pads. (9) Transmission system: which comprises of the traction motor, speed gearbox, etc. For the automation of the assembly system, the designed intelligent system consists of the integration of the following components: CCD camera, array of smart sensors, part component box, LED light source, control system, conveyor, grinding robot, robot for material handling and assembly, as well as automated guided vehicle for moving heavy assembled subassembly or moving heavy components. Some of the railcar components are structural members welded from steel plates, which requires

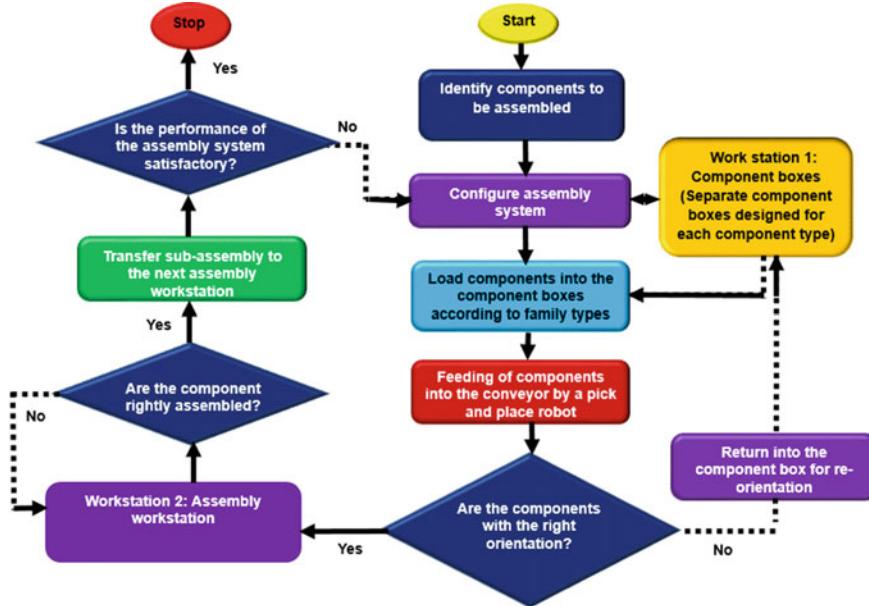


Fig. 1 Flowchart of the assembly operation

grinding operations. The accuracy of the grinding operation requires the automation of the process. The manual grinding process may be prone to error most especially in complex parts of the components. Hence, this study proposes the use of grinding robotic cells with the capacity of 400 kg payload embedded with a tactile sensor for measuring the surface and position of the work piece. The loading and unloading of the components or subassembly will be done using the automated guided vehicle. The purpose of the CCD camera with high-resolution is to capture and store images in real time. This will allow for collection, organization, storage and distribution of the captured images. Figure 1 presents the flowchart of the assembly operation.

The assembly system is configured as a continuous synchronous system consisting of sequence of workstations in a straight-line arrangement as shown in Fig. 2. The first workstation will contain the component boxes. The components are loaded into the component boxes. There are separate component boxes designed for each component type. Each component to be assembled is loaded according to their part family (type) into their respective component boxes in a random manner. A pick and place robot embedded with sensors has a component feeding mechanism to pick the components to be assembled from the component boxes in a programmed sequence of assembly operation and place them on the conveyor system. The conveyor system will then move the component to the assembly robot that will perform the assembly operation. Initially, the components will be in a random orientation, hence, the pick and place robot are embedded with a selector to ensure that only properly oriented components are selected into the conveyor. Only correctly oriented components will

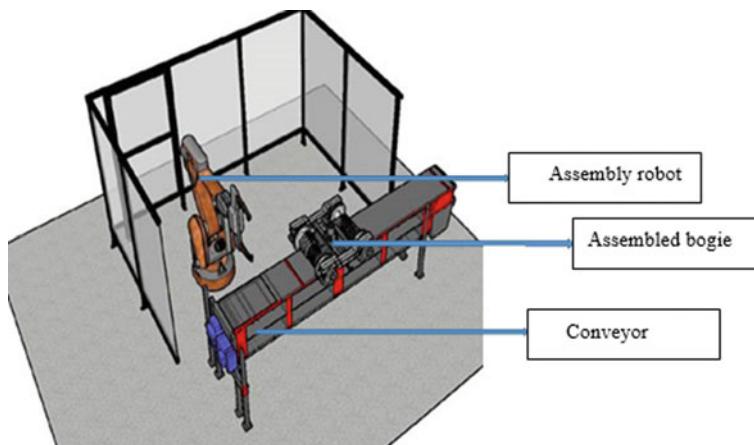


Fig. 2 The layout of the assembly floor

be selected into the conveyor's track while others with incorrect orientation will be returned to the component box for re-orientation. The conveyor will transfer the component received from the pick and place robot into the assembly workstation while maintaining proper orientation of the components during the movement.

The second robot, which is the assembly robot, has a placement device that places the components in the correct location in the assembly operation workstation. When the delivery rate of the component exceeds the cycle rate of the assembly system, a high-level optical or limit sensor placed on the conveyor can assist in turning off the feeding mechanism once the conveyor is full. Another sensor placed along the conveyor system at a certain distance from the initial one can assist in restarting the feeding mechanism once the queue on the conveyor system disappears.

Figure 3 shows the Proteus diagram of the intelligent system for assembly operation. It comprises of two precision stepper motors (Robots A and B), one for pick and place operations and the other for the assembly operation. Below each of the stepper motors is their drivers (L296H-Bridge). At the centre is the main controller (ATMEG) with the written code of the sequence of the assembly operations. D1 and D2 are diode LEDs to indicate the status of the component movements. D1 turns ON when robot A is placing a component on the conveyor. This implies that robot A is active while placing the component and turns OFF once it has placed the component successfully. D2 on the other hand, turns ON when robot B is in operation (when the component has arrived for assembly operation). A trans-receiver sensor is placed in between robot B to transmit a code to alert robot B to wait to receive a component once D1 turns ON and vice versa. This is a coordinating sensor, which directs robot A to pick and place a component on the conveyor and alert B to be ready to receive it.

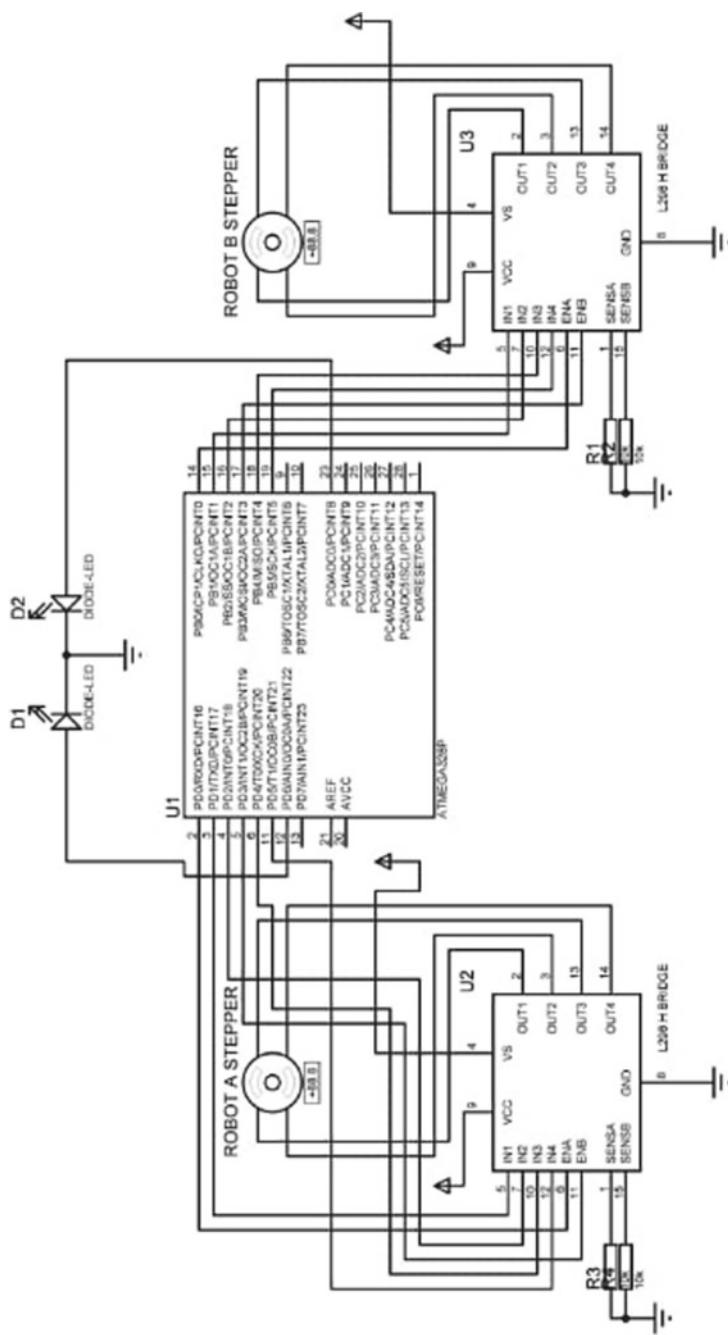


Fig. 3 The Proteus diagram of the robot circuitry

The probability that there is no defective component in workstation i (w_i) is expressed as Eq. 1 while Eq. 2 expresses the probability that there is a defective component in workstation i (w_i) which will likely affect the efficiency or sequence of the assembly operation.

$$P_i = 1 - d_i \quad (1)$$

$$P_i = w_i d_i \quad (2)$$

where d_i is the defective component in workstation i . Assuming the defective component does not affect the efficiency or sequence of assembly operation at workstation i , Eq. 3 holds thus:

$$P_i = (1 - w_i) d_i \quad (3)$$

In a situation whereby, a defective component is assembled to an existing subassembly, the probability of obtaining an acceptable subassembly P_a is given by Eq. 4.

$$P_a = \prod_{i=1}^n [w_i d_i + (1 - d_i)] \quad (4)$$

The probability that there is a downtime in the assembly due to defective component (P_d) is expressed by Eq. 5.

$$P_d = \sum_{i=1}^n P_i = \sum_{i=1}^n (w_i d_i) \quad (5)$$

Therefore, the average assembly time per assembly is expressed as Eq. 6.

$$T_a = T_c + \sum_{i=1}^n (w_i d_i) T_d \quad (6)$$

where T_c is the ideal cycle time (secs) and T_d is the average downtime per occurrence (seconds).

Equation 7 expresses the average rate of acceptable assembled product (R_p) while Eq. 8 presents the efficiency of the assembly line. The efficiency of the assembly line is the ratio of the ideal cycle time to the average assembly time.

$$R_p = \frac{1}{T_a} \quad (7)$$

$$e = \frac{T_c}{T_a} \quad (8)$$

The downtime of the assembly line is the ratio of the average downtime per cycle to the average assembly time expressed as Eq. 9.

$$D = \frac{T_d}{T_a} \quad (9)$$

3.2 *Simulation of the Virtual Assembly*

Simulating the virtual assembly process requires the creation of 3D models, tools and importing them to the Simlab soft platform. These models are textured to enhance its visibility and to enable its realism. Figure 4 shows a fully immersive virtual environment of the assembly robot performing grinding operation on different subassembly process of the bogie system. The creation of fully interactive and immersive three-dimensional (3D) visualization can be used for factory design and for planning factory operations and assembly operations.

It is vital to create a suitable assembly plan in the initial design stage. A good assembly plan integrates the factors to consider such as minimum assembly time, low cost, ergonomics and operator safety. Thus, a suitable plan and properly designed assembly process can improve production efficiency and product quality. It can also reduce the manufacturing cost and shorten product's time to market.

Technologies that allow for virtual assembly design, evaluation and analysis are not yet fully applied and embraced by the industries. Although this emerging virtual reality technology is not completely recognized concerning its applications within commercial industries, the technology as a whole is viewed as viable and valuable.

To fully utilize virtual assembly and simulation of the bogie system, it is crucial to understand some key factors that can enhance how the virtual assembly applications enable engineers to attain cohesive view of assembly issues and errors that may arise. It is also necessary to understand how the novel digital technology virtual assembly system will assist the engineers in making decisions and making the whole system effective and efficient. Thus, this novel digital system should be applied in real life design, and assembly. It may aid in meeting existing productions and future needs. Lastly, vital factors must be considered on the daily ease of use of the system looking at the ergonomics, safety issues, human and computer interfaces.

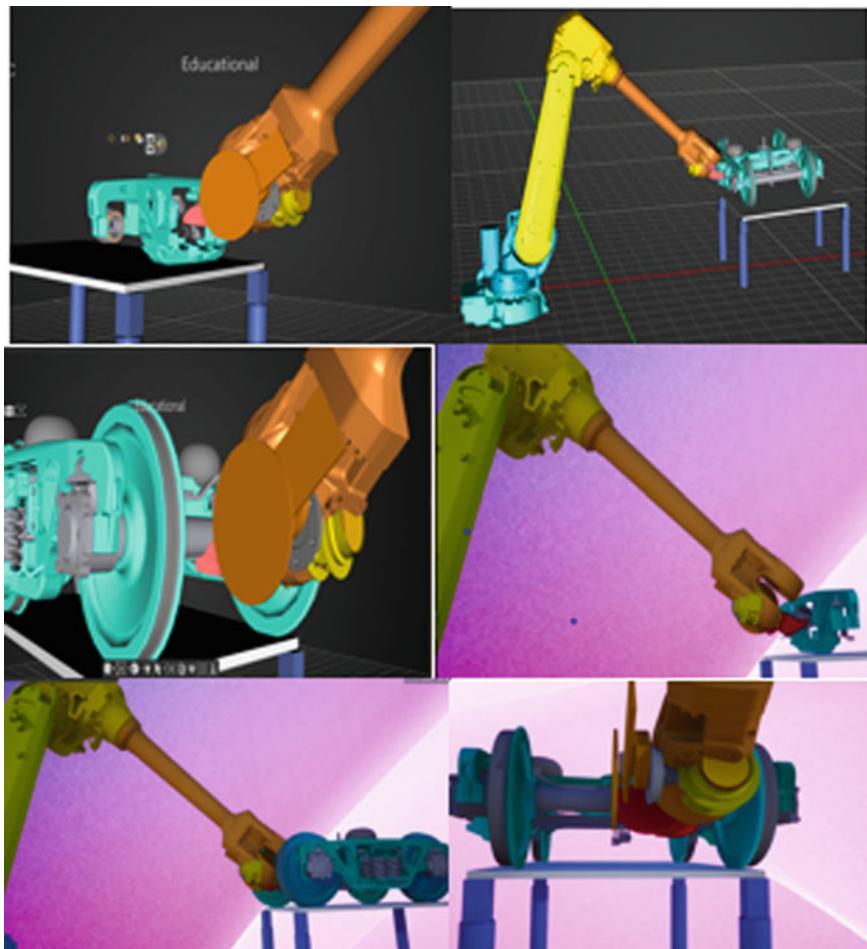


Fig. 4 Fully immersive virtual environment of the assembly robot performing grinding operation

4 Conclusions

The aim of this work was to design an intelligent system for the assembly operations of a railcar bogie. Furthermore, a 3D iterative process of design, modelling and simulation of the assembly process was carried in the Simlab software. The simulation carried out demonstrated the use of the proposed system to perform intelligent coordination of the bogie assembly process. Hence, this work provides an insight into the assembly process automation and intelligent assembly of a railcar bogie. It also highlights the possibility for the design and simulation of railcar bogie assembly process

with the aid of the VR technology. The work is limited to the design and simulation of the assembly process of a railcar bogie. Hence, future work can consider the development of the intelligent system and the performance evaluation.

Acknowledgements Funding: The authors disclosed receipt of the following financial support for the research: Technology Innovation Agency (TIA) South Africa, Gibela Rail Transport Consortium (GRTC), National Research Foundation (NRF grant 123575) and the Tshwane University of Technology (TUT)."

References

1. Daniyan, IA, Mpofu, K. Fameso, FO, Adeodu AO. Numerical simulation and experimental validation of the welding operation of the railcar bogie frame to prevent distortion. *Int J Adv Manuf Technol* 106:5213–5224 (2020)
2. Daniyan IA, Adeodu AO, Mpofu K, Maladhzi R, Kana-kana Katumba MG (2022) Application of lean Six Sigma methodology using DMAIC approach for the improvement of bogie assembly process in the railcar industry. *Heliyon* 8(e09043):1–14
3. Sasikala N, Kishore PVV (2020) Train bogie part recognition with multi-object multi-template matching adaptive algorithm. *J King Saud Univ Comput Inf Sci* 32:608–617
4. Gonzalo O, Seara JM, Guruceta E, Izpizua A, Esparza M, Zamakona I, Uterga N, Aranburu A, Thoelen J (2017) A method to minimize the workpiece deformation using a concept of intelligent fixture. *Robot Comput-Integr Manuf* 48:209–218
5. Glassman M, Kang MJ (2012) Intelligence in the internet age: the emergence and evolution of Open Source Intelligence (OSINT). *Comput Hum Behav* 28(2):673–682
6. Zeinoddini M, Arnavaz S, Zandi AP, Vaghlasloo YA (2013) Repair welding influence on offshore pipelines residual stress fields: an experimental study. *J Constr Steel Res* 86:31–41
7. Song S, Dong P (2017) Residual stresses at weld repairs and effects of repair geometry. *Sci Technol Weld Join* 22:265–277
8. Piln L, Bissacco G (2015) Development of on the machine process monitoring and control strategy in robot assisted polishing. *CIRP Ann Manuf Technol* 64:313–316
9. Vithanage RKW, Harrison CS, De Silva AKM (2019) Autonomous rolling-stock coupler inspection using industrial robots. *Robot Comput Integr Manuf* 59:82–91
10. Daniyan IA, Mpofu K, Ramatsetse BI, Zeferino E, Monzambe G, Sekano E (2021) Design and simulation of a flexible manufacturing system for manufacturing operations of railcar subassemblies. *Procedia Manuf* 54:112–117
11. Daniyan IA, Mpofu K, Ale F, Oyesola MO (2021) Design and simulation of a dual-arm robot for manufacturing operations in the railcar industry. *Int J Robot Autom* 36(6):434–447
12. Skosana XN, Mpofu K, Trimble J, van Wyk EA (2022) An empirical framework for developing and evaluating a virtual assembly training system in learning factories. *Interact Learn Environ*, pp 1–18
13. Abidi MH, Al-Ahmari A, Ahmad A, Ameen W, Alkhalefah H (2019) Assessment of virtual reality-based manufacturing assembly training system. *Int J Adv Manuf Technol* 105(9):3743–3759
14. Eschen H, Kötter T, Rodeck R, Harnisch M, Schüppstuhl T (2018) Augmented and virtual reality for inspection and maintenance processes in the aviation industry. *Procedia Manuf* 19:156–163
15. Koumaditis K, Chinello F, Mitkidis P, Karg S (2020) Effectiveness of virtual versus physical training: the case of assembly tasks, trainer's verbal assistance, and task complexity. *IEEE Comput Graphics Appl* 40(5):41–56

16. Towey D, Walker J, Austin C, Kwong CF, Wei S (2018) Developing virtual reality open educational resources in a Sino-foreign higher education institution: challenges and strategies. In: 2018 IEEE international conference on teaching, assessment, and learning for engineering (TALE). IEEE, pp 416–422
17. Bamodu O, Ye XM (2013) Virtual reality and virtual reality system components. In: Advanced materials research, vol 765. Trans Tech Pub. Ltd., pp 1169–1172
18. DeFanti T, Acevedo D, Ainsworth R, Brown M, Cutchin S, Dawe G, Doerr KU, Johnson A, Knox C, Kooima R, Kuester F (2011) The future of the CAVE. *Open Eng* 1(1):16–37
19. Van Wyk EA (2015) An evaluation framework for virtual reality safety training systems in the South African mining industry. PhD thesis
20. Seth A, Vance JM, Oliver JH (2011) Virtual reality for assembly methods prototyping: a review. *Virtual Real* 15(1):5–20
21. Dai F (1998) Introduction—beyond walkthroughs. In: Virtual reality for industrial applications. Springer, Berlin, Heidelberg, pp 1–9
22. Jayaram S, Connacher HI, Lyons KW (1997) Virtual assembly using virtual reality techniques. *Comput Aided Des* 29(8):575–584

Solving Fixed Charge Transportation Problem with Interval Parameters Using Generalized Reduced Gradient Method



Subhayan Das and Subhra Das 

Abstract Fixed charge transportation problem with interval parameters is studied in this paper. The total cost of transportation comprises of a fixed charge and transportation cost per unit of commodity transported from origin to destinations. Thus, unlike classical transportation problem, the objective function of fixed charge transportation problem is a non-linear function. An algorithm has been proposed for solving fixed charge transportation problem with interval parameters using generalized reduced gradient method subject to the supply and demand constraints. An optimal interval solution for the total cost of transportation is obtained using the proposed method within which the total cost will lie which satisfies all the supply and demand constraints of the problem. The optimal result obtained using the proposed algorithm is compared to the order relation \leq_{RC} and \leq_{HW} model reported by researchers in the past which computes optimal interval for the total cost for transportation based on the upper limit of supply and lower limit for demand. The coefficient of variations of the optimal range for total cost obtained from \leq_{RC} , \leq_{HW} and proposed model are 23%, 2% and 20%, respectively. The proposed model gives a better approximation to the optimal interval solution for the problem compared to \leq_{RC} . Although \leq_{HW} model provides a better approximation to the optimal interval solution compared to the proposed model, but it does not consider the entire set of constraints to evaluate the optimal interval for the cost function.

Keywords Fixed charge transportation problem · Interval parameters · Generalized reduced gradient method

S. Das (✉)

School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha, India

e-mail: nips.subhayan@gmail.com

S. Das

Amity School of Engineering and Technology, Amity University Haryana, Gurugram, India

1 Introduction

Transportation problem is a special type of linear programming problem (LPP) where the objective is to minimize the cost of transporting product from different sources to various destinations. French mathematician Gaspard Monge formulated this problem [1] in 1781. A. N. Tolstoi first proposed the mathematical formulation of the transportation problem (TP) in 1920. Although TPs can be written as a LPP but solving it using Simplex method is not suitable because of its structure. Several methods like north-west corner method, matrix minima, Vogel's approximation methods are used to find a feasible solution for the problem [2].

Transportation problems may have a single objective function or multiple objective function which are optimized subject to demand and supply constraints [3]. The fixed charge transportation problem [4] was formulated by Hirsch and Dantzig in 1954 and is a special type of TP where the cost of transportation of commodity from a source to the destination comprises of a fixed charge and a transportation cost per unit of the quantity transported. The objective function that needs to be optimized in this case is thus a non-linear function of the two cost parameters. A fixed cost transportation problem (FCTP) [5] can be expressed as a linear programming problem as follows: Consider m sources having a_i , $i = 1, 2, \dots, m$ units of the commodity which is transported to n destinations having demand b_j , $j = 1, 2, \dots, n$. The cost of transportation of a unit of commodity from i th source to j th destination comprises of the cost c_{ij} per unit of delivery and fixed charges f_{ij} which is independent of quantity of the commodity delivered. The mathematical formulation of the FCTP [6] is as follows:

$$\text{Minimize } Z = \sum_{i=1}^m \sum_{j=1}^n (c_{ij}x_{ij} + f_{ij}\alpha_{ij}) \quad (1)$$

Subject to the conditions

$$\sum_{j=1}^n x_{ij} \leq a_i, \quad i = 1, 2, 3, \dots, m$$

$$\sum_{i=1}^m x_{ij} \geq b_j, \quad j = 1, 2, 3, \dots, n$$

$$x_{ij} \geq 0 \text{ for all } i = 1, 2, 3, \dots, m, \quad j = 1, 2, 3, \dots, n$$

$$\alpha_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} > 0 \end{cases}$$

$$\sum_{i=1}^m a_i \geq \sum_{j=1}^n b_j$$

Total supply available at the sources should be either equal to or greater than the demand of the destinations. The supply, demand and the costs are all fixed in this type of problem. Algorithms are applied to optimize the objective function which satisfies the constraints. There are various traditional algorithms to obtain optimal solution while there are non-traditional algorithms like genetic algorithm, particle swarm algorithm, firefly algorithm, information utilization-based modified differential evolution algorithm [7], analytical hierarchy process [8] algorithms to solve various multi-criteria decision making problems. In real life situation, the supply at the source and demand at the destinations vary as per the market conditions and the cost of transportation also varies based on several conditions like hike in oil prices, increase in tax and so on. Researchers have formulated such type of transportation problem where the supply, demand, cost of transportation and fixed charges are variable and has classified the problem as fixed charge transportation problem with interval parameter [9]. They have applied order relation models like \leq_{RC} and \leq_{HW} to compute the optimal interval for the total cost of transportation by reducing the problem into an equivalent crisp problem with two objective functions which satisfies the constraints for supply and demand. The constraints considered here for computing the optimal solution considers only the upper limit for the supply at the sources and lower limit of demands at destinations to convert the interval parameters for supply and demand to crisp form.

The objective of the paper is to propose an algorithm to solve a fixed charge transportation problem with interval parameters which considers all the constraints of supply and demand to obtain the optimal solution unlike the \leq_{RC} and \leq_{HW} models reported in the literature which considers only the upper limit for supply constraints and lower limit of demand constraints to find the optimal interval for total cost of transportation. Solver in excel is used to solve the problem by applying the proposed algorithm.

2 Fixed Charge Transportation Problem (FCTP) with Interval Parameter

For a fixed charge transportation problem with interval parameters, the cost, fixed charges, supplies and demand are variable having an upper and lower bound for each of these parameters. The cost matrix corresponding to this type of problem with m sources and n destinations is shown in Table 1.

Here S_i represents the sources with a supply varying within the interval $[a_{iL}, a_{iR}]$ and D_j represents the destinations with demand varying within the interval $[b_{jL}, b_{jR}]$. The cost of transportation of unit commodity from i th source to j th destination vary within the interval $[C_{ijL}, C_{ijR}]$ and fixed charges vary within the interval $[f_{ijL}, f_{ijR}]$. The FCTP with interval parameters can be expressed mathematically using the order relation \leq_1 [9] as:

Table 1 Cost matrix for FTCP with interval parameters

	D_1	D_2	D_3	D_4		D_n	Supply
S_1	($[C_{11L}, C_{11R}], [f_{11L}, f_{11R}]$)	($[C_{12L}, C_{12R}], [f_{12L}, f_{12R}]$)	($[C_{13L}, C_{13R}], [f_{13L}, f_{13R}]$)	($[C_{12L}, C_{12R}], [f_{12L}, f_{12R}]$)	...	($[C_{1nL}, C_{1nR}], [f_{1nL}, f_{1nR}]$)	$[a_{1L}, a_{1R}]$
S_2	($[C_{21L}, C_{21R}], [f_{21L}, f_{21R}]$)	($[C_{22L}, C_{22R}], [f_{22L}, f_{22R}]$)	($[C_{13L}, C_{13R}], [f_{13L}, f_{13R}]$)	($[C_{24L}, C_{24R}], [f_{24L}, f_{24R}]$)	...	($[C_{2nL}, C_{2nR}], [f_{2nL}, f_{2nR}]$)	$[a_{2L}, a_{2R}]$
S_3	($[C_{31L}, C_{31R}], [f_{31L}, f_{31R}]$)	($[C_{32L}, C_{32R}], [f_{32L}, f_{32R}]$)	($[C_{33L}, C_{33R}], [f_{33L}, f_{33R}]$)	($[C_{34L}, C_{34R}], [f_{34L}, f_{34R}]$)	...	($[C_{3nL}, C_{3nR}], [f_{3nL}, f_{3nR}]$)	$[a_{3L}, a_{3R}]$
...
S_m	($[C_{m1L}, C_{m1R}], [f_{m1L}, f_{m1R}]$)	($[C_{m2L}, C_{m2R}], [f_{m2L}, f_{m2R}]$)	($[C_{m3L}, C_{m3R}], [f_{m3L}, f_{m3R}]$)	($[C_{m4L}, C_{m4R}], [f_{m4L}, f_{m4R}]$)	.	($[C_{mnL}, C_{mnR}], [f_{mnL}, f_{mnR}]$)	$[a_{mL}, a_{mR}]$
Demand	$[b_{1L}, b_{1R}]$	$[b_{2L}, b_{2R}]$	$[b_{3L}, b_{3R}]$	$[b_{4L}, b_{4R}]$...	$[b_{nL}, b_{nR}]$	

$$\text{Minimize } Z = \sum_{i=1}^m \sum_{j=1}^n ([c_{ijL}, c_{ijR}] x_{ij} + [f_{ijL}, f_{ijR}] \alpha_{ij}) \quad (2)$$

Subject to the conditions

$$\sum_{j=1}^n x_{ij} \leq_1 [a_{iL}, a_{iR}], \quad i = 1, 2, 3, \dots, m$$

$$\sum_{i=1}^m x_{ij} \geq_1 [b_{jL}, b_{jR}], \quad j = 1, 2, 3, \dots, n$$

$$x_{ij} \geq 0 \text{ for all } i = 1, 2, 3, \dots, m, \quad j = 1, 2, 3, \dots, n$$

$$\alpha_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} > 0 \end{cases}$$

where the inequality relations denoted by \leq_1 and \geq_1 are defined as follows [9]:

$$\begin{aligned} t \leq_1 [a, b] &\equiv \exists z \in [a, b]; \quad t \leq z \\ t \geq_1 [a, b] &\equiv \exists z \in [a, b]; \quad t \geq z. \end{aligned} \quad (3)$$

The decision variables in problem (2) takes values within a given range but decision makers always prefer a crisp value of these variables that minimizes the cost of transportation.

An algorithm based on the order relation \leq_{RC} is reported in literature [9] which converts the problem (2) into an equivalent crisp problem using order relation \leq_{RC} . The equivalent crisp problem considers optimizing two objective functions corresponding to the right bound Z_R and its center Z_C subject to the supply and demand constraints as defined below:

$$\text{Minimize } Z_R = \sum_{i=1}^m \sum_{j=1}^n (c_{ijR}x_{ij} + f_{ijR}\alpha_{ij}) \quad (4)$$

$$\text{Minimize } Z_C = \sum_{i=1}^m \sum_{j=1}^n (c_{ijC}x_{ij} + f_{ijC}\alpha_{ij}) \quad (5)$$

Subject to the constraints:

$$\sum_{j=1}^n x_{ij} \leq a_{iR}; \quad i = 1, 2, \dots, m \quad (6)$$

$$\sum_{i=1}^m x_{ij} \geq b_{jL}; \quad j = 1, 2, \dots, n$$

$$\alpha_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} > 0 \end{cases}$$

The optimal interval value for the above problem is obtained by considering only the right bound of the supply constraints and lower bound of the demand constraints. The result obtained by this method may not represent the actual scenario as it is not based on the entire constraint space.

3 Proposed Algorithm

The proposed algorithm solves the fixed charge transportation problem with interval parameter considering all the constraints for demand and supply.

Step 1: The problem is first written as an equivalent crisp problem for obtaining the optimal result. The fixed charge transportation problem with interval parameter is solved by solving the equivalent crisp problem involving the following two objective functions corresponding to the upper and lower bounds:

$$\text{Minimize } Z_R = \sum_{i=1}^m \sum_{j=1}^n (c_{ijR}x_{ij} + f_{ijR}y_{ij}) \quad (7)$$

$$\text{Minimize } Z_L = \sum_{i=1}^m \sum_{j=1}^n (c_{ijL}x_{ij} + f_{ijL}y_{ij}) \quad (8)$$

Subject to the constraints:

Supply Constraints

$$\sum_{j=1}^n x_{ij} \leq a_{iR}; \quad i = 1, 2, \dots, m \quad (9)$$

$$\sum_{j=1}^n x_{ij} \geq a_{iL}; \quad i = 1, 2, \dots, m \quad (10)$$

Demand Constraints

$$\sum_{i=1}^m x_{ij} \geq b_{jL}; \quad j = 1, 2, \dots, n \quad (11)$$

$$\sum_{i=1}^m x_{ij} \leq b_{jR}; \quad j = 1, 2, \dots, n \quad (12)$$

$$x_{ij} \geq 0 \text{ for all } i = 1, 2, 3, \dots, m, \quad j = 1, 2, 3, \dots, n$$

And

$$y_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} > 0 \end{cases}$$

Step 2: Find the optimal solution z_R^* to the problem with objective function corresponding to the upper bounds of the parameters (i.e., Minimize Z_R) subject to the constraints using generalized reduced gradient method [10].

Step 3: Find the optimal solution z_L^* to the problem with objective function corresponding to the lower bounds of the parameters (i.e., Minimize Z_L) subject to the constraints using generalized reduced gradient method.

Step 4: Optimal Solution for the total cost of transportation is obtained as an interval given by $[z_L^*, z_R^*]$.

4 Numerical Example

The fixed charge transportation problem with interval parameter discussed in this section has been taken from the work done by Safi and Razmjoo [9]. There are three sources O_1 , O_2 and O_3 which supplies raw material to four different destinations A, B, C and D. The demand at the destinations and the available supply at the sources are not constant but varies within a given range as shown in Table 2. A fixed charge is charged by the company which does the shipment, and it varies in the range of $[f_{ijL}, f_{ijR}]$. Apart from these fixed charges, the cost of transportation from i th source to j th destination also varies over a given range $[C_{ijL}, C_{ijR}]$. The problem is to find the best strategy of transportation of raw materials from the sources to the destinations which minimizes the total cost of transportation satisfying the constraints.

The FCTP with interval parameters is expressed as an equivalent crisp problem as:

$$\begin{aligned} \text{Min } Z_R = & 8x_{11} + 12x_{12} + 11x_{13} + 10x_{14} + 18x_{21} + 12x_{22} + 25x_{23} + 7x_{24} \\ & + 19x_{31} + 12x_{32} + 14x_{33} + 17x_{34} + 30y_{11} + 25y_{12} + 25y_{13} \\ & + 30y_{14} + 20y_{21} + 25y_{22} + 55y_{23} + 49y_{24} + 20y_{31} + 30y_{32} \\ & + 50y_{33} + 22y_{34} \end{aligned} \quad (13)$$

$$\begin{aligned} \text{Min } Z_L = & 4x_{11} + 8x_{12} + 9x_{13} + 8x_{14} + 10x_{21} + 10x_{22} + 11x_{23} + 5x_{24} \\ & + 7x_{31} + 8x_{32} + 8x_{33} + 13x_{34} + 10y_{11} + 19y_{12} + 19y_{13} + 20y_{14} \\ & + 16y_{21} + 15y_{22} + 25y_{23} + 38y_{24} + 10y_{31} + 22y_{32} + 30y_{33} + 20y_{34} \end{aligned} \quad (14)$$

Subject to the constraints

Supply constraints

Table 2 Cost matrix for the FCTP with interval parameters [9]

Sources	Destinations				Supply
	A	B	C	D	
O_1	([4, 8], [10, 30])	([8, 12], [19, 25])	([9, 11], [19, 25])	([8, 10], [20, 30])	[30, 33]
O_2	([10, 18], [16, 20])	([10, 12], [15, 25])	([11, 25], [25, 55])	([5, 7], [38, 40])	[27, 28]
O_3	([7, 19], [10, 20])	([8, 12], [22, 30])	([8, 14], [30, 50])	([13, 17], [20, 22])	[22, 25]
Demand	[20, 21]	[19, 24]	[23, 24]	[20, 22]	

$$\begin{aligned}
x_{11} + x_{12} + x_{13} + x_{14} &\geq 30 \\
x_{11} + x_{12} + x_{13} + x_{14} &\leq 33 \\
x_{21} + x_{22} + x_{23} + x_{24} &\geq 27 \\
x_{21} + x_{22} + x_{23} + x_{24} &\leq 28 \\
x_{31} + x_{32} + x_{33} + x_{34} &\geq 22 \\
x_{31} + x_{32} + x_{33} + x_{34} &\leq 25
\end{aligned} \tag{15}$$

Demand Constraints

$$\begin{aligned}
x_{11} + x_{21} + x_{31} &\geq 20 \\
x_{11} + x_{21} + x_{31} &\leq 21 \\
x_{12} + x_{22} + x_{32} &\geq 19 \\
x_{12} + x_{22} + x_{32} &\leq 24 \\
x_{13} + x_{23} + x_{33} &\geq 23 \\
x_{13} + x_{23} + x_{33} &\leq 24 \\
x_{14} + x_{24} + x_{34} &\geq 20 \\
x_{14} + x_{24} + x_{34} &\leq 22 \\
x_{ij} &\geq 0 \text{ for all } i = 1, 2, 3, \quad j = 1, 2, 3, 4
\end{aligned} \tag{16}$$

$$y_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} > 0 \end{cases}$$

The problem is solved using Solver in excel by taking each objective function separately and minimizing it w.r.t the constraints using generalized reduced gradient method. The optimal solution corresponding to the objective function $\text{Min } Z_R$ given by Eq. (13) subject to the supply and demand constraints given by Eqs. (15) and (16), respectively, is tabulated in Table 3.

The optimal solution corresponding to the objective function $\text{Min } Z_L$ given by Eq. (14) subject to the supply and demand constraints given by Eqs. (15) and (16), respectively, is tabulated in Table 4.

Table 3 Optimal solution for Z_R

Decision variables	x_{11}	x_{12}	x_{13}	x_{14}	x_{21}	x_{22}	x_{23}	x_{24}	x_{31}	x_{32}	x_{33}	x_{34}	Z_R
Optimal values	20	0	13	0	0	7	0	20	0	12	10	0	1011

Table 4 Optimal solution for Z_L

Decision variables	x_{11}	x_{12}	x_{13}	x_{14}	x_{21}	x_{22}	x_{23}	x_{24}	x_{31}	x_{32}	x_{33}	x_{34}	Z_L
Optimal values	20	10	0	0	0	0	7	20	0	9	16	0	681

Therefore, the optimal cost of transportation which satisfies all the constraints of supply and demand lies within the interval [681, 1011].

5 Validation of Result

The optimal interval solution for the total cost of transportation obtained following the proposed method is compared with that using the order relation \leq_{RC} and \leq_{HW} reported in [9]. The results are tabulated in Table 5.

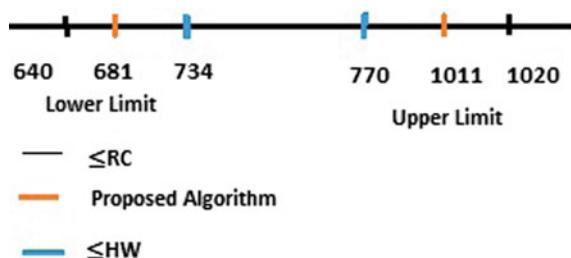
The result obtained by the proposed method is comparable with the solution obtained by other two methods while addressing all the supply and demand constraints to optimize the objective functions and is shown on the number line in Fig. 1.

It is observed that \leq_{RC} model provides a wider range for the optimal value of total cost of transportation with coefficient of variation of 23%. On the other hand, \leq_{HW} reported a better estimation for the total cost with only 2% coefficient of variation. The proposed method provides a better approximation for the range of the total cost of transportation compared to \leq_{RC} model with 20% coefficient of variation. Although

Table 5 Comparison of optimal solution

Method	Optimal interval solution for total cost of transportation		Remarks	References
	Lower limit	Upper limit		
\leq_{RC}	640	1020	Only upper limit of supply and lower bound of demand has been considered for optimization. Thus, ignores the lower limit of supply and upper limit of the demand while computing the optimal solution	[9]
\leq_{HW}	734	770		[9]
Proposed method	681	1011	Both the upper and lower limit of supply and demand has been considered to compute the optimal solution	

Fig. 1 The lower limit and upper limit of the optimal solution obtained by the order relation \leq_{RC} and \leq_{HW} and proposed algorithm is shown on the number line



\leq_{HW} reported a result better than the proposed method, but it does not consider all the constraints for demand and supply; thus, the result may not be representing the optimal value of Z subject to all constraints.

6 Conclusions

The fixed charge transportation problem with interval parameters involves is solved by converting it into an equivalent crisp problem with two objective functions corresponding to the upper and lower limits for the cost and fixed charges. The objective function obtained thus are non-linear and requires special treatment for optimizing them subject to the given constraints for supply and demand which varies over a given range of values. The constraints are also expressed by their equivalent crisp form by considering both the upper and lower bounds of the supply and demand.

The crisp problem thus obtained is like a linear programming problem with a non-linear objective function. Thus, it cannot be solved using simplex method and hence generalized reduced gradient method is used for minimizing the objective functions subject to the supply and demand constraints. The problem is solved using Solver in excel which reduces the complexity of calculation and provides an optimal solution which agrees with the result reported by researchers in the past.

The proposed algorithm is simple and considers all the constraints imposed on each of the variables while minimizing the objective function. Thus, it can be concluded that it provides a better solution for the problem.

References

1. Hadly G (1974) Linear programming. Addison-Wesley, Publishing Company, Inc.
2. Das S (2022) Analyzing impact of decision maker's strategy on performance parameters of bi-criteria transportation problem. *Int J Adv Oper Manag* 14:1–15
3. Aneja YP, Nair KPK (1979) Bicriteria transportation problem. *Manag Sci* 25(1):73–79
4. Hirsch WM, Dantzig GB (1968) The fixed charge problem. *Naval Res Log Q* 15(3):413–424
5. Robers P, Cooper L (1976) A study of the fixed charge transportation problem. *Comput Math Appl* 2:125–135
6. Zhu K, Fan Y, Shen J, Li Y, Yin M (2022) An uncertain programming model for fixed charge transportation problem with after-sale service. *Hindawi J Math* 2022:Article ID 8411876, 12 pp. <https://doi.org/10.1155/2022/8411876>
7. Kumar P, Pant M, Singh HP (2018) Solving nonlinear optimization problems using IUMDE algorithm. In: Pant M et al (eds) Soft computing: theories and applications. Advances in intelligent systems and computing, vol 584, pp 245–254. https://doi.org/10.1007/978-981-10-5699-4_24
8. Sharawat K, Dubey SK (2018) An approach to vendor selection on usability basis by AHP and fuzzy Topsis method. In: Pant M et al (eds) Soft computing: theories and applications. Advances in intelligent systems and computing, vol 584, pp 595–604. https://doi.org/10.1007/978-981-10-5699-4_56

9. Safi MR, Razmjoo A (2013) Solving fixed charge transportation problem with interval parameters. *Appl Math Model* 37:8341–8347
10. Faco JLD (1989) A generalized reduced gradient algorithm for solving large-scale discrete-time nonlinear optimal control problems. *Nonlinear Programming and Optimization*, Paris, France

Automated Solar PV Array Cleaning Based on Aerial Computer Vision Framework



Shreya Nallapaneni, Kairavi Shah, and Harsh S. Dhiman

Abstract Solar panels use solar energy radiation to generate electrical energy. Extracting the sun's maximum energy level minimizes installation costs and helps to fulfill peak electrical demand. However, physical conditions such as bird droppings, dust, and grime might hinder the solar panel's exposure to the sun. This accumulation of dirt might impair the panel's ability to extract maximum power. Module cleaning can improve overall efficiency and performance by 21% in residential modules and up to 60% in commercial installations. The present human-based cleaning methods for solar panels are inefficient in terms of time, water, and energy consumption. Conversely, an AI-based autonomous cleaning drone in conjunction with computer vision helps to boost efficiency by preventing soiling. In addition, cleaning drones will be around 25% less expensive than conventional panel cleaning processes.

Keywords Aerial imagery · Automation · Computer vision · Cleaning · Deep learning · Solar PV

1 Introduction

Renewable energy systems have seen a recent spurt in their installations globally owing to their abundance [4]. Solar energy systems harness the sun's thermal energy, which is copious, accessible, intermittent, and yet very economical. Photovoltaic

S. Nallapaneni · K. Shah · H. S. Dhiman (✉)

Department of Electrical Engineering, Adani Institute of Infrastructure Engineering, Ahmedabad, India

e-mail: hsdhiman1098@gmail.com; harsh.dhiman@sitpune.edu.in

S. Nallapaneni

e-mail: shreyan.ele18@aii.ac.in

K. Shah

e-mail: kairavishah.ele19@aii.ac.in

H. S. Dhiman

Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune, India

panels harness this thermal energy and convert it to electricity [7]. This is a unique form of photovoltaic solar energy system. Simply defined, solar panels are carefully arranged to gather solar energy and convert it to electrical energy that is sent into the grid. Another sort of solar energy system is the concentrated solar energy system, which consists of mirrors or lenses constructed in such a way that collected heat is concentrated to a single point. The photovoltaic (PV) system is the most popular type of solar energy system. The solar capacity of each country varies according to sun irradiation and available land [8]. This is a renewable energy alternative considering the fact that the energy source is the sun, which is a clean, renewable, abundant, and inexpensive source of energy. Solar photovoltaic farms can be built on the ground or on top of a building. Additionally, ground-based devices can be fixed arrays or multi-axis trackers. By allowing panels to track the sun's position throughout the day, axis trackers increase performance [1]. Solar panels transform captured thermal energy into direct current (DC) electricity. The inverter of the solar power plant becomes crucial when converting this direct current (DC) electricity to alternating current (AC) electricity.

Solar panels with an upward inclination are more prone to bird droppings and a build-up of general dust and grime that does not wash away with rain [10]. This reduces the amount of light striking the panel and lowering its output. Because solar panel manufacturers and installers base their energy forecasts on the optimal performance of clean solar panels. This accumulation of dirt might impair the panel's ability to meet those projections. As a result, it is vital to maintain clean solar panels in order to protect and preserve your investment [3].

This paper is organized as follows: Sect. 2 discusses the related works in solar module cleaning followed by Sect. 3 where we discuss the proposed methodology and Sect. 4 highlights the main conclusions.

2 Solar PV Cleaning Methodologies

Currently, there are a variety of cleaning methods available for photovoltaic modules. Module cleaning has existed for some time, ranging from manual to semi-automated. The advancement of artificial intelligence has resulted in an increase in the number of intelligent solutions for photovoltaic module cleaning [6]. The popular methodologies can be enumerated as follows:

- (a) **Non-automated Cleaning:** Automation is used sparingly for cleaning modules in this category. Almost all the activity is carried out by hand. The brush or cloth is primarily used to clean the modules. This type is typically used for small-scale roof top systems, such as commercial and residential, or for small-scale power plants in countries with cheap labor.
- (b) There is a significant amount of automation used to clean PV modules in this category, as well as a considerable manual input. The work is carried out either directly on the panels by the use of robots placed manually on the modules or

through the use of vehicles and cleaning apparatus available to clean the panels. Several of these varieties have existed for years. It can be further classified as follows:

- (i) **Robotic Cleaning System (Semi-automated):** Robots similar to those used in Fully Automated Cleaning may be used in this type of cleaning; however, in order to clean all the rows of a power plant, these bots must be manually shifted from row to row, or from one end of a tracker to the start of the next, or as the case may be. While there are numerous machine designs available in the market, there are typically no bridges connecting the rows and parking stations for robots.
- (ii) **Vehicle-driven Cleaning System:** This method of cleaning involves attaching the cleaning mechanism (typically a brush with controls) to a tractor or other compatible vehicle driven by a human. Each machine has its own safety system that regulates the pressure applied to the PV panels by the driver/operator's brush in order to avoid any damage to surface of the panel. A large turning radius is required at the ends of the row to allow the vehicle to easily turn around without wasting valuable cleaning time, which typically results in a larger land requirement than the former option.
- (c) **Fully Automated Cleaning:** This kind of cleaning system makes use of an Automatic Robotic Cleaning System (ARCS) to complete the work efficiently and with a high degree of reliability. Automation is often more advanced than the other two kinds. Cleaning is accomplished with the assistance of cleaning equipment known as robots or bots, which are permanently mounted on each row of the power plant. These robots travel along the panel edges, from one end of the module's row to the other, and vice versa. They are parked or docked on one side of each row at a docking station. Within the rows, such stations may also exist if they are longer than the robot's journey distance in one direction. The availability of bridges that fill up spaces enables robots to move from one array or tracker to another. The robots identify Return Stations set at predetermined locations and return to their docking stations. If engineered and produced appropriately, this form of cleaning should normally not require physical work to operate the cleaning process or place the devices. These machines can be operated at any time of day or night, although it is recommended that they be operated around sunset to achieve maximum cleaning efficiency in areas with heavy wetness. Additionally, they can be configured to operate on demand via a remote connection. As with any other machine or piece of mechanical equipment, the ARCS will require human intervention for its maintenance.
- (d) **Self-cleaning Glass:** This is a specific sort of glass that is available in the market that does not collect dust. This glass is exceptionally hydrophobic or hydrophilic, which means that if moisture comes into contact, it easily rolls off. While rolling away the wetness, it also sweeps away the dust, and even bird droppings that land on the glass surface are not deposited on it, particularly if it is at a certain angle.

According to studies, dust collected on the surface of the photovoltaic module results in a daily energy loss of approximately 4.4% over the course of a year [5, 11]. A decrease in the voltage and output power of the photovoltaic module is observed when dust particles are deposited on it, depending on the mass accumulated and the type of pollutant. Additionally, a greater reduction happens when the temperature of the photovoltaic module is increased. Furthermore, maintaining clean and cool PV modules results in inefficient system performance. Solar panels need to be cleaned on a biweekly basis to maintain peak efficiency, which is particularly difficult to achieve with large solar panel arrays. Cleaning dust particles from solar panels is a significant issue because it is a time-consuming procedure that needs a significant amount of people and money. The present human-based cleaning methods for solar panels are inefficient in terms of time, water, and energy consumption. This is because solar panel cleaning must be performed regularly, making the process more difficult and pricey. This is the primary reason for the need to design and deploy automatic solutions, such as an automatic cleaning equipment capable of cleaning and moving effortlessly across the glass surface of the panels.

3 Proposed Methodology

Solar Soiling that is defined as energy loss that owes to dust on panels can reduce energy output by 7–50%. Investors and individuals who rely on solar power may face substantial difficulties as a result of these energy decreases. Cleaning drone comes at the key moment when more engineering solutions are needed to keep up with the rapid increase of solar energy. As a result, using an autonomous drone in conjunction with artificial intelligence and computer vision will help to boost productivity by preventing soiling.

To establish the source of the efficiency reduction, the characteristics of the solar cell must be evaluated on a regular basis. Each panel module's efficiency will be monitored in real time, and the results will be sent to a computer system that will control the entire array of solar cells through a centralized system. As environmental conditions are not always optimal, the data is averaged over a period of time. Calculating the ratio of the solar cell's output energy to the energy received from the sun can be used to assess the cell's efficiency. As a result, in order to evaluate the solar cell's performance, efficiency must be carefully and precisely measured. The intensity of solar radiation changes throughout the year. It is constantly changing and very reliant on weather conditions. The intensity of the radiation, for example, is greatest in the summer and lowest in the winter. This pattern varies per geographical region as well. The efficiency threshold is set in reaction to climate change [2], keeping this in mind and in order to grow and preserve the efficiency of the panels.

When the efficiency of a given cell goes below the target value, the control mechanisms are automatically triggered. In order to find abnormalities, the efficiency of one cell is compared to that of others. The settling of various particles causes a continuous loss of efficiency. These particles tend to gather if left unchecked, resulting

in a considerable reduction in overall efficiency. The efficiency is affected differently by each particle. The mass, texture, color, and other features of each particle influence the efficiency decline. Each solar panel's efficiency is calculated continuously in proportion to its output voltage and current, and the results are sent to a centralized system. Because efficiency is directly tied to meteorological conditions, the average of data collected over a specific period is used. The ground subsystem activates the drone to fulfill its tasks and if the drone's efficiency falls below a specified level. A Human Machine Interface (HMI) panel is installed on a ground subsystem to display abnormalities and battery status and allow the operator to take control of operations in the event of an emergency. The drone's camera gets an image of the ineffective cell, which is then analyzed using various vision techniques. These methods are used to find the abnormalities' coordinates and then control them using the drone's end effectors. The following hardware will be required for cleaning module:

1. Dual Camera Setup (Thermal and RGB Camera): View thermal and high-definition color photos while in flight. In low light and gloomy circumstances, it is ideal for inspecting, locating, and identifying subjects.
 2. LIDAR Sensor: Light Detection and Ranging (LiDAR) sensors scan the ground and measure changing distances using light energy generated by a laser. The result is a large collection of elevation data that may be used to create high-resolution maps and 3D models.
 3. Nozzle: A nozzle provides a good spraying performance, as it can spray smaller droplets of cleaning fluid more uniformly.
 4. Cleaning Arm Extender: In case of requirement, this arm will extend and clean the solar panel using microfibers attached to it.
 5. Plan the flight's path automatically using a flight planning algorithm and feed it to the drone.
 - (a) GPS sensors are devices with antennas that employ a satellite-based navigation system to deliver location, velocity, and time information using satellites orbiting the earth.
 - (b) Triangulation and marking of the coordinates for the survey on the images.
 - (c) Strategic coordinates are marked using visible markers with Ground control points.

Drone flight path software develops mission plans automatically based on the drone operator's flying parameters, including start and finish times, desired speed, altitude, and position. With this hardware equipments, it is possible to create a fully automatic solar cleaning drone.
 6. Calculate the efficiency of the solar panels: By comparing the output of each solar panel to the input it receives, the efficiency of each panel can be monitored separately, and if the efficiency falls below a particular level, the drone is triggered to carry out its operations by the ground subsystem.
- In addition, the drone is equipped with a thermal camera, which aids in the detection of hot spots, cracks, and momentary shadowing problems on the panel caused by dust and bird droppings. The temperature difference between the cells of the

panels is so exact that defects can be detected early. Panel fault coordinates are provided through the connection between the thermal camera and the ground subsystem, enabled for panel cleansing. The battery life status is also displayed on the ground subsystem's HMI, and when the battery runs out, the operator is notified. The saved images are processed using image processing algorithm on the server.

The drone that flies over the solar panels gets an image of the solar panels and their cells using the camera on the top of the drone. The drone is activated and advanced to the specific cell when the efficiency of the solar cells falls below the minimum value of the threshold. It reads the image from the overhead camera and analyzes it using computer vision algorithms. Considering a pattern-matching method for image processing is a part of the procedure. Pattern matching finds the parts of a gray-scale image that match a known reference sequence, which is presented as a reference model. To guarantee that the desired output model matches the reference model, a model template is developed. To account for this, an image of the solar cells on the panel is captured during panel setup and used as a template for pattern matching. The cleaning mechanism is activated if the pattern-matching score falls below a predetermined threshold value. Figure 1 illustrates the path planning and movement of the drone on the solar panel area.

7. Cleaning action is initiated: The cleaning fluid is sprayed over the selected area of the solar module using the nozzle attached to the drone. Most of the time, a simple water spray will suffice to clean a dirty module. In the event of bird droppings or other similar issues, a robotic arm extension is provided, which will extend itself out of the drone and reach up to the solar panel, cleaning the dirty area.
8. Return to charging dock: The quadcopter is docked near the solar panels in a waterproof docking station. It leaves its station, takes off, and heads straight for the panels. There, it will align itself over each panel, spraying it with cleaning fluid, using the advanced technology already described. The quadcopter will return to its docking station once the cleaning is completed, where it will be safely stationed again. A robotic system will then refill its cleaning fluid tank and replace its battery with a fully charged battery. When the time comes again, the drone will be ready for a further cleaning session (Fig. 2).

Over the span of time, the total cost of operation and maintenance has a significant impact on solar system rates and efficiency [9]. Given the current booming environment of automation, these technologies could be a boon in resolving the issue of high-cost O&M and particularly solar panel cleaning. Nowadays, maintenance is handled by engaging human labor, which is both costly and sometimes dangerous to the personnel. However, here are a few possible challenges for deploying and maintaining automatic cleaning modules which can be summarized as:

1. All robotic cleaning solutions are investment heavy.
2. For each array, a different robotic cleaning module must be purchased and installed. Since the land acquired for solar power generation is not always rectangular, this is a major challenge.
3. Windy atmosphere can hamper the stability of drone.

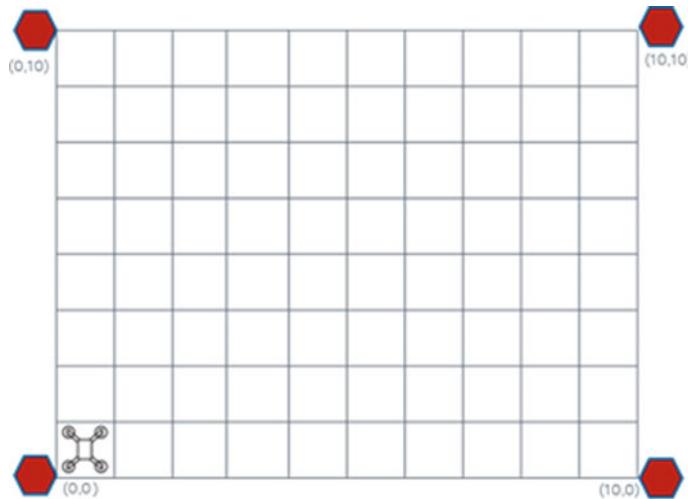


Fig. 1 Sample coordinates for the drone

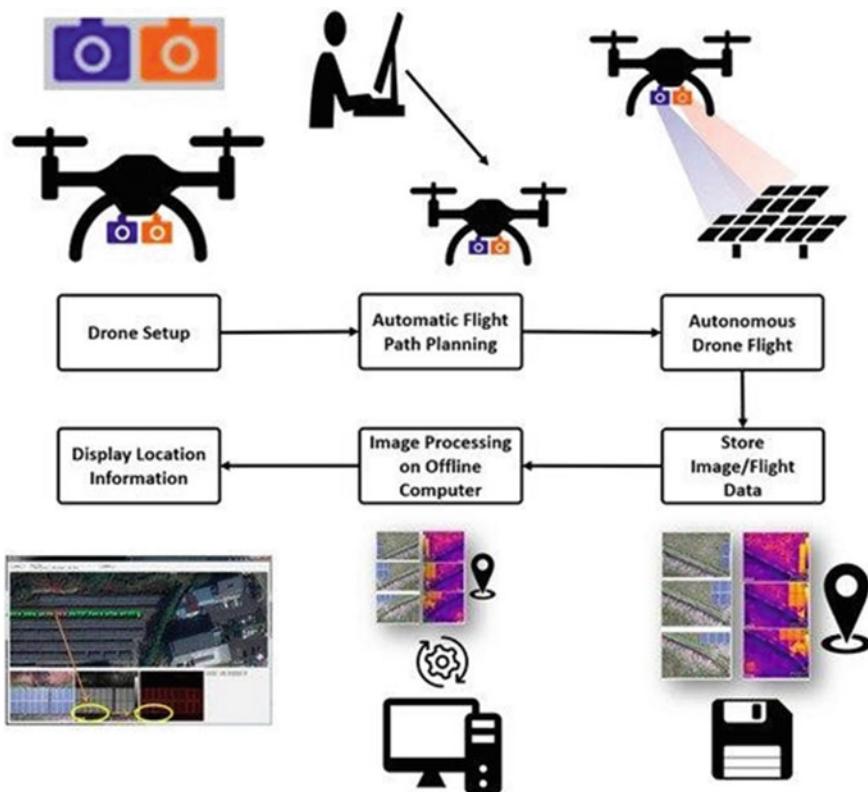


Fig. 2 Flowchart of the proposed cleaning framework

4. Since the dry robotic cleaning module directly rubs the cleaner with the solar panel, it might end up scratching the panel in a longer run.
5. Most of the solar PV plants are located in desert land, far away from residential area, procuring water for wet cleaning can be a big challenge.
6. To maintain the robotic cleaning module, an experienced human resource will be required.

4 Conclusion

Investment in solar PV industry has seen a recent spurt. With increasing installations globally, the operation and maintenance (O&M) cost of the panels must be controlled. Solar panel cleaning process ensures that the solar cells are maintained throughout, which is normally a tough task. By integrating machine vision and artificial intelligence tools, the cleaning solution is expected to improve the overall efficiency of the solar PV system. As a result, successfully utilizing this resource while maintaining a high level of efficiency is crucial. Adapting autonomous drones to examine and clean solar cells can help save money and time in hazardous environments. Furthermore, the proposed system may work credibly regardless of external conditions, resulting in greater accuracy. Following that, the quality and quantity of data acquired considerably improves the inspection's efficiency. It also entails developing the planned work as a real-time system and focusing on identifying qualities that outperform existing technical systems in terms of efficiency.

References

1. Alvarez DL, Al-Sumaiti AS, Rivera SR (2020) Estimation of an optimal PV panel cleaning strategy based on both annual radiation profile and module degradation. *IEEE Access* 8:63832–63839. <https://doi.org/10.1109/ACCESS.2020.2983322>
2. Azouzoute A, Zitouni H, Ydrissi ME, Hajjaj C, Garoum M, Bennouna EG, Ghennoui A (2021) Developing a cleaning strategy for hybrid solar plants PV/CSP: case study for semi-arid climate. *Energy* 228:120565. <https://doi.org/10.1016/j.energy.2021.120565>
3. Deb D, Brahmabhatt NL (2018) Review of yield increase of solar panels through soiling prevention, and a proposed water-free automated cleaning solution. *Renew Sustain Energy Rev* 82:3306–3313. <https://doi.org/10.1016/j.rser.2017.10.014>
4. Dhiman HS, Deb D, Guerrero JM (2019) Hybrid machine intelligent SVR variants for wind forecasting and ramp events. *Renew Sustain Energy Rev* 108:369–379
5. Ekinci F, Yavuzdeger A, Nazlıgül H, Esenboğa B, Mert BD, Demirdelen T (2022) Experimental investigation on solar PV panel dust cleaning with solution method. *Solar Energy* 237:1–10. <https://doi.org/10.1016/j.solener.2022.03.066>
6. Enaganti PK, Dwivedi PK, Sudha R, Srivastava AK, Goel S (2020) Underwater characterization and monitoring of amorphous and monocrystalline solar cells in diverse water settings. *IEEE Sens. J.* 20(5):2730–2737. <https://doi.org/10.1109/JSEN.2019.2952428>
7. Jones RK, Bara A, Saeeri AA, Al Qahtani A, Al Amoudi AO, Al Shaya Y, Alodan M, Al-Hsaien SA (2016) Optimized cleaning cost and schedule based on observed soiling conditions

- for photovoltaic plants in central Saudi Arabia. *IEEE J. Photovoltaics* 6(3):730–738. <https://doi.org/10.1109/JPHOTOV.2016.2535308>
- 8. Kaldellis J, Zafirakis D (2012) Experimental investigation of the optimum photovoltaic panels' tilt angle during the summer period. *Energy* 38(1):305–314. <https://doi.org/10.1016/j.energy.2011.11.058>
 - 9. Khan MU, Abbas M, Khan MM, Kousar A, Alam M, Massoud Y, Jafri SHM (2021) Modeling and design of low-cost automatic self cleaning mechanism for standalone micro PV systems. *Sustain Energy Technol Assess* 43:100922. <https://doi.org/10.1016/j.seta.2020.100922>
 - 10. Micheli L, Fernández EF, Muller M, Almonacid F (2020) Extracting and generating PV soiling profiles for analysis, forecasting, and cleaning optimization. *IEEE J Photovoltaics* 10(1):197–205. <https://doi.org/10.1109/JPHOTOV.2019.2943706>
 - 11. Salamat T, Ramahi A, Alamara K, Juaidi A, Abdallah R, Abdelkareem MA, Amer EC, Olabi AG (2022) Effect of dust and methods of cleaning on the performance of solar PV module for different climate regions: comprehensive review. *Sci Total Environ* 827:154050. <https://doi.org/10.1016/j.scitotenv.2022.154050>

An Ensemble Framework for Glaucoma Classification Using Fundus Images



Achirangshu Patra , Arijit Nandi , Mayaluri Zefree Lazarus , and Satyabrata Lenka 

Abstract Glaucoma is an irreversible progressive vision condition that can lead to permanent sightlessness. With null early-stage symptoms, it is critical to prevent in advanced stages of glaucoma. Artificial intelligence has shown significant escalation, in many fields, especially in medical image diagnosis, wherein highly accurate automated disease diagnosis of a large dataset in less time has now become feasible. Researchers have utilized many machine learning and deep learning approaches for glaucoma detection from retinal fundus images, but their performance varies and it depends on the input dataset. According to the “No Free Lunch theorem”, a single classifier is not suitable for classifying all datasets. So, to overcome this situation, we have proposed an ensemble learning approach that utilizes different machine learning approaches for Glaucoma detection. A benefit of using ensemble learning is to improve the average prediction performance by combining predictions from multiple models. In this paper, we developed a machine learning model ensemble approach [Majority-Voting-Ensemble (MVE)] consisting of a Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP) classifiers. Our approach involves the Histogram of Oriented Gradients (HOG) technique to collect efficient features for glaucoma detection. The efficiency of our proposed approach is evaluated using two

A. Patra

Department of Mechatronics Engineering, C.V. Raman Global University,
Bhubaneswar, Odisha, India

A. Nandi 

Department of Computer Science, Universitat Politècnica de Catalunya (BarcelonaTech),
Barcelona, Spain

e-mail: [jит.ari172@gmail.com](mailto:jit.ari172@gmail.com)

Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain

M. Z. Lazarus · S. Lenka

Department of Electrical and Electronics Engineering, C.V. Raman Global University,
Bhubaneswar, Odisha, India

e-mail: zefree.lazarus@cgu-odisha.ac.in

S. Lenka

e-mail: 21080013@cgu-odisha.ac.in

popularly used benchmark datasets, called ORIGA and REFUGE. The results show that the proposed ensemble approach is capable of outperforming the base classifiers for glaucoma classification from fundus images.

Keywords Glaucoma classification · Image processing · Machine learning · Ensemble learning · Majority voting ensemble

1 Introduction

Glaucoma [1] is defined as optic neuropathy due to the escalation of intraocular pressure (IOP) characterized by loss of sight. Glaucoma is the second most widely recognized reason for blindness among eye diseases [2]. It is known as a silent killer of sight because one does not even aware that they have started to lose vision, as it starts with loss of peripheral vision. Since sight loss from glaucoma cannot be regained, so glaucoma screening or detection is needed in the early stages to preserve eye health for a better quality of life [3]. However, the clinical diagnosis by ophthalmologists is time-consuming and costly, and not worthy of mass detection. Figure 1 shows the retinal fundus image for glaucoma positive and glaucoma negative cases.

Clinically available eye examinations practiced to determine Glaucoma are (1) Tonometry-Intraocular pressure (IOP) measurement. (2) Ophthalmoscopy (funduscopy)—The color and shape of the retina, optic nerve-based examination by seeing inside of the eye. (3) Perimetry (function-based visual field test)—The complete systematic way of measuring the field of vision function. (4) Gonioscopy determine the angle where the iris meets the cornea. (5) Pachymetry test to measure the thickness of the cornea. Clinical assessment depends on the availability of specialized perimetric apparatus which is not normally present in every healthcare. Manual assessment is a skill-based costly and time-consuming process. Retinal images contain essential information about the health of an eye. Thus, it is required to develop an automated process for detecting symptoms of glaucoma from an image. In this paper, we have proposed a Majority voting ensemble method to classify glaucoma based on the histogram of gradients feature extraction process. The main contributions: the proposed approach consists of three machine learning models that are independently tested, combining their results by the majority voting ensemble method to improve advantage of our ensemble method is that it can be easily accompanied by other machine learning models, which will increase the adaptiveness in different scenarios.

2 Related Work

Several works have been proposed, and many deep learning and machine learning techniques have also been explored for glaucoma detection. The authors of [4] devel-

oped an automated method for glaucoma detection using a support vector machine classifier. Several popular preprocessing techniques like Contrast-Limited Adaptive Histogram Equalization (CLAHE) [5], Illumination Correction and Intensity adjustment [6], noise removal, and contrast adjustment are applied before classification to get better efficiency. In [7] four machine learning algorithms: C5.0, RF, SVM, and KNN are analyzed to make a learning model with aim of glaucoma detection. In method [8] they employed an SVM-based approach for glaucomatous classification using retinal fundus images. For feature extraction, a 2-dimensional variational mode decomposition tool has been used. Some papers applied some classic machine learning models to predict glaucoma. This [9] paper evaluates the performance of the Random Forest, support vector machine, XGboost, and C5.0—prediction models, by using a tenfold cross-validation method. Among these four models, the XGboost model stands best in terms of accuracy, sensitivity, specificity, and AUC. In [10] this paper researchers integrated the performance of support vector machine (SVM), random forest (RF), K-nearest neighbors (KNN), and logistic regression. In [11], authors have developed a novel statistical feature extraction method also has been applied to fundus retinal images to collect 12 different features. A novel deep learning model has been proposed for retinal blood vessel semantic segmentation process which helps to diagnose various ophthalmologic diseases.

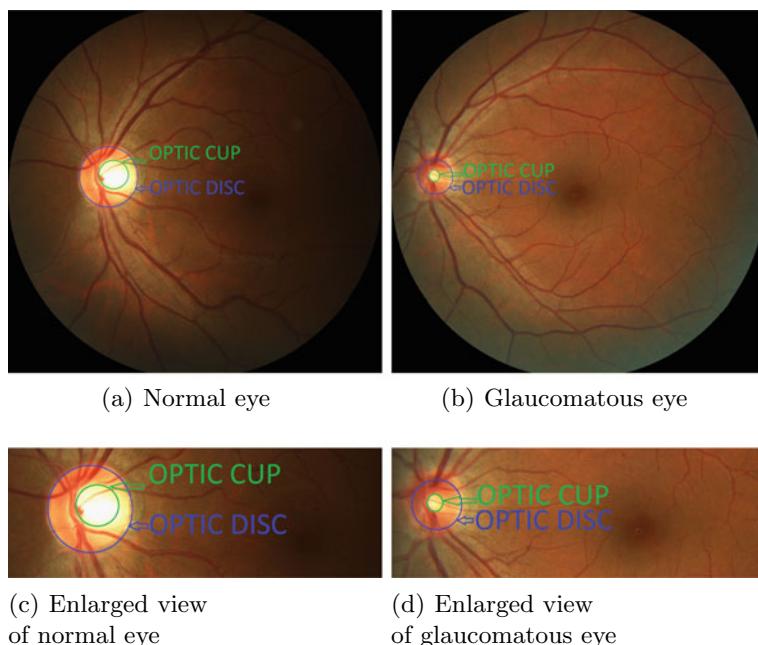


Fig. 1 Retinal Fundus Glaucoma Challenge (REFUGE) image glaucoma/non-glaucoma. The area surrounded by the blue circle is the optic disc (OD); the central bright zone surrounded by the green circle is the optic cup (OC); right: increased cup caused by glaucoma, and the area between them is called the neuroretinal rim

3 Preliminaries

3.1 Feature Extraction

Features are key points, patterns, characteristics, or main facts about the content of an image that helps to identify itself. To make the task of classification made easy feature extraction provides an important role by obtaining the most relevant information from the image and representing that information in a lower dimensionality space.

The efficiency of machine learning techniques depends on the quality of the feature selected so it is very important to choose a proper method. We have chosen HOG as it doesn't provide region information rather it focuses on the global features of an image. In particular, this type of extraction process is applied to extract the most important features to make a simplified representation of the image. It is better than any feature extraction process as it uses edge detection by gradient calculation and histograms of gradients, with magnitudes as weights. In addition to optic disc to cup ratio comparison, the local retinal nerve fiber thinning can be an important parameter for detecting glaucoma [12]. Also, analysis of macular ganglion cell loss gives additional information about the glaucomatous eye [12]. Our model consists of two levels [13], (A) feature extraction, (B) glaucoma classification. In the proposed method Histogram of Oriented Gradients has been used to extract features from the Retinal Fundus Glaucoma images (REFUGE). From Fig. 2 we can visualize the HOG features extraction process.

We are using OpenCV to read pixel-level information from the raw image. Resize the image into an exact likeness of 128×64 pixels. This dimension was used in

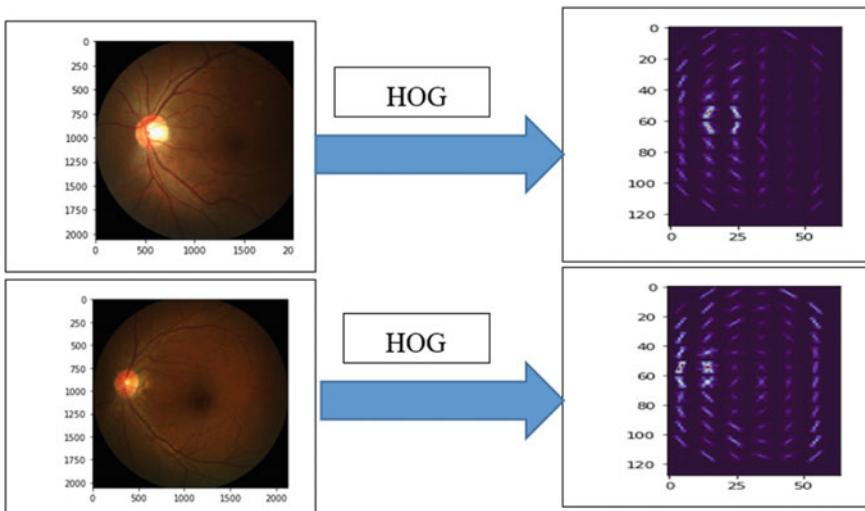


Fig. 2 Visualization of HOG features

the paper and was urged by the author as their main goal with this model was to acquire improved results on the task of pedestrian discovery or detection. After Computing the hog features, we have applied Principal Components Analysis (PCA) for dimensionality reduction. We are not only considering the cup to disc ratio, there are some other important features such as local retinal nerve fiber thinning, ganglion cell loss in the macular region, for detecting glaucoma. Because of glaucoma, there are effects in some other regions of our eye, so it's necessary to select features from the entire retinal fundus image; hence HOG is best suitable for that. We have chosen a hundred features to be extracted with HOG for this experiment.

3.2 Classifier

Support Vector Machine We have used a support vector machine (SVM) which is a supervised learning model, for classifying the glaucomatous eye from the non-glaucomatous eye, one of the most well-known machine learning algorithms. We have chosen an SVM classifier as this algorithm is formulated for binary classification problems and gives the best result.

Random Forest A non-parametric classifier, meta estimator generates a class prediction and the class accompanying the maximum votes enhances the final forecasting of the proposed model. This is another classifier of our proposed ensemble model. As we are trying to make a lightweight machine learning model RF classifier is the best fit for that.

Multilayer Perceptron Multilayer perceptron (MLP) is a supplement of a feed-forward neural network where the mapping between inputs and outputs is non-linear. This classifier model is composed of an input layer to receive the signal and propagate it to the output layer, an output layer that makes a resolution or prediction based on the given dataset, and in between those two, an arbitrary number of hidden layers are present. After getting the output error must be calculated to minimize it by the process of backpropagation and update the model.

Ensemble Learning The intention behind ensemble learning [8] is to integrate multiple classifiers efficiently, that can be used for improving prediction performance or the decision-making process. Figure 3 illustrates the overall flowchart of the ensemble learning process.

– *Majority voting ensemble*: Due to the large diversity of data, the ensemble model is used to reduce the generalization error of the prediction. It gives better performance compared to any single model used for glaucoma classification as it works by combining the prediction of Multilayer perceptron (MLP), Support vector machine (SVM), and Random Forest classifier. Hard voting is a simple form of voting by the majority [14].

We predict the class level Y by plurality voting of n number of classifier set $Q = \{C_1, C_2, \dots, C_n\}$. Class label $L = \{L_1, L_2, \dots, L_m\}$. For each instance, let consider,

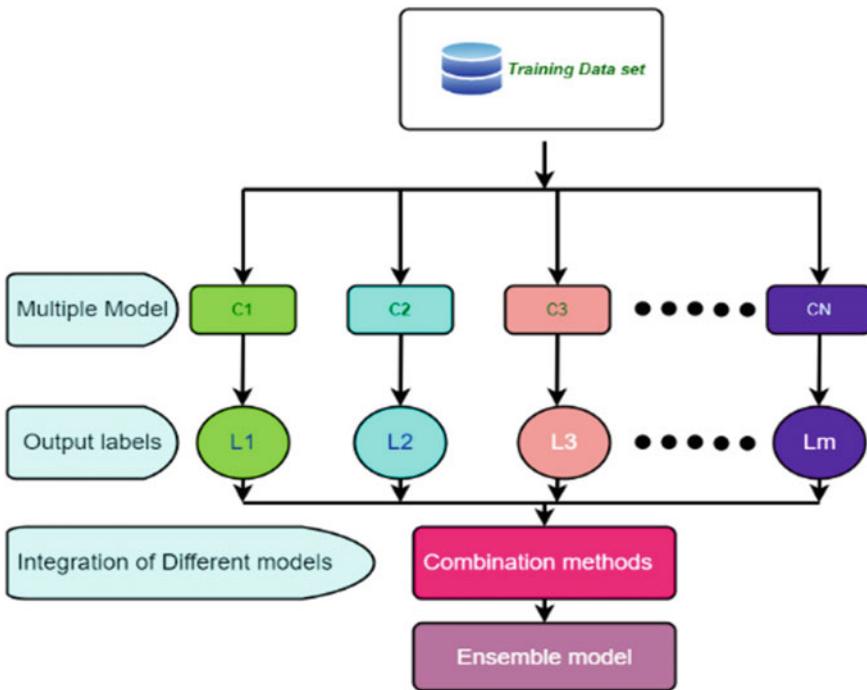


Fig. 3 Ensemble learning

$$X^{L_p}(C_j(i)) = \begin{cases} 1 & \text{if } C_j(i) = L_p \text{ where } L_p \text{ belongs to set } L \\ 0 & \text{if } C_j(i) = L_q \text{ where } L_q \text{ belongs to set } L \text{ and } p \neq q \end{cases} \quad (1)$$

After combining the prediction of M number of classifiers with majority voting. The ensemble prediction is

$$\hat{Y} = \operatorname{argmax} \sum_{j=1}^M X^{L_p}(C_j(i)) \quad (2)$$

4 Proposed Approach

For this experiment, we have selected three classification models for the ensemble model (CF_1, CF_2, CF_3) and (P_1, P_2, P_3) are the predicted outputs of the selected classifiers. The final predicted class level \hat{Y}_{predict} is generated after applying the

majority voting ensemble model. The pseudo code of our proposed MVE is presented below:

Algorithm 1 Pseudo-code of our approach MVE

Result: Final class prediction \rightarrow Glaucoma/Non-Glaucoma

1. Input: ORIGA and REFUGE Dataset (retinal fundus images);

2. Feature extraction: HOG features from images ;

3. Dataset prep: K-fold cross-validation (K=7) ;

4. Initialization: Classification models MLP, SVM and RF;

4. Prediction result matrix: $P[|folds|]$;

Part 1: Model fitting ;

for $i \leftarrow 1$ **to** $|af|$ **do**

for $j \leftarrow 1$ **to** $|folds|$ **do**

 1. Train each base classifier ;

 2. Test each classifier ;

end

end

Part 2: Ensembling (Majority voting);

1. Final class prediction ($\hat{Y}_{predict}$)= $majorityvote(P)$;

2. Test ensemble model.

5 Experimental Results

5.1 Dataset Description

In this experiment, two open-source datasets are used as input. The model is trained and tested with these datasets. The analyses of the data sets are noticed and mentioned in Table 1.

The ORIGA retinal images were collected by Singapore Malay Eye Study (SiMES) over a period of three years from 2004 to 2007. They have made this online depository intending to assist researchers in developing their imaging technology and developing new glaucoma mass testing tools. Retinal Fundus Glaucoma Challenge—To overcome the issue of the limited size of the available data set this challenge was held with MICCAI 2018. The challenge included two main goals, namely optic disc/cup segmentation and separating the glaucomatous eye images from the healthy eye images.

From the dataset description (in Table 1), we can see that REFUGE and ORIGA have class imbalance problems. Classification with an imbalanced dataset is challenging as the base model lacks training in minority class labels. So there is a high possibility that the decision of the classifier can be biased and that the model's performance can be hampered while testing. To resolve the class imbalance in the REFUGE and ORIGA datasets we have used a popular statistical method called

Synthetic Minority Oversampling Technique (SMOTE) [15] to increase the number of minority class labels (i.e., Glaucoma class labels) to match the number of samples in the majority class.

5.2 Experimental Setup

The entire program was developed and tested on Ubuntu 18.04.6, 64bit OS (i5-10300H processor, CPU@2.50GHz with RAM 8Gb). The code was written in Python 3.6.13 and the packages used to develop this model are TensorFlow 1.8.0, Keras 2.0.1, and Scikit-learn-1.1.2. The developed code can be found on Github: <https://github.com/officialarijit/Glaucoma-ML-DL>.

5.3 Performance Metric

The accuracy, sensitivity, precision, *F*1-score, Positive predictive Value, Negative Predictive Value, and others have been widely used to understand the utility of any machine learning model. The following attributes are essential to understanding their utility of them:

- **Accuracy:** The percentage of predictions refers to the ratio of correct predictions to total predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

- **Sensitivity/Recall:** It is the true positive rate (TPR). It refers to the strength and ability of the model to correctly recognize those cases accompanying Glaucoma:

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR} \text{ (False Negative Rate)} \quad (4)$$

- **Precision:** It is a ratio of true positives to total predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

Table 1 Detailed information about the two datasets used for this research experiment

Dataset	Total no. of images	Glaucoma	Non-glaucoma
ORIGA [9]	650	168	482
REFUGE [10]	400	40	360

- **F1-Score:** The $F1$ -score is defined as the harmonic mean of precision and recall.

$$F1\text{-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

- **Positive Predictive Value (PPV):** is a parameter that represents the proportion of positive tests among the total positive.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

- **Negative Predictive Value (NPV):** means the proportion of negative tests that are truly negative.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (8)$$

- **LR^+ :** The positive likelihood ratio means the probability that a patient is diagnosed with glaucoma divided by the probability of a positive result in patients without glaucoma.

$$LR^+ = \frac{\text{sensitivity}}{1 - \text{specificity}} \quad (9)$$

- **LR^- :** The negative likelihood ratio means the probability that a person with glaucoma tested negative divided by the probability that a person without glaucoma tested negative.

$$LR^- = \frac{1 - \text{sensitivity}}{\text{specificity}} \quad (10)$$

where specificity = $\frac{\text{TN}}{\text{TN} + \text{FP}}$.

- **AUC Score (Area Under Receiver Operating Characteristic (ROC) Curve):** is used to evaluate the performance of the model over its total operating range. It is used to determine the degree of separability for the model.

Here, True Positive (TP): the patient has been diagnosed with glaucoma and the prediction is positive; True Negative (TN): the patient does not have glaucoma disease and the prediction is negative; False Positive (FP): the patient does not have glaucoma disease but the prediction is positive; False Negative (FN): the patient has been diagnosed with glaucoma but the prediction is negative. In this evaluation process, we have used the K -fold cross-validation approach, where K is 7.

6 Results and Discussion

Here in this section, we have presented our experimental results and discussions for RFUGE and ORIGA datasets.

6.1 REFUGE Dataset

In Tables 2, 3, 4 and 5 the performance of MLP, RF, SVM and our proposed MVE are presented, respectively, for the REFUGE dataset.

The following mean accuracy has obtained from Tables 2, 3, 4 and 5: 97.91% (MLP), 99.86% (SVM), 99.16% (RF), and 99.86% (MVE), respectively. From the mean accuracy value comparison, we can see that the ensemble model gives 99.86%

Table 2 Performance metrics of MLP on REFUGE dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.9803	0.9791	0.9791	0.9791	1	0.9605	–	0.0416	0.9791
Max	1	1	1	1	1	0.9272	–	0.0784	1
Min	0.9639	0.9611	0.961	0.9611	1	0.9272	–	0	0.9607
Std	0.0121	0.0131	0.0131	0.0131	0	0.0242	–	0.0263	0.0131

Table 3 Performance metrics of SVM on REFUGE dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.9986	0.9986	0.9986	0.9986	1	0.9973	–	0.0028	0.9985
Max	1	1	1	1	1	1	–	0.0196	1
Min	0.9904	0.9902	0.9902	0.9902	1	0.9811	–	0	0.9901
Std	0.0033	0.0033	0.0033	0.0033	0	0.0066	–	0.0068	0.0034

Table 4 Performance metrics of RF on REFUGE dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.9917	0.9916	0.9916	0.9916	0.9916	0.9917	–	0.0086	0.9916
Max	1	1	1	1	1	1	–	0.0407	1
Min	0.9611	0.9611	0.9611	0.9611	0.9607	0.9615	24.9803	0	0.9611
Std	0.0131	0.0131	0.0131	0.0131	0.0142	0.014	–	0.0147	0.0131

Table 5 Performance metrics of MVE on REFUGE dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.9986	0.9986	0.9986	0.9986	1	0.9973	–	0.0028	0.9985
Max	1	1	1	1	1	1	–	0.0196	1
Min	0.9904	0.9902	0.9902	0.9902	1	0.9811	–	0	0.9901
Std	0.0033	0.0033	0.0033	0.0033	0	0.0066	–	0.0068	0.0034

which is higher than the considered base classifier (0.7% higher than RF, 1.95% higher than MLP). As the change in accuracy is higher compared to the change in standard deviation so we have prioritized the accuracy as a marker for performance analysis of our proposed MVE. As accuracy is not enough to evaluate the effectiveness of our proposed model for glaucoma detection in REFUGE dataset, we considered *F1*-score (harmonic mean of precision and recall) to be an important marker to declare the better performing model. A classifier with higher *F1* value implies a model which can almost perfectly classify the glaucomatous images from retinal fundus images. In Table 5, the *F1* value (0.9986) tends to be 1 and the value is higher than other base models RF and MLP. At the same time, the AUC score of our proposed MVE is 0.9985 close to 1, which is also higher than the base models, implies better performing classifier in case glaucoma detection. So, from the comparison, our proposed MVE approach has outperformed all the considered base classifiers for glaucoma classification from retinal fundus images in REFUGE dataset. Hence, our proposed MVE approach can be declared as an effective and better classifier.

6.2 ORIGA Dataset

In Tables 6, 7, 8 and 9 the performance of MLP, RF, SVM and our proposed MVE are presented, respectively, for the ORIGA dataset.

In the case of accuracy comparison from Table 6 for MLP, Table 7 for SVM, Table 8 for RF and finally Table 9 for our proposed MVE approach in ORIGA dataset, we can see that our proposed MVE approach is 0.21% higher than RF, 0.29% higher

Table 6 Performance metrics of MLP on ORIGA dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.8713	0.8651	0.8645	0.8651	0.9142	0.8284	18.3174	0.2095	0.8651
Max	0.9219	0.913	0.9124	0.913	0.983	0.863	58	0.3333	0.913
Min	0.8194	0.8043	0.802	0.8043	0.8636	0.75	6.42	0.1564	0.8043
Std	0.0327	0.0334	0.0339	0.0334	0.0419	0.0956	17.6184	0.0552	0.0335

Table 7 Performance metrics of SVM on ORIGA dataset

	Precision	Recall	<i>F1</i>	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.5869	0.8558	0.8557	0.8558	0.8548	0.8589	6.2482	0.1662	0.8558
Max	0.8862	0.8832	0.883	0.8832	0.8923	0.92066	8.2857	0.8747	0.8695
Min	0.7835	0.7826	0.7824	0.7826	0.7671	0.8	3.2941	0.1189	0.7826
Std	0.0315	0.0313	0.0314	0.0313	0.0377	0.0365	1.4177	0.0493	0.0314

Table 8 Performance metrics of RF on ORIGA dataset

	Precision	Recall	<i>F</i> 1	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.8855	0.8827	0.8825	0.8827	0.589	0.912	6.4526	0.0979	0.8828
Max	0.9246	0.9202	0.92	0.9202	0.909	0.9677	10	0.1428	0.9202
Min	0.8562	0.855	0.8547	0.8552	0.8266	0.875	4.7692	0.0333	0.855
Std	0.0247	0.0242	0.0243	0.0242	0.0293	0.0335	1.763	0.0398	0.02419

Table 9 Performance metrics of MVE on ORIGA dataset

	Precision	Recall	<i>F</i> 1	Accuracy	PPV	NPV	LR ⁺	LR ⁻	AUC score
Mean	0.8861	0.8848	0.8847	0.8848	0.8971	0.875	9.7987	0.1461	0.8849
Max	0.927	0.927	0.927	0.927	0.9393	0.9264	12.6144	0.2419	0.927
Min	0.8271	0.826	0.8259	0.826	0.8461	0.8051	5.5	0.0782	0.826
Std	0.0354	0.0363	0.0364	0.0363	0.0328	0.0455	3.4606	0.0618	0.03633

than SVM, and 1.97% higher than MLP. A model having high recall value is better and effective for an output-sensitive case like glaucoma detection. The recall value (presented in Table 9) of our proposed MVE is 0.8848 which is higher compared to other base classifiers (MLP, SVM and RF). As, the *F*1-score represents the harmonic mean of precision and recall. So a model with higher *F*1-score can be declared as a better performing classifier. From the *F*1-score comparison presented in Tables 6, 7, 8 and 9, we can declare that our proposed MVE is efficient in the glaucoma detection from retinal images in ORIGA dataset. Also, from the PPV comparison value, our proposed model has achieved 0.8971 which is higher than the SVM (0.8548), and RF (0.8590). Hence it indicates that our proposed MVE is capable of detecting the patients with glaucoma-positive at a much higher perfection that the base models. In case of LR⁻ value comparison our proposed MVE is 0.1461 which is less than 0.1662 (SVM), 0.2095 (MLP), and 0.0979 (RF), that indicates our proposed MVE is having less false negative in glaucoma detection, makes our model model reliable and effective. Also, the AUC value of our ensemble model is higher than the other three base models. Hence from all the comparison we can conclude that our proposed MVE is better than the considered base models in glaucoma classification from the ORIGA dataset.

6.3 State-of-the-Art Comparison

Here we have compared our proposed approach with the state-of-the-art (SOA) approaches from previous literature. The previous literature selection is based on

Table 10 State-of-the-art comparison for glaucoma classification

SOA literature	Year	Method	Dataset			ORIGA		
			REFUGE	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
[16]	2020	Convolutional neural network (CNN)	–	–	–	78.32	58.06	92.44
[18]	2021	EfficientNetB7	99.2	98	97	–	–	–
[17]	2020	CNN	90	96	84	–	–	–
[19]	2020	GBDT + ADASYN	–	–	–	84.3	89.4	79.3
Our proposed	2022	Ensemble approach MVE	99.86	99.86	100	88.48	88.48	91.50

(1) used similar benchmark dataset (REFUGE and ORIGA) for model development and testing; (2) the main objective is to screen Glaucoma from fundus retinal images. In [16] paper an 18-layer convolutional neural network has been designed and trained for glaucoma detection. In comparison, our model gives 10.16% higher accuracy, and 30.42% higher sensitivity. Mean values are considered for comparison whereas the type of value was not mentioned in the existing state of the art. The performance parameter of the CNN architecture [17] and EfficientNetB7 [18] give lower values compared to our proposed model. In [19] an adaptive synthetic sampling (ADASYN) algorithm has been applied to solve the data imbalance problem after that a gradient boosting decision tree (GBDT) classifier is used for glaucoma classification. In Table 10 the comparison is presented, the better performing model is in bold.

The accuracy and specificity of our proposed MVE model are higher than all the considered state-of-the-art model. Though sensitivity is slightly less than [19], but our proposed model has outperformed other approaches in other considered performance matrices. So we can conclude that our proposed MVE approach is better in Glaucoma screening.

7 Conclusion

We have developed an ensemble model based on a majority voting approach (called MVE) for glaucoma detection, in which MLP, SVM, and RF classifiers are used as base classifiers. The efficiency of our proposed model is evaluated with the help of two popular benchmark datasets REFUGE and ORIGA. Model selection is an essential step for the ensemble approach and it also varies with application to application. By seeing the performance analysis, we can conclude that our proposed MVE approach is capable of detecting glaucoma with good performance and is an effective way to improve the glaucoma detection test performance. The proposed model has shown its effectiveness with the REFUGE and ORIGA dataset as the highest accuracy obtained is 99.86%.

8 Future Works

We will analyze our model with different types of noise like Gaussian noise, salt-and-pepper noise, etc. to test the robustness of our proposed model. Also, we will implement a weighted ensemble approach and transfer learning-based approaches to improve the performance in Glaucoma detection.

References

1. Lee DA, Higginbotham EJ (2005) Glaucoma and its treatment: a review. *Am J Health-Syst Pharmacy* 62(7):691–699. <https://doi.org/10.1093/ajhp/62.7.691>
2. Kingman S (2004) Glaucoma is second leading cause of blindness globally. *Bull World Health Organ* 82(11):887–888
3. Orlando JI et al (2020) Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal* 59:101570. <https://www.sciencedirect.com/science/article/pii/S1361841519301100>
4. Dey A, Bandyopadhyay S (2016) Automated glaucoma detection using support vector machine classification method. *Br J Med Med Res* 11:1–12
5. Balasubramanian T, Krishnan S, Mohanakrishnan M, Rao KR, Kumar CV, Nirmala K (2016) Hog feature based SVM classification of glaucomatous fundus image with extraction of blood vessels. In: 2016 IEEE annual India conference (INDICON), pp 1–4
6. Nirmala K, Venkateswaran N, Kumar CV (2017) Hog based naive Bayes classifier for glaucoma detection. In: TENCON 2017—2017 IEEE region 10 conference, pp 2331–2336
7. Kim SJ, Cho KJ, Oh S (2017) Development of machine learning models for diagnosis of glaucoma. *PLOS ONE* 12(5):1–16. <https://doi.org/10.1371/journal.pone.0177726>
8. Parashar DR, Agarwal DK (2021) SVM based supervised machine learning framework for glaucoma classification using retinal fundus images. In: 2021 10th IEEE international conference on communication systems and network technologies (CSNT), pp 660–663
9. Oh S, Park Y, Cho KJ, Kim SJ (2021) Explainable machine learning model for glaucoma diagnosis and its interpretation. *Diagnostics* 11(3). <https://www.mdpi.com/2075-4418/11/3/510>
10. Ravishyam D, Samiappan D (2021) Comparative study of machine learning with novel feature extraction and transfer learning to perform detection of glaucoma in fundus retinal images. In: Sharma TK, Ahn CW, Verma OP, Panigrahi BK (eds) *Soft computing: theories and applications*. Springer, Singapore, pp 419–429
11. Varshney H, Kant U, Gupta H, Verma OP, Sharma TK, Ansari IA (2021) Semantic segmentation of retinal blood vessel with autoencoders. *Soft computing: theories and applications*. Springer, Singapore, pp 563–573
12. Zhou B, Mohammadi F, Lim JS, Forouzesh N, Ghasemzadeh H, Amini N (2021) Analysis of macular thickness deviation maps for diagnosis of glaucoma. In: Bebis G, Athitsos V, Yan T, Lau M, Li F, Shi C, Yuan X, Mousas C, Bruder G (eds) *Advances in visual computing*. Springer, Cham, pp 53–64
13. Zhu Q, Yeh MC, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE Computer Society conference on computer vision and pattern recognition (CVPR’06), vol 2, pp 1491–1498
14. Nandi A, Jana ND, Das S (2020) Improving the performance of neural networks with an ensemble of activation functions. In: 2020 international joint conference on neural networks (IJCNN), pp 1–7
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority oversampling technique. *J Artif Int Res* 16(1):321–357
16. Elangovan P, Nath MK (2021) Glaucoma assessment from color fundus images using convolutional neural network. *Int J Imaging Syst Technol* 31(2):955–971. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22494>
17. Sharma A, Agrawal M, Roy SD, Gupta V (2020) Automatic glaucoma diagnosis in digital fundus images using deep CNNs. Springer, Singapore, pp 37–52. https://doi.org/10.1007/978-981-15-2620-6_3

18. Shoukat A, Akbar S, Hassan SAE, Rehman A, Ayesha N. An automated deep learning approach to diagnose glaucoma using retinal fundus images. In: 2021 international conference on frontiers of information technology (FIT), pp 120–125
19. Guo F, Li W, Tang J, Zou B, Fan Z (2020) Automated glaucoma screening method based on image segmentation and feature extraction. *Med Biol Eng Comput* 58(10):2567–2586. <https://doi.org/10.1007/s11517-020-02237-2>

A Note on Laguerre-Based Appell-Type Daehee Polynomials and Numbers



Waseem A. Khan, Azhar Iqbal, and Mohd Nadeem

Abstract In this paper, we introduce a new class of generalized Laguerre-based Appell-type Daehee polynomials and then derive diverse explicit and implicit summation formulae and symmetric identities by using series manipulation techniques. Multifarious summation formulas and identities are given earlier for some well-known polynomials such as Daehee polynomials and Appell-type Daehee polynomials are generalized.

Keywords Laguerre polynomials · Daehee polynomials · Identities

Mathematics Subject Classification 11B68 · 33C99

1 Introduction

The two variable Laguerre polynomials (2-VLP) $\mathfrak{L}_\omega(\xi, \eta)$ [2] are defined by

$$\frac{1}{(1 - \eta\omega)} \exp\left(\frac{-\xi\omega}{1 - \eta\omega}\right) = \sum_{v=0}^{\infty} \mathfrak{L}_v(\xi, \eta)\omega^v \quad (|\eta\omega| < 1), \quad (1.1)$$

which is equivalently [1] given by

W. A. Khan (✉) · A. Iqbal

Department of Mathematics and Natural Sciences, Prince Mohammad Bin Fahd University, P.O. Box 1664, Al Khobar 31952, Saudi Arabia

e-mail: wkhan1@pmu.edu.sa

A. Iqbal

e-mail: aiqbal@pmu.edu.sa

M. Nadeem

Department of Natural and Applied Sciences, Glocal University, Saharanpur, Uttar Pradesh 247121, India

$$\exp(\xi\omega)C_0(\eta\omega) = \sum_{v=0}^{\infty} \mathbb{L}_v(\xi, \eta) \frac{\omega^v}{v!}. \quad (1.2)$$

From (1.1) and (1.2), we have

$$\mathbb{L}_v(\xi, \eta) = v! \sum_{s=0}^v \frac{(-1)^s \xi^s \eta^{v-s}}{(s!)^2 (v-s)!} = \eta^v \mathbb{L}_v(\xi/\eta). \quad (1.3)$$

Thus, we have

$$\mathbb{L}_v(\xi, \eta) = \frac{(-1)^v \xi^v}{v!}, \quad \mathbb{L}_v(0, \eta) = \eta^v, \quad \mathbb{L}_v(\xi, 1) = \mathbb{L}_v(\xi), \quad (1.4)$$

where $\mathbb{L}_v(\xi)$ are one variable Laguerre polynomials [1].

The Bernoulli polynomials are defined by (see [5–8])

$$\frac{\omega}{e^\omega - 1} e^{\xi\omega} = \sum_{v=0}^{\infty} \mathbb{B}_v(\xi) \frac{\omega^v}{v!}. \quad (1.5)$$

In the case $\xi = 0$, $\mathbb{B}_v = \mathbb{B}_v(0)$.

The second kind of Bernoulli polynomials are given by (see [9–11])

$$\frac{\omega}{\log(1+\omega)} (1+\omega)^\xi = \sum_{v=0}^{\infty} \mathbb{B}_v(\xi) \frac{\omega^v}{v!}. \quad (1.6)$$

Suppose $\xi = 0$, $\mathbb{B}_v = \mathbb{B}_v(0)$.

Let p be a hard and fast abnormal high number and v_p be the normalized exponential valuation of \mathbb{C}_p with $|p|_p = p_p^{v_p} = \frac{1}{p}$. For $\cup D(\mathbb{Z}_p)$ be area of uniformly differentiable function on \mathbb{Z}_p . For $f \in \cup D(\mathbb{Z}_p)$, the bosonic p -adic vital on \mathbb{Z}_p is described via

$$I_0(f) = \int_{\mathbb{Z}_p} f(\xi) d\mu_0(\xi) = \lim_{N \rightarrow \infty} \sum_{\zeta=0}^{p^N-1} f(\zeta) \mu_0(\zeta + p^N \mathbb{Z}_p) = \lim_{N \rightarrow \infty} \frac{1}{p^N} \sum_{\zeta=0}^{p^N-1} f(\zeta). \quad (1.7)$$

By (1.1), we have (see [13–15])

$$I_0(f_1) = I_0(f) + f'(0),$$

where $f_1(\zeta) = f(\zeta + 1)$.

The generating function of the Daehee polynomials are given by (see [14])

$$\frac{\log(1+\omega)}{\omega}(1+\omega)^\xi = \sum_{v=0}^{\infty} \mathbb{D}_v(\xi) \frac{\omega^v}{v!}. \quad (1.8)$$

At $\xi = 0$, $\mathbb{D}_v = \mathbb{D}_v(0)$.

The generating function of the Appell-type Daehee polynomials are given by (see [16])

$$\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} e^{\xi\omega} = \sum_{v=0}^{\infty} \mathbb{D}_v(\xi|\lambda) \frac{\omega^v}{v!}. \quad (1.9)$$

In $\xi = 0$, $\mathbb{D}_v(\lambda) = \mathbb{D}_v(0|\lambda)$.

The first kind of Stirling numbers is defined by (see [1–10])

$$(\xi)_v = \sum_{\sigma=0}^v S_1(v, \sigma) \xi^\sigma, \quad (v \geq 0). \quad (1.10)$$

The second kind of Stirling numbers are given by (see [11–23])

$$\xi^v = \sum_{\sigma=0}^v S_2(v, \sigma) (\xi)_\sigma. \quad (1.11)$$

From (1.10) and (1.11), we have (see [1–23])

$$\frac{1}{\sigma!} (\log(1+\omega))^\sigma = \sum_{v=\sigma}^{\infty} S_1(v, \sigma) \frac{\omega^v}{v!}, \quad (1.12)$$

and

$$\frac{1}{\sigma!} (e^\omega - 1)^\sigma = \sum_{v=\sigma}^{\infty} S_2(v, \sigma) \frac{\omega^v}{v!} \quad (\sigma \geq 0). \quad (1.13)$$

For each $\rho \geq 0$, $S_\rho(v)$ [3] is defined by

$$S_\rho(v) = \sum_{\theta=0}^v \theta^\rho, \quad (1.14)$$

and

$$\sum_{\rho=0}^{\infty} S_\rho(v) \frac{\omega^\rho}{\rho!} = 1 + e^\omega + e^{2\omega} + \cdots + e^{v\omega} = \frac{e^{(v+1)\omega} - 1}{e^\omega - 1}. \quad (1.15)$$

2 Laguerre-Based Appell-Type Daehee Polynomials

Let us assume that $\lambda, \omega \in \mathbb{C}_p$ with $|\lambda\omega|_p < p^{-\frac{1}{p-1}}$. We define Laguerre-based Appell-type Daehee–Hermite polynomials as

$$\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} e^{\xi\omega} C_0(\eta\omega) = \sum_{v=0}^{\infty} {}_L\mathbb{D}_v(\xi, \eta|\lambda) \frac{\omega^v}{v!}, \quad (2.1)$$

and

$$\left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega} C_0(\eta\omega) = \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(\xi, \eta|\lambda) \frac{\omega^v}{v!}. \quad (2.2)$$

Letting $\xi = \eta = 0$, $\mathbb{D}_v(\lambda) = {}_L\mathbb{D}_v(0, 0|\lambda)$ and $\mathbb{D}_v^{(r)}(\lambda) = {}_L\mathbb{D}_v^{(r)}(0, 0|\lambda)$ are the Appell-type Daehee numbers and Appell-type Daehee numbers of order r .

Theorem 2.1 *Let $v \geq 0$. Then*

$${}_L\mathbb{D}_v(\xi, \eta|\lambda) = \sum_{\rho=0}^v \binom{v}{\rho} {}_L\mathbb{D}_\rho^{(r)}(0, \eta|\lambda) \xi^{v-\rho}. \quad (2.3)$$

Proof By (1.2) and (2.1), we acquire (2.3). \square

Theorem 2.2 *We have*

$$\mathbb{L}_v(\xi, \eta) = \sum_{s=0}^v \sum_{l=0}^s \binom{s}{l} \binom{v}{s} \lambda^l \mathbb{D}_l b_{s-l} {}_L\mathbb{D}_{v-s}(\xi, \eta|\lambda). \quad (2.4)$$

Proof From (1.2), (1.6) and (2.1), we have

$$\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} e^{\xi\omega} C_0(\eta\omega) = \sum_{v=0}^{\infty} {}_L\mathbb{D}_v(\xi, \eta|\lambda) \frac{\omega^v}{v!}. \quad (2.5)$$

Since

$$\begin{aligned} e^{\xi\omega} C_0(\eta\omega) &= \frac{\log(1+\lambda\omega)^{\frac{1}{\lambda}}}{\log(1+\omega)} \sum_{v=0}^{\infty} {}_L\mathbb{D}_v(\xi, \eta|\lambda) \frac{\omega^v}{v!} \\ &= \frac{\log(1+\lambda\omega)}{\lambda\omega} \frac{\omega}{\log(1+\omega)} \sum_{v=0}^{\infty} {}_L\mathbb{D}_v(\xi, \eta|\lambda) \frac{\omega^v}{v!} \\ &= \left(\sum_{l=0}^{\infty} \mathbb{D}_l \frac{\lambda^l \omega^2 q^{1l}}{l!} \right) \left(\sum_{s=0}^{\infty} b_s \frac{\omega^s}{s!} \right) \left(\sum_{v=0}^{\infty} {}_L\mathbb{D}_v(\xi, \eta|\lambda) \frac{\omega^v}{v!} \right) \end{aligned}$$

$$\begin{aligned}
&= \left(\sum_{j=0}^{\infty} \left(\sum_{l=0}^s \binom{s}{l} \lambda^l \mathbb{D}_l b_{s-l} \right) \frac{\omega^s}{s!} \right) \left(\sum_{v=0}^{\infty} {}_L \mathbb{D}_v(\xi, \eta | \lambda) \frac{\omega^v}{v!} \right) \\
&= \sum_{v=0}^{\infty} \left(\sum_{s=0}^v \sum_{l=0}^s \binom{s}{l} \binom{v}{s} \lambda^l \mathbb{D}_l b_{s-l} {}_L \mathbb{D}_{v-s}(\xi, \eta | \lambda) \right) \frac{\omega^v}{v!} \\
&\quad \sum_{v=0}^{\infty} \mathbb{L}_v(\xi, \eta) \frac{\omega^v}{v!} \\
&= \sum_{v=0}^{\infty} \left(\sum_{s=0}^v \sum_{l=0}^s \binom{s}{l} \binom{v}{s} \lambda^l \mathbb{D}_l b_{s-l} {}_L \mathbb{D}_{v-s}(\xi, \eta | \lambda) \right) \frac{\omega^v}{v!}, \quad (2.6)
\end{aligned}$$

complete the proof. \square

Theorem 2.3 *We have*

$${}_L \mathbb{D}_v^{(r)}(\xi, \eta | \lambda) = \sum_{k=0}^v \sum_{s=0}^k \binom{k}{s} \binom{v}{k} \mathbb{D}_s^{(r)} \lambda^{k-s} \mathbb{B}_{k-s}^{(k-r+1)}(1) \mathbb{L}_{v-s}(\xi, \eta). \quad (2.7)$$

Proof By (2.2), we have

$$\left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega} C_0(\eta\omega) = \sum_{v=0}^{\infty} {}_L \mathbb{D}_v^{(r)}(\xi, \eta | \lambda) \frac{\omega^v}{v!}.$$

Now

$$\begin{aligned}
&\left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega+\eta\omega^2} = \left(\frac{\log(1+\omega)}{\omega} \right)^r \left(\frac{\lambda\omega}{\log(1+\lambda\omega)} \right)^r e^{\xi\omega+\eta\omega^2} \\
&= \left(\sum_{s=0}^{\infty} \mathbb{D}_s^{(r)} \frac{\omega^s}{s!} \right) \left(\sum_{k=0}^{\infty} \mathbb{B}_k^{(k-r+1)}(1) \frac{\lambda^k \omega^k}{k!} \right) \left(\sum_{v=0}^{\infty} \mathbb{L}_v(\xi, \eta) \frac{\omega^v}{v!} \right) \\
&= \left(\sum_{k=0}^{\infty} \sum_{s=0}^k \binom{k}{s} \mathbb{D}_s^{(r)} \lambda^{k-s} \mathbb{B}_{k-s}^{(k-r+1)}(1) \frac{\omega^k}{k!} \right) \left(\sum_{v=0}^{\infty} \mathbb{L}_v(\xi, \eta) \frac{\omega^v}{v!} \right) \\
&= \sum_{v=0}^{\infty} \left(\sum_{k=0}^v \sum_{s=0}^k \binom{k}{s} \binom{v}{k} \mathbb{D}_s^{(r)} \lambda^{k-s} \mathbb{B}_{k-s}^{(k-r+1)}(1) \mathbb{L}_{v-s}(\xi, \eta) \right) \frac{\omega^v}{v!}, \quad (2.8)
\end{aligned}$$

yields the proof (2.7). \square

Theorem 2.4 *We have*

$${}_L \mathbb{D}_v^{(r)}(\xi, \eta | \lambda) = \sum_{\phi=0}^v \sum_{\tau=0}^{\phi} \binom{v}{\tau} (\xi)_{\phi} S_2(\tau, \phi) {}_L \mathbb{D}_{v-\tau}(0, \chi | \lambda). \quad (2.9)$$

Proof In (2.2),

$$\begin{aligned}
\sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(\xi, \eta|\lambda) \frac{\omega^v}{v!} &= \left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega} C_0(\eta\omega) \\
&= \left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r C_0(\eta\omega)(e^\omega - 1 + 1)^\xi \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(0, \eta|\lambda) \frac{\omega^v}{v!} \sum_{s=0}^{\infty} (\xi)_s \frac{1}{s!} (e^\omega - 1)^s \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(0, \eta|\lambda) \frac{\omega^v}{v!} \sum_{\rho=0}^{\infty} (\xi)_\rho \sum_{\theta=\rho}^{\infty} S_2(\theta, \rho) \frac{\omega^\theta}{\theta!} \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(0, \eta|\lambda) \frac{\omega^v}{v!} \sum_{\tau=0}^{\infty} \sum_{\phi=0}^{\tau} (\xi)_\phi S_2(\tau, \phi) \frac{\omega^\tau}{\tau!} \\
&= \sum_{v=0}^{\infty} \left(\sum_{\tau=0}^v \sum_{\phi=0}^{\tau} \binom{v}{\tau} (\xi)_\phi S_2(\tau, \phi) {}_L\mathbb{D}_{v-\tau}^{(r)}(0, \chi|\lambda) \right) \frac{\omega^v}{v!}. \quad (2.10)
\end{aligned}$$

In view of (2.11), we get (2.10). \square

Theorem 2.5 *We have*

$${}_L\mathbb{D}_v^{(r)}(\xi + \alpha, \eta|\lambda) = \sum_{k=0}^v \sum_{\phi=0}^{\tau} \binom{v}{\tau} (\xi)_\phi S_2(\tau + \alpha, \phi + \alpha) {}_L\mathbb{D}_{v-\tau}(0, \eta|\lambda). \quad (2.11)$$

Proof Let $\xi \rightarrow \xi + \alpha$, Eq. (2.2) as

$$\begin{aligned}
\sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(\xi + \alpha, \eta|\lambda) \frac{\omega^v}{v!} &= \left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega} C_0(\eta\omega) e^{\alpha\omega} \\
&= \left(\frac{\log(1+\omega)}{\log(1+\lambda\omega)^{\frac{1}{\lambda}}} \right)^r C_0(\eta\omega) e^{\alpha\omega} (e^\omega - 1 + 1)^\xi \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v(0, \eta|\lambda) \frac{\omega^v}{v!} e^{\alpha\omega} \sum_{s=0}^{\infty} (\xi)_s \frac{1}{s!} (e^\omega - 1)^s \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(0, \eta|\lambda) \frac{\omega^v}{v!} e^{\alpha\omega} \sum_{s=0}^{\infty} (\xi)_s \sum_{k=s}^{\infty} S_2(k, s) \frac{\omega^k}{k!} \\
&= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(0, \eta|\lambda) \frac{\omega^v}{v!} \sum_{\nu=0}^{\infty} \sum_{\rho=0}^{\nu} (\xi)_\rho S_2(\nu + \alpha, \rho + \alpha) \frac{\omega^\nu}{\nu!}
\end{aligned}$$

$$= \sum_{v=0}^{\infty} \left(\sum_{\nu=0}^v \sum_{\kappa=0}^{\tau} \binom{v}{\tau} (\xi)_{\phi} S_2(\tau + \alpha, \phi + \alpha) {}_L \mathbb{D}_{v-\tau}^{(r)}(0, \eta | \lambda) \right) \frac{\omega^v}{v!}, \quad (2.12)$$

yields the result. \square

Theorem 2.6 *We have*

$${}_L \mathbb{D}_v^{(r)}(\xi, \eta | \lambda) = \sum_{l=0}^v \binom{v}{l} {}_L \mathbb{D}_{v-l}^{(r-k)}(\xi, \eta | \lambda) \mathbb{D}_l^{(k)}(\lambda). \quad (2.13)$$

Proof In (2.2), we see

$$\begin{aligned} \sum_{v=0}^{\infty} {}_L \mathbb{D}_v^{(r)}(\xi, \eta | \lambda) \frac{\omega^v}{v!} &= \left(\frac{\log(1 + \omega)}{\log(1 + \lambda\omega)^{\frac{1}{\lambda}}} \right)^r e^{\xi\omega} C_0(\eta\omega) \\ &= \left(\frac{\log(1 + \omega)}{\log(1 + \lambda\omega)^{\frac{1}{\lambda}}} \right)^{r-k} \left(\frac{\log(1 + \omega)}{\log(1 + \lambda\omega)^{\frac{1}{\lambda}}} \right)^k e^{\xi\omega} C_0(\eta\omega) \\ &= \sum_{\omega=0}^{\infty} {}_L \mathbb{D}_v^{(r-k)}(\xi, \eta | \lambda) \frac{\omega^v}{v!} \sum_{l=0}^{\infty} {}_L \mathbb{D}_l^{(k)}(0, 0 | \lambda) \frac{\omega^l}{l!} \\ &= \sum_{v=0}^{\infty} \left(\sum_{l=0}^v \binom{v}{l} {}_L \mathbb{D}_{v-l}^{(r-k)}(\xi, \eta | \lambda) \mathbb{D}_l^{(k)}(\lambda) \right) \frac{\omega^v}{v!}. \end{aligned} \quad (2.14)$$

In view of (2.14), we get the required result. \square

3 General Identities

Theorem 3.1 *We have*

$$\begin{aligned} \sum_{v=0}^{\nu} \binom{v}{\nu} a^{v-s} b^{\nu} {}_L \mathbb{D}_{v-\nu}^{(r)}(b\xi, b\eta | \lambda) {}_L \mathbb{D}_{\nu}^{(r)}(a\xi, a\eta | \lambda) \\ = \sum_{\nu=0}^{\nu} \binom{\nu}{\nu} b^{\nu-s} a^{\nu} {}_L \mathbb{D}_{\nu-\nu}^{(r)}(a\xi, a\eta | \lambda) {}_L \mathbb{D}_{\nu}^{(r)}(b\xi, b\eta | \lambda). \end{aligned} \quad (3.1)$$

Proof Let

$$\mathbf{A}(\omega) = \left(\frac{\log(1+a\omega) \log(1+b\omega)}{(\log(1+\lambda\omega)^{\frac{a}{\lambda}})(\log(1+\lambda\omega)^{\frac{b}{\lambda}})} \right)^r e^{(a+b)\xi\omega} C_0(a\eta\omega) C_0(b\eta\omega).$$

$$\mathbf{A}(\omega) = \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} a^{v-v} b^v {}_L\mathbb{D}_{v-v}^{(r)}(b\xi, b\eta|\lambda) {}_L\mathbb{D}_v^{(r)}(a\xi, a\eta|\lambda) \right) \frac{\omega^v}{v!}.$$

And

$$\mathbf{A}(\omega) = \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \binom{v}{v} b^{v-v} a^v {}_L\mathbb{D}_{v-v}^{(r)}(a\xi, a\eta|\lambda) {}_L\mathbb{D}_v^{(r)}(b\xi, b\eta|\lambda) \right) \frac{\omega^v}{v!},$$

yields the result. \square

Theorem 3.2 We have

$$\begin{aligned} & \sum_{v=0}^v \sum_{\alpha=0}^{a-1} \sum_{\beta=0}^{b-1} \binom{v}{v} a^{v-s} b^v {}_L\mathbb{D}_{v-v}^{(r)} \left(b\xi + \frac{b}{a}\alpha + \beta, \eta|\lambda \right) {}_L\mathbb{D}_v^{(r)}(a\xi, \eta|\lambda) \\ &= \sum_{v=0}^v \sum_{\beta=0}^{a-1} \sum_{\alpha=0}^{b-1} \binom{v}{v} b^{v-v} a^v {}_L\mathbb{D}_{v-v}^{(r)} \left(a\xi + \frac{a}{b}\alpha + \beta, \eta|\lambda \right) {}_L\mathbb{D}_v^{(r)}(b\xi, \eta|\lambda). \quad (3.2) \end{aligned}$$

Proof Let

$$\begin{aligned} \mathbf{B}(\omega) &= \left(\frac{\log(1+a\omega) \log(1+b\omega)}{(\log(1+\lambda\omega)^{\frac{a}{\lambda}})(\log(1+\lambda\omega)^{\frac{b}{\lambda}})} \right)^r \\ &\quad \frac{(e^{ab\omega} - 1)^2}{(e^{a\omega} - 1)(e^{b\omega} - 1)} e^{ab(\xi+\zeta)\omega} C_0(a\eta\omega) C_0(b\eta\omega) \\ &= \left(\frac{\log(1+a\omega)}{\log(1+\lambda\omega)^{\frac{a}{\lambda}}} \right)^r e^{ab\xi\omega} C_0(a\eta\omega) \sum_{i=0}^{a-1} e^{b\omega\alpha} \left(\frac{\log(1+b\omega)}{\log(1+\lambda\omega)^{\frac{b}{\lambda}}} \right)^r \\ &\quad e^{ab\eta\omega} C_0(b\eta\omega) \sum_{l=0}^{b-1} e^{a\omega\nu} \\ &= \sum_{v=0}^{\infty} \left(\sum_{v=0}^v \sum_{\alpha=0}^{a-1} \sum_{\beta=0}^{b-1} \binom{v}{v} a^{v-s} b^s {}_L\mathbb{D}_{v-v}^{(r)} \right. \\ &\quad \left. \left(b\xi + \frac{b}{a}\alpha + \beta, \eta|\lambda \right) {}_L\mathbb{D}_v^{(r)}(a\xi, \eta|\lambda) \right) \frac{\omega^v}{v!}. \quad (3.3) \end{aligned}$$

On the other hand, we have

$$\mathbf{B}(\omega) = \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \sum_{\theta=0}^{a-1} \sum_{\alpha=0}^{b-1} \binom{v}{\rho} b^{v-\rho} a^{\rho} {}_L\mathbb{D}_{v-\rho}^{(r)} \left(a\xi + \frac{a}{b}\alpha + \beta, a\eta | \lambda \right) {}_L\mathbb{D}_{\rho}^{(r)}(b\xi, \eta | \lambda) \right) \frac{\omega^v}{v!},$$

yields the desired result. \square

Theorem 3.3 *We have*

$$\begin{aligned} & \sum_{\rho=0}^v \binom{v}{\rho} a^{v-\rho} b^{\rho} {}_L\mathbb{D}_{v-\rho}^{(r)}(b\xi, b\xi | \lambda) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(b-1), a\xi {}_L\mathbb{D}_{\rho-\theta}^{(r)}(a\eta, a\xi | \lambda) \\ & \sum_{\rho=0}^v \binom{v}{\rho} b^{v-\rho} a^{\rho} {}_L\mathbb{D}_{v-\rho}^{(r)}(a\xi, a\xi | \lambda) \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(a-1) {}_L\mathbb{D}_{\rho-\theta}^{(r)}(b\eta, b\xi | \lambda). \end{aligned} \quad (3.4)$$

Proof Let

$$\begin{aligned} \mathbf{C}(\omega) &= \left(\frac{\log(1+a\omega) \log(1+b\omega)}{(\log(1+\lambda\omega)^{\frac{a}{\lambda}})(\log(1+\lambda\omega)^{\frac{b}{\lambda}})} \right)^r \\ &\quad \frac{(e^{ab\omega} - 1)}{(e^{a\omega} - 1)(e^{b\omega} - 1)} e^{ab(\xi+\eta)\omega} C_0(a\xi\omega) C_0(b\xi\omega) \\ &= \sum_{v=0}^{\infty} {}_L\mathbb{D}_v^{(r)}(b\xi, b\xi | \lambda) \frac{(a\omega)^v}{v!} \sum_{\theta=0}^{\infty} S_{\theta}(b-1) \sum_{\rho=0}^{\infty} {}_L\mathbb{D}_{\rho}^{(r)}(a\eta, a\xi | \lambda) \frac{(b\omega)^{\rho}}{\rho!} \\ &= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \binom{v}{\rho} a^{v-\rho} b^{\rho} {}_L\mathbb{D}_{v-\rho}^{(r)}(b\xi, b\xi | \lambda) \right. \\ &\quad \left. \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(b-1) {}_L\mathbb{D}_{\rho-\theta}^{(r)}(a\eta, a\xi | \lambda) \right) \frac{\omega^v}{v!}. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbf{C}(\omega) &= \sum_{v=0}^{\infty} \left(\sum_{\rho=0}^v \binom{v}{\rho} b^{v-\rho} a^{\rho} {}_L\mathbb{D}_{v-\rho}^{(r)}(a\xi, a\xi | \lambda) \right. \\ &\quad \left. \sum_{\theta=0}^{\rho} \binom{\rho}{\theta} S_{\theta}(a-1) {}_L\mathbb{D}_{\rho-\theta}^{(r)}(b\eta, b\xi | \lambda) \right) \frac{\omega^v}{v!}. \end{aligned}$$

The complete proof of this theorem. \square

4 Concluding Remarks

In this paper, we have got described the Laguerre-based totally Appell-type Daehee polynomials, their precise forms, and miscellaneous houses and positive symmetry identities, in particular, a few interesting series representations. We have deduced a few applicable homes with the aid of the usage of the shape and the relations glad by using the presently generalized Laguerre polynomials.

References

1. Andrews LC (1985) Special functions for engineers and mathematicians. Macmillan Co., New York
2. Dattoli G, Torre A (1998) Operational methods and two variable Laguerre polynomials. *Atti Accad Sci Torino Cl Sci Fis Mat Natur* 132:3–9
3. Deeba E, Rodrigues D (1991) Stirling series and Bernoulli numbers. *Am Math Mon* 98:423–426
4. Dolgy DV, Khan WA (2021) A note on type two degenerate poly-Changhee polynomials of the second kind. *Symmetry* 13(579):1–12
5. Haroon H, Khan WA (2018) Degenerate Bernoulli numbers and polynomials associated with Hermite polynomials. *Commun Korean Math Soc* 33(2):651–669
6. Khan WA, Araci S, Acikgoz M, Haroon H (2017) A new class of partially degenerate Hermite–Genocchi polynomials. *J Nonlinear Sci Appl* 10(9):5072–5081
7. Khan WA, Pathan MA (2019) On generalized Lagrange–Hermite–Bernoulli and related polynomials. *Acta Comment Univ Tarta Math* 23(2):211–224
8. Khan WA, Haroon H (2016) Some symmetric identities for the generalized Bernoulli, Euler and Genocchi polynomials associated with Hermite polynomials. *Springer Plus* 5:1–21
9. Khan WA, Muhiuddin G, Muhyi A, Al-Kadi D (2021) Analytical properties of type 2 degenerate poly-Bernoulli polynomials associated with their applications. *Adv Differ Equ* 2021(420):1–18
10. Khan WA, Muhyi A, Ali R, Alzobydi KAH, Singh M, Agarwal P (2021) A new family of degenerate poly-Bernoulli polynomials of the second kind with its certain related properties. *AIMS Math* 6(11):12680–12697
11. Khan WA, Acikgoz M, Duran U (2020) Note on the type 2 degenerate multi-poly-Euler polynomials. *Symmetry* 12:1–10
12. Khan WA, Ali R, Alzobydi KAH, Ahmed A (2021) A new family of degenerate poly-Genocchi polynomials with its certain properties. *J Funct Spaces* 2021:8. Article ID 6660517
13. Kim DS (2013) A note on the λ -Daehee polynomials. *Int J Math Anal* 7(62):3069–3080
14. Kim DS, Kim T (2013) Daehee numbers and polynomials. *Appl Math Sci* 7(120):5969–5976
15. Kim T, Lee S, Mansour T, Seo JJ (2014) A note on q -Daehee polynomials and numbers. *Adv Stud Contemp Math (Kyungshang)* 24(2):155–160
16. Kwon JK, Rim SH, Park JW (2015) A note on the Appell-type Daehee polynomials. *Glob J Pure Appl Math* 11(5):2745–2753
17. Muhiuddin G, Khan WA, Duran U, Al-Kadi D (2021) Some identities of the multi poly-Bernoulli polynomials of complex variable. *J Funct Spaces* 2021:8. Article ID 7172054
18. Muhiuddin G, Khan WA, Duran U (2021) Two variable type 2 Fubini polynomials. *Mathematics* 9(281):1–13
19. Muhiuddin G, Khan WA, Muhyi A, Al-Kadi D (2021) Some results on type 2 degenerate poly-Fubini polynomials and numbers. *Comput Model Eng Sci* 29(2):1051–1073. <https://doi.org/10.32604/cmes.2021.016546>
20. Muhiuddin G, Khan WA, Al-Kadi D (2021) Construction on the degenerate poly-Frobenius–Euler polynomials of complex variable. *J Funct Spaces* 2021:9. Article ID 3115424. <https://doi.org/10.1155/2021/3115424>

21. Pathan MA, Khan WA (2015) Some implicit summation formulas and symmetric identities for the generalized Hermite–Bernoulli polynomials. *Mediterr J Math* 12:679–695
22. Pathan MA, Khan WA (2016) A new class of generalized polynomials associated with Hermite and Euler polynomials. *Mediterr J Math* 13:913–928
23. Pathan MA, Khan WA (2021) A new class of generalized polynomials associated with Hermite and poly-Bernoulli polynomials. *Miskolc Math J* 22(1):317–330

Type-II Fuzzy Kernel-Based Multi-layer Extreme Learning Machine



Avatharam Ganivada and Sayima Mukhtar

Abstract Kernel-based multi-layer extreme learning machine (KMLELM) is developed by integrating ML-ELM and a kernel function (RBF). The model is aimed at partially solving the issues of H-ELM and ML-ELM such as eliminating the number of hidden nodes in every layer, achieving optimal model generalization, reducing the reconstruction error using invertible matrix, and decreasing the execution time by compacting the memory storage of all the hidden layers. The KMLELM can deal with small datasets effectively, but it becomes slow while dealing with large datasets. The KMLELM has also limitation of handling uncertainty which arises from overlapping regions between class boundaries. To overcome all these issues, a new type-II fuzzy KMLELM for representation learning and classification is developed by integrating type-II fuzzy set theory into KMLELM. The type-II fuzzy set is used for handling uncertainty. Further, applications of representation learning and classification to the proposed model are provided.

Keywords Type-II fuzzy set · Extreme learning machine · Fuzzy kernel function · Uncertainty modeling

1 Introduction

Extreme learning machine (ELM) is a feed-forward neural network consisting of a single hidden layer. The learnable parameters of ELM like weights and bias are initialized randomly between the nodes of the input and hidden layers. The ELM analytically calculates the output between the hidden and the output layers using a least square method [1]. It is shown in the theoretical studies that ELM maintains the SLFN's universal approximation capability though it works with randomly generated hidden nodes. The learning speed of ELM is much faster than traditional learning

A. Ganivada (✉) · S. Mukhtar

School of Computer and Information Sciences, University of Hyderabad, Hyderabad, Telangana 500046, India

e-mail: avatharg@uohyd.ac.in

URL: https://scis.uohyd.ac.in/People/profile/ga_profile.php

methods like gradient-based back propagation algorithm. It avoids the tuning of the parameters (weights and bias) which leads to learn the parameters faster than the traditional algorithms. Because of the universal approximation capability, fast training speed, and good generalization performance, ELM attracts the attention of many researchers.

Extreme learning machine is integrated with auto-encoders to learn the feature representation. Auto-encoders are neural networks that are trained to copy the input to the outputs. In this network, the hidden layer has lower dimensional space than the input layer. The auto-encoder is to learn a lower dimensional representation for a higher dimensional data [2]. During training, the network does capture the most important features of the input data. However, simple extreme learning machine can not extract the abstract features for a complex dataset. Because of its shallow architecture, the multi-layer extreme learning machine (MLELM) is developed in [3]. It consists of multiple hidden layers where each hidden layer is constructed by ELM-based auto-encoders. In MLELM, each hidden layer acts as a feature detector. Here, the lower hidden layers detect simple features, and these features are then sent to the higher hidden layers to extract the more abstract features. Another variant of MLELM, called hierarchical extreme learning machine (H-ELM), is discussed in [4]. It consists of two individual procedures that are representation learning and classification. Once the representation learning is completed, it is used as input for the individual ELM for classification.

Kernel-based multi-layer extreme learning machine (KMLELM) [5] is based on ML-ELM. It is defined by employing the kernel learning in MLELM. The KMLELM resolves the several drawbacks of MLELM partially. It can handle with unstable and suboptimal performance, caused by manual tuning of number of hidden nodes. The model becomes optimal using the kernel learning method. Using the fuzzy sets and extreme learning machine, a classification approach is provided to classify the clinical datasets. By applying the process of fuzzification to the features of clinical datasets, the model achieves better generalization performance than ELM. It also reduces the training time [6]. More recently, kernel-based extreme learning machine resolves the multi-label classification problem to some extent by incorporating the ordinary fuzzy set theory into it. However, the fuzzy multi-layer extreme learning machine (FMLELM) could not provide satisfactory results for multi-label datasets. The training time of KMLELM depends on the number of training datasets because of the kernel trick. It deals with small-scale datasets effectively, but it requires large memory for for large-scale datasets and gradually becomes slow during the training process. Therefore, the model has slow training speed [7].

Different from the existing KMLELM, in this paper, we propose a new method of type-II fuzzy kernel-based multi-layer extreme learning machine for classification. The real-life dataset is generally associated with uncertainty. Therefore, a type-II fuzzy set is used for handling uncertainty. In order to improve the performance of the KMLELM, we integrate the type-II fuzzy set theory and the KMLELM and develop a type-II fuzzy kernel-based multi-layer extreme learning machine (type-II FKMELM). Further, applications of the proposed model to representation learning and classification are generated. The fuzzy logic handles the fuzziness effectively

and thus minimizes the effect of uncertainty and improves generalization ability of the model. The type-II fuzzy set is to reduce the fuzziness of a membership function of an ordinary type-I fuzzy set. Here, the membership function of type-II fuzzy set is also a type-I fuzzy set [6, 8].

The paper is organized as follows: Sect. 2 details existing multi-layer extreme learning machines and kernel-based multi-layer extreme learning machines. The proposed type-II fuzzy kernel multi-layer extreme learning machine (type-II FKM-LELM) is discussed in Sect. 3. Section 4 discusses the experimental results of the proposed model. Conclusions of this work are provided in Sect. 5.

2 Related Work

The aforesaid two existing models of multi-layer extreme learning machine and kernel-based multi-layer extreme learning machine are discussed in this section. These are used for comparing the performance of the proposed model.

2.1 Multi-layer Extreme Learning Machine

The multi-layer extreme learning machine (MLELM) for representation learning is developed in [3]. In MLELM, there are multiple hidden layers followed by the final hidden layer which is for classification [9]. The hidden layers are to extract the useful information of the input data. Each hidden layer is constructed by ELM-AE (Fig. 1).

Let $S = \{(x_k, t_k) | k = 1, 2, \dots, n\}$ denote k number of training patterns and target values. Here, a pattern $x_k = \{x_{k1}, \dots, x_{kd}\}$ and $t_k = \{t_{k1}, t_{k2}, \dots, t_{kc}\}$. A pattern is a d dimensional vector and c denotes the number of classes. All the patterns belong to c number of classes.

Let $x^{(i)} = \{x_1^{(i)}, \dots, x_n^{(i)}\}$ be i th data representation of an input $x_k^{(i)}$, $k = 1, \dots, n$ and $\Gamma^{(i)} = [\gamma_1^{(i)}, \dots, \gamma_n^{(i)}]$ be i th transformation matrix. Here, $\gamma_k^{(i)}$ is the transformation vector that is used for representation learning corresponding to an input $x_k^{(i)}$. The transformation matrix $\Gamma^{(i)}$ is defined as

$$H^{(i)}\Gamma^{(i)} = X^{(i)}, \quad (1)$$

where $H^{(i)}$ is the output matrix of the i th layer with respect to $X^{(i)}$. The transformation matrix $\Gamma^{(i)}$ can be solved using

$$\Gamma^{(i)} = (H^{(i)})^T \left(\frac{1}{c} + H^{(i)}(H^{(i)})^T \right)^{-1} X^{(i)} \quad (2)$$

Final representation of $X^{(i)}$ is defined as

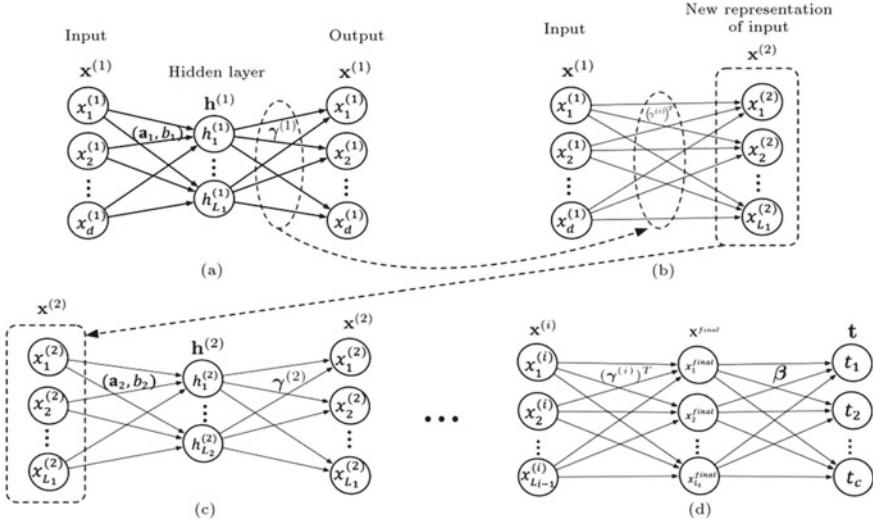


Fig. 1 **a** ELM-AE produces transformation matrix $\gamma^{(1)}$ for representation learning. **b** New input representation is obtained by $g(x^{(1)}(\gamma^{(1)})^T)$. **c** x^2 is input for next new data representation. **d** $x^{(\text{final})}$ is final output that is used for classification after completing the representation learning

$$X^{(\text{final})} = g(X^{(i)}(\Gamma^{(i)})), \quad (3)$$

where $X^{(\text{final})}$ is the final data representation which is used as an input to find the output weight matrix β in order to learn the ELM classifier.

$$X^{(\text{final})}\beta = T \quad (4)$$

T is target values corresponding to given input matrix of size $n \times c$, and β can be defined as

$$\beta = X^{(\text{final})} \left(\frac{1}{c} + X^{(\text{final})}(X^{(\text{final})})^T \right)^{-1} T \quad (5)$$

2.2 Kernel-Based Multi-layer Extreme Learning Machine

To achieve the better generalization and much faster learning speed, kernel learning is employed in an extreme learning machine [5, 10]. Kernel learning is popular for avoiding the tuning of parameters and thus achieves the optimal performance [10, 11]. The KMLELM for representation learning is based on kernel learning and MLELM. The KMLELM describes two separate learning procedures. The first procedure is an unsupervised representation learning and the second procedure is based on classification (supervised procedure) using kernel ELM (KELM). The idea

of KELM is to apply the kernel function to the input matrix $X^{(i)}$ in order to obtain the kernel matrix $\Omega^{(i)}$. It is given as

$$\Omega^{(i)} = K^i(X_k, X_j) \cong \exp(-||X_k - X_j||)/2\sigma_i^2; k, j = 1, 2, \dots, n. \quad (6)$$

As similar to Eq. 1, the i th transformation matrix $\Gamma^{(i)}$ in KELM-AE is learned as

$$\Omega^{(i)}\Gamma^{(i)} = X^{(i)}, \quad (7)$$

where $\Gamma^{(i)}$ can be calculated using the exact inverse of the kernel matrix using the following equation.

$$\Gamma^{(i)} = \left(\frac{1}{c} + \Omega^{(i)} \right)^{-1} X^{(i)} \quad (8)$$

The kernel-based extreme learning machine is described in detail in [5].

3 Proposed Type-II Fuzzy Kernel Multi-layer Extreme Learning Machine (Type-II FKMELM)

In kernel-based extreme learning method, we incorporate the concepts of type-II fuzzy set into the kernel function (see Eq. 6). We explain the concepts of type-II fuzzy set followed by the designing process of fuzzy kernel function which is employed in multi-layer learning machine. We initially normalize the aforesaid input matrix as

$$\mu(X) = \frac{X^{(i)} - X_{\min}}{X_{\max} - X_{\min}} \quad (9)$$

where as $X_{(\max)}$ and $X_{(\min)}$ are the maximum and minimum values of the input matrix, respectively.

A type-II fuzzy set of a set $A \subseteq X^{(i)}$ is given as

$$A_{\text{type-II}} = x^{(k)}, \widehat{\mu_A(x^{(k)})}, \quad (10)$$

where $\widehat{\mu_A(x^{(k)})}$ is type-2 membership function which has lower and upper memberships. The upper and lower membership functions, characterized by a type-II fuzzy set, of a pattern $x^{(k)} \in A \subseteq X^{(i)}$ are given as

$$\mu^{(\text{upper})}(\widehat{x^{(k)}}) = [\widehat{\mu_A(x^{(k)})}]^\alpha \quad (11)$$

$$\mu^{(\text{lower})}(\widehat{x^{(k)}}) = [\widehat{\mu_A(x^{(k)})}]^{\frac{1}{\alpha}} \quad (12)$$

where the value of α is chosen as 0.8 using the trial and error method satisfying $0 \leq \alpha \leq 1$. Next, a membership function is obtained based on the upper and lower memberships using a Hamacher T conorm.

$$\widehat{\mu'(g^k)} = \frac{\mu^{(\text{upper})}(\widehat{x^{(k)}}) + \mu^{(\text{lower})}(\widehat{x^{(k)}}) + (\lambda - 2)\mu^{(\text{upper})}(\widehat{x^{(k)}})\mu^{(\text{lower})}(\widehat{x^{(k)}})}{1 - (1 - \lambda)\mu^{(\text{upper})}(\widehat{x^{(k)}})\mu^{(\text{lower})}(\widehat{x^{(k)}})} \quad (13)$$

$$\mu'(\widehat{g}) = \frac{\sum_{k=0}^n \mu'(\widehat{g^k})}{n} \cong \frac{\sum_{k=0}^n \sum_{l=0}^d \mu'(g^{kl})}{nd}, \quad (14)$$

where n is the total number of patterns and d is the number of attributes (features) of a pattern.

Now apply the membership function $\mu'(g)$ and kernel function $K^i(x_k, x_j)$ to the input matrix $X^{(i)}$ to obtain a type-II fuzzy-based kernel matrix, denoted by $\Omega^{(i)}$.

$$\Omega^{(i)} = \mu'(g) \exp \frac{-K^i(x_k, x_j)}{\sigma_i^2} \quad (15)$$

The i th transformation matrix $\Gamma^{(i)}$ is defined as

$$\Omega^{(i)} \Gamma^{(i)} = X^{(i)} \quad (16)$$

The transformation matrix $\Gamma^{(i)}$ for input matrix $X^{(i)}$ is defined as

$$\Gamma^{(i)} = \left(\frac{1}{c} + \Omega^{(i)} \right)^{-1}. \quad (17)$$

Here, a regularization parameter c is employed in $\Gamma^{(i)}$ to reduce the overfitting for better generalization [10]. To reduce the reconstruction error, transformation matrix $\Gamma^{(i)}$ is obtained by exact inverse of kernel matrix rather than pseudoinverse. In Eq. 16, $\Gamma^{(i)}$ is used for representation learning. The learning procedure is obtained using

$$X^{(\text{final})} = g(X^{(i)}(\Gamma^{(i)})), \quad (18)$$

where $X^{(\text{final})}$ is the final representation of an input vector. It is used as hidden layer output to compute the output weight β . It is computed as

$$X^{(\text{final})} \beta = T \quad (19)$$

T is target values corresponding to given input matrix of size $n \times c$, and β can be defined as

$$\beta = \left(\frac{1}{c} + \Omega^{(i)} \right)^{-1}. \quad (20)$$

The algorithmic form of the proposed type-II fuzzy kernel-based extreme learning machine is provided as follows:

Algorithm 1 Proposed type-II fuzzy-based FKMELM

Input: Given input matrix X^i , regularization parameter c , kernel parameter σ for all layers and activation g_i .

Output: Output weights

- 1: Obtain $\gamma^1, \gamma^2, X^{(\text{final})}$, $\gamma_{\text{unified}} = \text{KMELM}(X^i, c, \sigma, g_i)$
- 2: Normalize the input data between 0 and 1.
- 3: Calculate the upper and lower memberships using Eqs. 11 and 12, respectively.
- 4: Calculate the membership function using the Hamacher T conorm Eq. 14.
- 5: Construct the kernel matrix and apply the membership function to obtain Ω^i (use Eq. 15).
- 6: Calculate the output weight $\beta = (\frac{1}{c} + \Omega^{(i)})^{-1}$ (see Eq. 20).

4 Experimental Results

The performance of type-II fuzzy-based KMELM is evaluated over three benchmark datasets which are publicly available at OpenML (<https://www.openml.org>). The datasets are downloaded from UCI repository [12]. The characteristics of datasets are provided in Table 1.

4.1 Experiment Setup

The experiments are conducted on a desktop computer with a core i7-11700 2.50-GHz processor and 16-GB RAM running MATLAB R2019b. In all experiments, the number of hidden layers is set equal to 3. For the proposed method, tuning the number of hidden nodes in each layer is avoided. However, the number of hidden nodes for ML-ELM is set as $100 \times r$ ($r = 2, \dots, 15$), and the regularization parameter is set as C^j , $\{j = -7, -5, \dots, 7\}$ in each layer.

4.2 Training and Testing Accuracy of the Type-II FKMELM

The results of Type-II FKMELM, as compared to MLELM and KMELM, are given in Table 2. It also shows the training time, training accuracy, and test accuracy

Table 1 Properties of data

S. No.	Dataset	Instances	Features
1	Cmc	1473	10
2	Segment	2310	25
3	Har	10,299	562

Table 2 Performance comparison for different algorithms on three datasets

Dataset	Algorithm	Training accuracy	Testing accuracy (%)	Training time (s)
Cmc	MLELM	39	32	0.78
	KMELLM	41	46.30	0.40
	Type-II fuzzy KMELLM	99.90	93.57	0.38
Segment	ML-ELM	40	37	2.00
	KMELLM	41	46.30	9.43
	Type-II fuzzy KMELLM	99.00	97.57	8.23
Har	MLELM	31	28.61	32
	KMELLM	32	29	36
	Type-II fuzzy KMELLM	96.31	95.24	54

Bold values are the best values of the algorithm

values. The test accuracy and training time are compared for each algorithm on different three benchmark datasets. As we can see, the proposed type-II KMELLM shows the tremendous change in training and test accuracy values for all datasets and achieves impressive performance as compared to the existing algorithms, MLELM and KMELLM. The results in the table indicate that the proposed type-II ML-KELM outperforms the MLELM and KMELLM.

5 Conclusion

In this study, a method of type-II fuzzy kernel multi-layer extreme learning machine (type-II FKMELLM) for feature representation and classification is developed. The type-II fuzzy set defines a fuzzy kernel for multi-layer extreme learning machine. A component of type-II fuzzy set, i.e., Hamacher T-norm, to find membership values is used. The type-II fuzzy set enables the model to deal with uncertain information. The proposed model is tested on three real-life datasets. Classification accuracy of the proposed model for the three datasets is found to be superior to two existing models namely multi-layer extreme learning machine and kernel-based multi-layer extreme learning machine. The type-II fuzzy kernel-based multi-layer extreme learn-

ing machine obtains high classification accuracy for the three datasets. It is highly improved to approximately 40%. The proposed method has benefits of eliminating the random projection of input weight and bias. The model also needs less execution time and less memory storage. Experimental results reveal that the proposed model performs very well for different datasets with varying number of samples from 1473 to 10,299. Degradation of generalization (underfitting) could be a limitation of the proposed type-II FKMELM because there is no weight optimization based on back propagation using the gradient descent method.

Acknowledgements This work is done under a research project funded by Institution of Eminence, University of Hyderabad, Govt. of India.

References

1. Guang BH, Qin YZ, Chee KS (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70:489–501. <https://doi.org/10.1016/j.neucom.2005.12.126>
2. Liyanarachchi LCK, Yan Y, Guang BH, Zhengyou Z (2016) Dimension reduction with extreme learning machine. *IEEE Trans Image Process* 25:3906–3918. <https://doi.org/10.1109/TIP.2016.2570569>
3. Kasun LLC, Zhou H, Huang G-B, Vong CM (2013) Representational learning with ELMs for big data. *IEEE Intell Syst* 28(6):31–34
4. Jie Xiong T, Chenwei D, Guang BH (2016) Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst* 27:809–821. <https://doi.org/10.1109/TNNLS.2015.2424995>
5. Wong CM, Vong CM, Wong PK, Cao J (2018) Kernel-based multilayer extreme learning machines for representation learning. *IEEE Trans Neural Netw Learn Syst* 29:757–762. <https://doi.org/10.1109/TNNLS.2016.2636834>
6. Kindie BN, Nehemiah HK, Arputharaj K (2016) Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets. *Inform Med Unlocked* 2:1–11. <https://doi.org/10.1016/J.IMU.2016.01.001>
7. Yanika K, Punyaphol H, Pakarat M, Khamron S (2019) Kernel extreme learning machine based on fuzzy set theory for multi-label classification. *Int J Mach Learn Cybern* 10:979–989. <https://doi.org/10.1007/S13042-017-0776-3/TABLES/6>
8. Nilesh NK, Jerry MM (1998) Introduction to type-2 fuzzy logic systems. In: IEEE international conference on fuzzy systems proceedings & IEEE world congress on computational intelligence, Anchorage, AK, pp 915–920. <https://doi.org/10.1109/FUZZY.1998.686240>
9. Migel DT, Mark DM (2015) Deep extreme learning machines for classification. In: Cao J, Mao K, Cambria E, Man Z, Toh KA (eds) *Proceedings of ELM-2014 in adaptation, learning and optimization*. Springer, pp 345–354. https://doi.org/10.1007/978-3-319-14063-6_29
10. Guang BH, Hongming Z, Xiaojian D, Rui Z (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern Part B Cybern* 42:513–529. <https://doi.org/10.1109/TSMCB.2011.2168604>
11. Alessia M, Marco T, Nello C (2009) Support vector machines. *WIREs Comput Stat* 1:283–289. <https://doi.org/10.1002/wics.49>
12. Dua D, Graff C (2017) UCI machine learning repository. <http://archive.ics.uci.edu/ml>

Energy Trading in Smart Grids Using Game Theoretic Approach



Anshul Agarwal 

Abstract The smart grid is the emerging power grid that uses the advanced communication, control, and automation technologies to generate, transmit, distribute, and trade energy. The flow of energy in the smart grid is bidirectional and the consumers are evolved into active participants in energy trading. The energy when produced in surplus is either sold back to the grid (main energy supplier) or is stored in energy storage systems and traded within the grid with other users using peer-to-peer model. In this paper, an approach is developed that integrates the direct trading mechanism between the users and utility, as well as the peer-to-peer trading model in a smart grid. The advantage of trading with other prosumers (ones that consume and produce energy) in the peer-to-peer (P2P) model in comparison to the direct trading with the utility is studied. The game theory auction-based approach is introduced in the P2P energy trading model. It was concluded that the integration of forecasting models with the auction models increased the overall cost savings of prosumers in the P2P energy trading model.

Keywords Game theory · Peer-to-peer energy trading · Auction · Prosumers · Support vector regression (SVR)

1 Introduction

Smart grid is a power grid which uses the emerging technologies and techniques in fields of communication, control, and automation to efficiently generate, transmit, distribute, and trade energy. According to the Climate Action Tracker [1], 60% of India's electricity comes from burning fossil fuels, mostly coal and natural gas. The renewable energy resources can be an alternative to the conventional energy sources for providing energy. This is done by integrating both the renewable and conventional energy resources and adapting a bidirectional or two-way communication infrastructure in the smart grids [2]. The smart grid benefits both the utilities

A. Agarwal (✉)

Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT) Nagpur, Nagpur, Maharashtra 440010, India

e-mail: anshulagarwal@cse.vnit.ac.in

and prosumers (ones who produce and consume energy [3]) in various ways. For instance, prosumers can also trade the excessive energy produced locally to generate profit [4, 5]. To encourage and promote the use of renewable energy sources, various policies and programs are also implemented by the government [6, 7]. This is possible because of the two-way communication infrastructure provided by the smart grid. The research and studies mentioned in [8] provide an overview of different technologies and advanced communication required in building a smart grid infrastructure. The smart grids are complex systems, and hence to implement the energy trading process, it needs reliable solutions and mechanisms to facilitate this process [9].

Energy trading in a P2P market can be adopted by the smart grids. Different issues related to power limit distribution and PMU allocation in smart grids have been discussed in the literature [10–12]. Prosumers can either use a centralized approach [13] or a decentralized approach [14] for communication in a P2P market platform. Reference [15] have proposed optimal siting and sizing of Distributed Generation (DG) and capacitors in the distribution system to minimize the power losses. Reference [16] have proposed a new approach to optimize the power consumption in the computer center. But there is still a need for improvement in P2P approaches to make it more acceptable and pragmatic for the consumers to adapt.

The work is done to integrate an auction-based game theory approach in the P2P energy trading mechanism in smart grid—this increases cost savings of the prosumers and consumers as compared to only trading with the main energy supplier or utility directly. A forecasting model is also introduced into the system to assess the users in trading energy within the grid. Other aspects that are considered while providing a system for trading are demand, storage capacity and transmission line cost.

2 Methodology

The developed model is designed in a way that enables the prosumers and consumers to trade the energy in the P2P design framework using the auction models as well as with the main energy supplier or utility directly. We have designed a grid with buildings and a main energy supplier to simulate the process. The four basic P2P auction models are studied in the simulation. A forecasting model is also introduced in the smart grid to predict the prices and predict the consumption and production of the buildings. The developed system can be divided into the following modules.

2.1 Grid Design and Energy Trading Mechanism

The simulation contains a grid of buildings of size $10 * 10$. The buildings in the designed grid are placed along the width and height of the grid on a particular location (x, y) . The grid also has a main energy supplier at location $(-1, -1)$. While trading with the utility, the consumers and prosumers calculate their (Manhattan)

distance from the utility. It is assumed that longer distance between the supplier and the buyer is less power efficient, and therefore, costs more money to send power through (line payments). The distance $d(i, j)$, denoting the distance between seller placed at the position (x_i, y_i) and buyer placed at the position (x_j, y_j) , is calculated as follows:

$$d(i, j) = |x_i - x_j| + |y_i - y_j| \quad (1)$$

The buildings calculate its surplus energy based on the production and consumption. If the building has surplus energy after self-consumption, then it stores the energy in the energy storage system that can store up to one hour worth of energy produced. The energy that cannot be stored is sent back to the grid. In the P2P energy trading mechanism, the buildings trade energy with each other. The buildings are divided into buyers and sellers based on the amount of energy stored by the building (E_a). If $E_a < 0$, the buildings are buyers else they are categorized as sellers. In case the building is unable to fulfill its demand from P2P trading then it buys the energy from the grid directly on an average once a day. An auction-based game theoretic approach is designed in this P2P energy trading market to help the prosumers or consumers to trade the energy.

2.2 Auction Mechanism

The four basic auction models considered in this work for the P2P energy trading are—English, Dutch, First Price and Vickrey that have different properties and set of rules for buyers and sellers participating in the auction. The auctions are conducted in an hourly manner based on the auction periods. Each auction model follows a different set of rules in allocation of energy at a price and so the outcome of each auction varies.

Open-Bid Auctions

In the proposed open-bid auctions the P2P energy trading process takes place between buyers and sellers recognizing that no sellers communicate with each other and only communicate with the buyer. The set of sellers in the system is represented as $S = \{S_1, S_2, S_3, S_n\}$ where $S_i \in S$ and each seller contains stored energy that they are willing to sell as $E_a = \{E_{a1}, E_{a2}, E_{a3}, \dots, E_{an}\}$, where $E_{ai} \in E_a$ is the energy amount of seller S_i . Each seller sets its starting price which is denoted as the values $P_s = \{P_1, P_2, P_3, \dots, P_{an}\}$ where $P_i \in P_s$, which is the price of the seller S_i . This max distance D is used to set its maximum starting price or bid that will initiate the auction. The set of buyers is represented as $B = \{B_1, B_2, B_3, B_m\}$ where $B_j \in B$. These private values are denoted as $P_b = \{P_1, P_2, \dots, P_m\}$ where $P_j \in P_b$, which is the private value of the buyer B_j . The distance between a seller S_i and a buyer B_j is calculated using distance formula Eq. 1 and is represented as $d(i, j)$. The distance cost dp is added based on this distance.

English Auction Model

English auction is an open-bid auction. It follows the iterative approach for bidding where the participants make increasingly higher bids. The minimum increment MI_i is calculated based on the initial price and the increment percentage IP which is set to 0.02 that means 2% increase from the original price. A lowering increment is also initialized LI which is considered to lower the price if nobody bids after the start of the auction.

$$MI_i = P'_i * IP, LI = LO_i \quad (2)$$

The set B is iterated, and a buyer say B_j starts the auction and its minimum cost value MC_j is set to its private value. Each seller $S_i \in S$ is at a certain distance $d(i, j)$ from the buyer B_j and so the cost is calculated by the buyer B_j to each of these sellers based on distance and price of that seller that is open to the buyers as:

$$Ci, j = P'_i + (d(i, j) * dp) \quad (3)$$

The buyer can offer the bid to a particular seller S_i only if: $Ci, j < MC_j$

$$Bid = P_i + MI_i \quad (4)$$

The bid is placed by the buyer B_j if $Bid > MB_i$. The process takes place iteratively with each buyer following the same set of rules and the bids are updated and increased in the ascending order following the above procedure. If the higher bid is placed by the new buyer, the previous buyer drops out and the buyer with the highest bid is now present with the seller. The seller returns the bidder with the highest bid after the end of the auction. If there is no bid received after setting the starting bid, the seller decrements its value based on the $LI = LO_i$ as:

$$P'_i = P'_i - LI \quad (5)$$

Dutch Auction Model

Dutch auction is a descending price auction. It follows an open-bid mechanism. A starting bid is set, and the seller decrements the price by the decrementing value until there is a bid from a willing buyer. The first buyer to bid wins the auction and receives the amount of energy at the given price. Each seller in set S sets its maximum starting price or bid that will initiate the auction. The starting bid is now the price of the seller P_i

$$SB_i = P'_i + (D * dp), P_i = SB_i \quad (6)$$

The reserve price is set by the seller and the bidding will end once it reaches this reserve price. The reserve price is the initial set price $RP_i = P'_i$. The lower increment (DP) is set to 0.05 that means 5% decrease from the original price.

$$LI_i = P'_i * DP \quad (7)$$

The buyer B_j checks which seller has not sold the energy yet and calculates the cost based on distance and price from seller S_i as:

$$Ci, j = P'_i + (d(i, j) * dp) \quad (8)$$

The buyer offers the bid to a particular seller S_i only if $Ci, j < MC_j$. The energy is sold by seller S_i to the first buyer who bids after passing the above condition. After each iteration if the energy is not sold the lowered price is compared by the seller with its reserve price and if the price is less than reserve price then the auction end or else the lowered price is now the current price of the seller, and the bidding continues. The buyer again compares its price with this new value and places the bid accordingly.

Sealed Bid Auctions

The buyers in the sealed bid auctions submit their bid in sealed format simultaneously. All bids will be collected by the seller and then the highest bidder wins the auction and pays the price. In this mechanism there is only one iteration where the bids are submitted. As it is a sealed bid here the sellers don't initiate the auction with any starting price or bid. The buyers will follow certain rules and place the bid to the seller. The set of buyers is represented as $B = \{B_1, B_2, B_3, B_m\}$ where $B_j \in B$. As discussed earlier, each buyer has its own independent private value; these private values are denoted as $P_b = \{P_1, P_2, \dots, P_m\}$ where $P_j \in P_b$, which is the private value of the buyer B_j .

First Price Auction Model

The First Price auction is referred to as blind auction. This auction follows a sealed bid mechanism. The winning bidder will have to pay his/her price or placed bid, so the bidding value placed by the buyer is slightly lower than its true value. The expected cost of each seller is calculated by the buyer based on the expected price calculated. This price is used by the buyer to make the decision of making a bid to the seller. The buyer puts its bid to the respective seller in this case seller S_i by using its private value information and considering its distance from the seller as:

$$Bid(i, j) = Pj + (d(i, j) * dp) \quad (9)$$

A multiplier value is calculated randomly based on the bid multiplier value and the final bid value is placed that is less than its private value.

$$Bid = Bid(i, j) * \text{multiplier} \quad (10)$$

This procedure is followed by all the buyers, and they place their bid to the respective sellers accordingly. The highest bid value received is compared by the seller with its initial price and if the bid is greater the buyer placing the highest bid wins and pays its bidding price.

Vickrey Auction Model

The Vickrey auction is the second price auction and follows the sealed bid mechanism. In this auction the winning buyer pays the second highest bid. The process of collection of bids and initializing the auction is the same as first price auction except while bidding the bidder bids its true private value P_j . The seller in the Vickrey auction model returns the buyer bidding the highest bid. The buyer who won the auction will now pay the second highest bid.

After the end of each of the auctions mentioned above based on the amount of energy sold, the total energy amount and balance of buyers and sellers are updated.

3 Result and Discussion

The model is simulated, and a grid is defined for each simulation. The model runs for 100 days per simulation and the buildings trade the energy using both the mechanisms. There are five such grids simulated for 100 days each. For each simulation an auction is selected from the English, Dutch, First price and Vickrey. The forecasting model used in our study is support vector regression (SVR). These auctions are simulated using the forecasting models in the three ways. The buildings can trade without using a forecasting model (M0 model). The building uses a forecasting model to predict prices while trading the energy (M1 model). While placing a bid, each buyer say B_j (as mentioned in the above section) calculates the minimum cost (MC_j) that is its true value on which they are ready to bid and buy the amount of energy auctioned. The SVR model is used here by the buyers to help them calculate this price by considering an hour-ahead predicted price of the grid. The feature in the data is the period of the day and on this basis the dependent variable price is calculated. The building uses the production and consumption prediction model (M2 model) to predict its consumption and production when trading the energy. If a building or prosumer decided to use this model then after the total energy (TE) is calculated, the buildings use the prediction model designed to calculate the predicted consumption and production value for that time of the grid. If there is more amount of energy predicted to be produced, then that is added to the amount of energy stored for trading and if there is more energy predicted to be consumed then that amount is deducted from the stored energy. This final amount of energy calculated will then be used by buildings or prosumers for energy trading in P2P.

When the energy was sold to the utility the buildings or prosumers were the sellers and utility was the buyer. Based on the amount of energy sold following data were collected for all the prosumers or buildings taking part as sellers in the trading of energy to the utility.

- DGS Dollars earned selling to the grid
- LGS Line payment cost selling to the grid
- EGS Energy sold to the grid.

Average grid sell cost for all the buildings is calculated as:

$$GCS = (DGS + LGS)/EGS \quad (11)$$

When the energy was bought from the utility the buildings or prosumers are the buyers and utility is the seller. Based on the amount of energy bought from the grid following data were collected for all the prosumers or buildings taking part as buyers in the trading of energy to the utility.

- DGB Dollars spent buying from the grid
- LGB Line payment cost buying from the grid
- EGB Energy bought from the grid.

Average grid buy cost for all the buildings is calculated as:

$$GCB = (DGB + LGB)/EGB \quad (12)$$

Total Average grid cost can be calculated as:

$$GC = (GCS + GCB)/2 \quad (13)$$

When the energy is traded in the P2P model the sellers with surplus energy sell it to the buyers who need it to fulfill its demand. Based on the amount of energy traded in P2P the following data were collected for all the prosumers or buildings.

- DA Dollars traded in auction
- EA Energy traded in auction
- LA Line payments cost in auction while selling the energy
- TA Total auctions.

Average auction buy cost for all the buildings is calculated as:

$$AB = (DA + LA)/EA \quad (14)$$

Average auction sell cost for all the buildings is calculated as:

$$AS = DA/EA \quad (15)$$

Average auction cost in P2P energy trading is calculated as:

$$AC = (AS + AB)/2 \quad (16)$$

Using the above calculated average values, the following metrics are used to measure the performance of the grids when P2P energy trading was used with respect to direct trading with utility.

3.1 Average Cost % Advantage

The first metric shows the advantage of trading through auctions compared to trading with the grid in terms of average cost in trading the energy (Fig. 1). The metric is calculated as follows:

$$\text{Auction Trading \% advantage : } \alpha A = (GC/AC - 1) * 100 \quad (17)$$

Table 1 represents the advantage of trading using an auction mechanism in terms of average cost of the energy traded. When there was no forecasting model used

Fig. 1 Auction % advantage (metric)

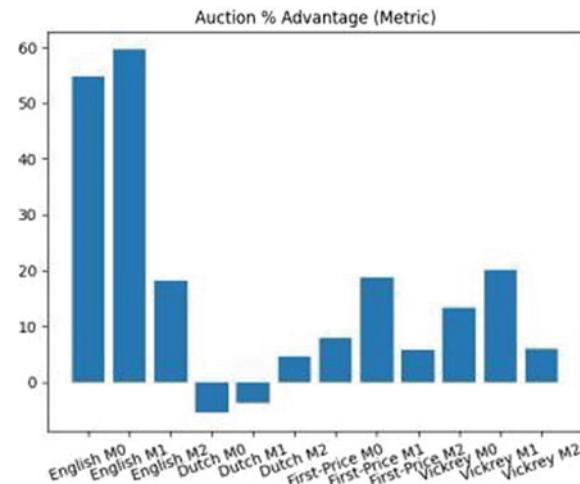


Table 1 Auction trading % advantage

Models auctions	M0	M1	M2
English	54.7954	59.5540	18.1101
Dutch	-5.5392	-3.6322	4.5686
First price	7.8485	18.7264	5.7266
Vickrey	13.3230	20.1658	5.9960

the English auction showed an efficiency in terms of average cost of 54.79% which increased to 59.55% with the integration of the price forecasting model. However, the integration with the energy generation and consumption model leads to decrease in efficiency. The English auction was followed by Vickrey auction that again showed an increase in efficiency in terms of average cost when integrated with model M1. The First price auction had an efficiency of 18.72% which is nearly the same as the Vickrey auction with the M1 model. The Dutch auction performed poorly with both the models M0 and M1, but the efficiency increased when integrated with model M2. This metric takes into consideration the average grid cost of energy and average cost of energy while trading using an auction model and concludes the efficiency in terms of cost saved per kwh.

3.2 *Total \$ Savings*

This metric computes the total amount of dollars saved by auctioning compared to the scenario where the entire trading is done directly with the main energy supplier or utility.

$$\text{Auction Trading \$ savings} : \gamma A = (\text{GC} - \text{AC}) * \text{DA} \quad (18)$$

Cost Savings % or Returns

This metric computes the percent of total dollars saved or made while trading through auctioning compared to if all trading was done directly with the main energy supplier.

Auction Trading % savings (Fig. 2):

$$\beta A = (\gamma A / \text{DGS} + \text{DGB} + \text{DA} + \text{LGS} + \text{LGB} + \text{LA}) * 100 \quad (19)$$

Table 2 of Auction % saving shows that the clear dominant auctioning system is the English auction, followed by First Price, Vickrey, and then Dutch being not great at all. The “Auction % Savings” is the metric that shows the total percentage of dollars traded that was saved by various buildings through trading with auctions instead of trading with the main energy supplier. M1 model integration works best for both First price and Vickrey auction and using M2 model increases the cost savings or returns as compared to without using any forecasting model. In case of sealed bid auctions, we can conclude that Blind (First price) when integrated with model M1 has 0.82% savings whereas with model M2 has 0.63% savings which is greater than the efficiency when traded without using a forecasting model. The Vickrey auction also improved cost savings % when integrated with model M1. The cost savings increased from 0.41% when traded without a forecasting model to 0.71% with model M1 and 0.64% with model M2. In the case of Dutch auctions, the cost savings increase when integrated with model M2. Another interesting thing to note is that Dutch auctions seem to be a very terrible method unless using model type M2. Using this graph,

Fig. 2 Auction % savings (metric)

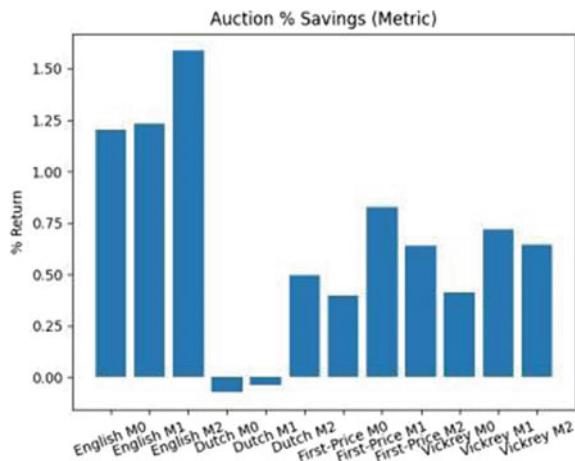


Table 2 Auction trading % saving

Models auctions	M0	M1	M2
English	1.2037	1.2333	1.5869
Dutch	-0.0748	-0.0402	0.4981
First price	0.3968	0.8270	0.6392
Vickrey	0.4115	0.7171	0.6446

we can see that English auction with model type M2 are the overall most efficient auctioning method.

4 Conclusion

This research work was undertaken to design the auction-based approach in the smart grid for energy trading in the P2P model. The advantages of trading in the P2P model using the four different auction models over the direct trading mechanism were compared based on different performance metrics. These results can be used to encourage prosumers to be part of the P2P model over the direct trading with the utility as it is proved to be efficient. Overall, this study strengthens the idea of integration of forecasting models in energy trading using an auction-based game theoretic approach.

References

1. "Research on energy efficiency, CO₂ emissions, energy consumption, forecast." <https://www.enerdata.net/>
2. "Electricity Generation from Renewable Energy." <https://www.nationalgeographic.org/maps/electricity-generation-renewable-energy-sources/>
3. Cai Y, Huang T, Bompard E, Cao Y, Li Y (2017) Self-sustainable community of electricity prosumers in the emerging distribution system. *IEEE Trans Smart Grid* 8(5):2207–2216. <https://doi.org/10.1109/tsg.2016.2518241>
4. Chen S, Shroff NB, Sinha P (2022) Energy trading in the smart grid: from end-user's perspective. Presented at the 2013 Asilomar conference on signals, Systems and computers. Accessed 10 Jun 2022 [Online]. Available <https://doi.org/10.1109/acssc.2013.6810288>
5. "What is electricity trading?" 29 Dec 2020. <https://www.drax.com/power-generation/what-is-electricity-trading/>
6. "Renewable Energy: Distributed Generation Policies and Programs." <https://www.energy.gov/eere/slsc/renewable-energy-distributed-generation-policies-and-programs>
7. O. of Scientific and Technical Information (OSTI), E. Doris, S. Busche, and S. Hockett, "Net Metering Policy Development and Distributed Solar Generation in Minnesota: Overview of Trends in Nationwide Policy Development and Implications of Increasing the Eligible System Size Cap," Dec. 2009. Accessed 10 June 2022 [Online]. Available <https://doi.org/10.2172/969897>
8. Yan Y, Qian Y, Sharif H, Tipper D (2013) A survey on smart grid communication infrastructures: motivations, requirements and challenges. *IEEE Commun Surv Tutorials* 15(1):5–20. <https://doi.org/10.1109/surv.2012.021312.00034>
9. Editor (2019) Smart grid: vision for India. 04 Dec 2019. <https://www.electricalindia.in/smart-grid-vision-for-india/>
10. Agarwal A, Khandeparkar K (2021) Distributing power limits: Mitigating blackout through brownout. *Sustain Energy, Grids Netw* 26:100451, ISSN 2352-4677. <https://doi.org/10.1016/j.segan.2021.100451>
11. Agarwal A, Ramamritham K (2020) A comparison of novel optimization model and algorithm for solving PMU deployment issues in the grid. *Sādhanā* 45:284. <https://doi.org/10.1007/s12046-020-01522-y>
12. Agarwal A, Ramamritham K (2020) Tackling issues related to PMU deployment in the grid using a novel algorithm. In: 2020 IEEE First international conference on smart technologies for power, energy and control (STPEC), 2020, pp 1–6. <https://doi.org/10.1109/STPEC49749.2020.9297698>
13. Jadhav AM, Patne NR, Guerrero JM (2019) A novel approach to neighborhood fair energy trading in a distribution network of multiple microgrid clusters. *IEEE Trans Industr Electron* 66(2):1520–1531. <https://doi.org/10.1109/tie.2018.2815945>
14. Morstyn T, Teytelboym A, Mcculloch MD (2019) Bilateral contract networks for peer-to-peer energy trading. *IEEE Trans Smart Grid* 10(2):2026–2035. <https://doi.org/10.1109/tsg.2017.2786668>
15. Manikanta G, Mani A, Singh HP, Chaturvedi DK (2019) Simultaneous placement and sizing of DG and capacitor to minimize the power losses in radial distribution network. In: Ray K, Sharma T, Rawat S, Saini R, Bandyopadhyay A (eds) Soft computing: theories and applications. Advances in Intelligent Systems and Computing, vol 742. Springer, Singapore. https://doi.org/10.1007/978-981-13-0589-4_56
16. Verma KK, Saxena V (2019) Energy consumption of university data centre in step networks under distributed environment using Floyd–Warshall algorithm. In: Ray K, Sharma T, Rawat S, Saini R, Bandyopadhyay A (eds) Soft computing: theories and applications. Advances in Intelligent Systems and Computing, vol 742. Springer, Singapore. https://doi.org/10.1007/978-981-13-0589-4_10

Design of Disturbance Observer-Based Dynamic Sliding Mode Control



S. S. Nerkar and B. M. Patre

Abstract This paper proposes the design and implementation of observer-based dynamic sliding mode control (DSMC). The designed disturbance observer (DO) estimates the uncertainties and disturbance in an integrated manner. It generates smooth and chattering-free control. The disturbance observer-based DSMC technique is validated through experimentation on DC motor setup. The results show the effectiveness of the combination of the controller–observer design for position control of DC motor against lumped uncertainty. The overall stability of the observer-based control system is proved by Lyapunov theory.

Keywords Dynamic sliding mode control · Disturbance observer · DC motor · Lyapunov theory

1 Introduction

Sliding mode control (SMC) is a robust control technique to counter the presence of uncertainties and disturbances in the system. But the main drawback of the SMC is that it uses the discontinuous control to achieve the control objective. This rapidly changing control actions induce stress and wear in mechanical parts and the system and could damage it. Thus, chattering in the SMC restricts it for the real-life applications. SMC requires that the full state vector to be available for the control to apply effectively. But states may not be available always. In the range of uncertainty, if it cannot be determined or not known exactly, sliding condition may not be satisfied. Many methods have been presented to reduce the chattering like continuous approx-

S. S. Nerkar (✉)

Department of Instrumentation Engineering, Government College of Engineering, Jalgaon, Maharashtra, India

e-mail: sachinnerkar@rediffmail.com

B. M. Patre

Department of Instrumentation Engineering, S.G.G.S. Institute of Engineering and Technology, Nanded, Maharashtra, India

imation by boundary layer technique [1], to use higher order sliding mode control (HOSM) [2]. Dynamic sliding mode control (DSMC) is also designed for chattering elimination, where the control developed by sliding mode is filtered and then applied to the actual plant [3].

The DSMC adopts a special control structure, in which an integrator as a filter is placed in front of the system as depicted in Fig. 1. A sliding mode control w is designed for the augmented system consisting of system and the filter. Being a sliding mode control, the auxiliary control signal w has chattering; however, the actual control signal u applied to the system is smooth since the chattering in w is filtered by integrator. Here, a low-pass filter eliminates the chattering along with maintaining the control accuracy. In case of a system where measurement noise is present, DSMC effectively filters out the chattering due to it. Overall, DSMC provides a systematic methodology to eliminate chattering with the required control accuracy [4–6].

In the design of DSMC, the main problem is to form the sliding surface; as we are designing the SMC for the augmented system, the sliding surface should be the function of the states of the augmented system and not the actual system. Since the augmented system is one dimension larger than the original system, the new sliding variable in DSMC contains an uncertainty term due to the external disturbance and/or parametric uncertainty. And also from the structure of DSMC, one can observe that the lumped uncertainty involves in the augmented system. Evaluation of the new sliding variable in DSMC becomes difficult because of this reason. To overcome the problem of bounds of uncertainty, instead of using the bounds of uncertainty in the control, one can estimate it by using the methods like time delay control (TDC), inertial delay control (IDC) and can use it in the control design. Hence, there is a need to design an observer to estimate the lumped uncertainty for the design of sliding surface in DSMC. Different types of state observers are developed to estimate the states as well as uncertainties in the systems like uncertainty and disturbance estimator (UDE), supertwisting disturbance observer (STDO), inertial delay observer (IDO), etc. [7–12]. In this paper, disturbance observer (DO) is designed for estimation of the lumped uncertainty in the system. The state and uncertainty observers are more useful for the control design of higher order real-time applications where it is difficult to measure all the states using the sensors. The control strategies are verified by the practical experimentation on Quanser's DC servomotor plant under measurement noise and actuator unmodeled dynamics. The main contributions of the work are summarized as follows:

- Design of DO-based DSMC to estimate available system states and uncertainties for achieving smooth control.
- Due to estimation of states and uncertainty, the observer-based DSMC design reduces the need sensors.
- Experimental results validate the efficacy of robustness of control strategies against the presence of disturbance, uncertainties and sensor noise in the system parameters.

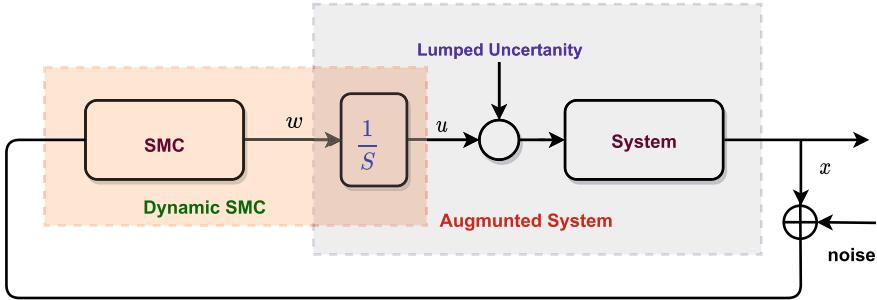


Fig. 1 DSMC design with measurement noise

The paper is organized as follows. Section 2 presents the observer-based design approach with numerical simulation and stability analysis of the observer-based design. Section 3 contains application of the strategy in Sect. 3 to the system to Quanser's DC servomotor plant along with the experimental results and performance analysis. Finally, in Sect. 4, paper is concluded.

2 Control Design

In the design of DSMC, as the lumped uncertainties are included in between the filter and the system, the sliding surface requires the unknown lumped uncertainty in its design. Hence, design of sliding surface in DSMC is a tedious and requires the unknown parameter called lumped uncertainty. Therefore, it is required to design an observer for estimation of lumped uncertainty for the design of sliding surface. For the systems where the states are unavailable for measurement, there is a need to design an observer which estimates the states of the systems. Due to concurrent estimation of states and uncertainty, it reduces the number of sensors and make system robust despite the presence disturbance. To overcome the problem of bounds of uncertainty, instead of using the bounds of uncertainty in the control, one can estimate the uncertainty and disturbance in the system by using DO and can use it in the control design.

Consider the DSMC design for linear time invariant uncertain system with relative degree two

$$\begin{aligned}\dot{x} &= Ax + Bu + Be \\ y &= Cx\end{aligned}\tag{1}$$

where $x \in \mathbb{R}^n$ is state vector, $y \in \mathbb{R}^p$ is output of the system, A, B, C are corresponding matrices of appropriate dimension, $u \in \mathbb{R}^m$ and e are external disturbance and uncertainties to the system called lumped or parametric uncertainty, satisfying the matching conditions given by

$$\begin{aligned}\Delta A &= BD \\ \Delta B &= BE \\ d &= BF\end{aligned}\tag{2}$$

Assumption 1 Uncertainties or external disturbances are varying slowly $\dot{d}(t) \cong 0$, and all the states are available for measurement.

2.1 Sliding Surface

In the design of DSMC, integrator and system form an augmented system, which has one dimension higher than the original system. As a result, the relative degree (r) of the extended system becomes $r + 1$. Hence, for the augmented system, one can choose the sliding variable as

$$\sigma = y'' + \lambda_1 y' + \lambda_0 y\tag{3}$$

where λ_0, λ_1 are the positive integers and

$$\begin{aligned}y &= Cx \\ y' &= CAx \\ y'' &= CA^2x + CABu + CABe\end{aligned}\tag{4}$$

Since y'' cannot be calculated based on only x and u , on account of e being unknown, a robust DO is suggested to estimate term e and thereby construct a sliding variable.

2.2 Disturbance Observer

A DO is proposed for the estimation of state and uncertainty of the auxiliary system. Define an auxiliary state

$$q = \begin{bmatrix} y' \\ y'' \end{bmatrix}\tag{5}$$

where the first component y' is accessible if x is known, but the second component is not accessible, due to the lumped uncertainty in y'' . For this purpose, a robust DO is used to estimate q , so that y'' is made available to evaluate the sliding surface. By taking the derivative of y'' from Eq. (4), one can form the dynamic model of the form

$$\dot{q} = A_2 q + B_2 u_2 + B_2 e^*\tag{6}$$

where e^* is an unknown lumped uncertainty,

$$e^* = CA^2Be + CAB\dot{e} \quad (7)$$

$$u_2 = CA^3x + CA^2Bu + CABw \quad (8)$$

System matrices are given by

$$A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad C_2 = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad (9)$$

The following observer is suggested to estimate the state vector q of the dynamic system formed in Eq. (6)

$$\dot{\hat{q}} = A_2\hat{q} + B_2u_2 + L(y' - C_2\hat{q}) + B_2\hat{e}^* \quad (10)$$

where

$$\dot{\hat{e}}^* = M(y' - C_2\hat{q}) \quad (11)$$

Error in state and uncertainty estimation,

$$\tilde{q} = q - \hat{q} \quad (12)$$

$$\dot{\tilde{e}}^* = e^* - \hat{e}^* \quad (13)$$

Using Eqs. (5), (10) and (11), the error dynamics are obtained as

$$\dot{\tilde{q}} = (A_2 - LC_2)\tilde{q} + B_2\dot{\tilde{e}}^* \quad (14)$$

$$\dot{\tilde{e}}^* = -MC_2\tilde{q} + \dot{e}^* \quad (15)$$

Here, as per Assumption 1, the lumped uncertainty is a slow varying signal, that is,

$$\dot{e}^* \cong 0 \quad (16)$$

then Eq. (15) becomes

$$\dot{\tilde{e}}^* = -MC_2\tilde{q} \quad (17)$$

State space representation of Eqs. (14) and (17) is given as

$$\begin{bmatrix} \dot{\tilde{q}} \\ \dot{\tilde{e}}^* \end{bmatrix} = \begin{bmatrix} (A_2 - LC_2) & B_2 \\ -MC_2 & 0 \end{bmatrix} \begin{bmatrix} \tilde{q} \\ \tilde{e}^* \end{bmatrix} \quad (18)$$

The representation in (18) is converted in the form of the pole placement problem. Find the observer gain matrix K_{ob}

$$\dot{P} = (H - G_1 K_{\text{ob}})P \quad (19)$$

where

$$P = \begin{bmatrix} \tilde{q} \\ \tilde{e}^* \end{bmatrix} H = \begin{bmatrix} A_2 & B_2 \\ 0 & 0 \end{bmatrix} G_1 = \begin{bmatrix} C_2 & 0 \end{bmatrix} K_{\text{ob}} = \begin{bmatrix} L \\ M \end{bmatrix} \quad (20)$$

By proper choice of the eigen values of the matrix $(H - K_{\text{ob}}G_1)$ using the observer gain K_{ob} , the pair (H, G_1) is observable. So, the observer error dynamics are $\tilde{e} \rightarrow 0$ at $t \rightarrow 0$.

From the observer, one can get the state and uncertainty estimate (\hat{q}, \hat{e}^*) . Observer is able to estimate the states and the uncertainty with small value of the estimation error, which can be compensated by the designed control.

The new sliding surface $\hat{\sigma}$ can be obtained by using the estimate of the unknown terms in Eq. (3), and estimate of the sliding surface can be obtained as

$$\hat{\sigma} = \hat{y}'' + \lambda_1 y' + \lambda_0 y \quad (21)$$

The new sliding surface is used to design the sliding surface for the augmented system.

2.3 Disturbance Observer-Based Control Design

From the designed sliding surface, the sliding mode control for the augmented system can be designed, and it is obtained by the equations as

$$u = \int w \, dx \quad (22)$$

The SMC for the augmented system is designed so as to satisfy the sliding condition as

$$\hat{\sigma} \dot{\hat{\sigma}} \leq 0 \quad (23)$$

Differentiating Eq. (21) and assuming that the auxiliary control as

$$w = w_{\text{eq}} + w_n \quad (24)$$

where w_{eq} is nominal control which addresses the known part of the system and w_n is the discontinuous control to compensate the estimation error, then the equation is

$$\dot{\hat{\sigma}} = CA^3x + CA^2Bu + \lambda_1 y'' + \lambda_0 y' + CABw_{\text{eq}} + CABw_n + e^* \quad (25)$$

Select the control for known dynamics as

$$w_{\text{eq}} = -(CAB)^{-1}(CA^3x + CA^2Bu + \lambda_1\hat{y}'' + \lambda_0y' + \hat{e}^* + K\hat{\sigma}) \quad (26)$$

where K is a small positive constant from Eqs. (25) and (26) we get,

$$\dot{\hat{\sigma}} = CABw_n + \tilde{e}^* \quad (27)$$

To satisfy the sliding condition in (23), w_n is selected as

$$w_n = -(CAB)^{-1}K_1 \text{sign}(\hat{\sigma}) \quad (28)$$

where K_1 is the positive constant, and $K_1 > |\tilde{e}^*|$ as it is large, it is to compensate the large estimation errors.

2.4 Numerical Simulation

In order to demonstrate the effectiveness of DO-based DSMC control strategy, a following uncertain second-order plant is considered. It is analyzed through numerical simulation as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u + d \end{aligned} \quad (29)$$

For simulation 20% uncertainties in the system parameters and $\sin(t)$, external disturbances are considered. While the constant parameters are $\lambda_0 = 15$, $\lambda_1 = 5$, $K = 1$ and $K_1 = 10$, the initial states for simulation considered are $[1 \ 0]^T$. Apart from the disturbance, here uniform measurement noise of absolute magnitude 0.01 is added to the above lumped uncertainty in order to verify the effectiveness of the control.

From Fig. 2a, it is observed that the system states not only converge to origin but also achieve desired dynamics despite the added noise in the states. The output of the controller w has chattering as shown in Fig. 2b but input to the actual system u is a smooth control signal as shown in Fig. 2c.

2.5 Stability Analysis

Consider the Lyapunov function as

$$V = \frac{1}{2}\sigma^2 \quad (30)$$

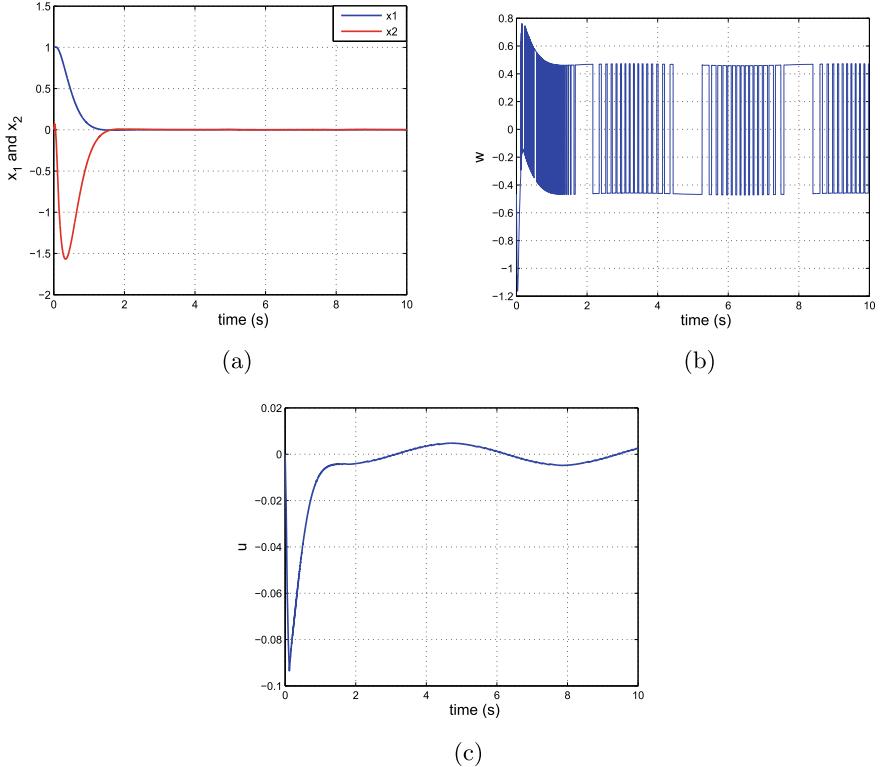


Fig. 2 DO-based DSMC control. **a** Plant states. **b** Auxillary control w . **c** Control u

Taking derivative, we get

$$\dot{V} = \sigma \dot{\sigma} \quad (31)$$

From Eq. (3),

$$\begin{aligned} \dot{\sigma} &= CA^3x + CA^2Bu + CA^2Be + CABw + CAB\dot{e} \\ &\quad + \lambda_1 y'' + \lambda_0 y' \end{aligned} \quad (32)$$

During sliding motion, the control term w_{eq} from Eq. (26) and substituting it in Eq. (32) we get,

$$\begin{aligned} \dot{\sigma} &= CA^3\tilde{x} + CA^2B\tilde{e} + CAB\dot{\tilde{e}} + \lambda_1 \tilde{y}'' \\ &\quad + \lambda_0 \tilde{y}' - K\hat{\sigma} + CABw_n \end{aligned} \quad (33)$$

The observer error \tilde{x} converges to zero asymptotically, if K_{ob} is chosen appropriately.

Then, $\dot{\sigma}$ is

$$\dot{\sigma} = CABw_n - K\hat{\sigma} + \tilde{e}^* \quad (34)$$

So, Eq. (31) becomes

$$\dot{V} = \sigma\tilde{e}^* - K\sigma\hat{\sigma} - \sigma CAB(CAB)^{-1}K_1 CAB\text{sign}(\hat{\sigma}) \quad (35)$$

$$\begin{aligned} \dot{V} &= \sigma\tilde{e}^* - K\sigma\hat{\sigma} - \sigma K_1 CAB \frac{|\hat{\sigma}|}{CAB\hat{\sigma}} \\ \dot{V} &\leq |\sigma| \left(|\tilde{e}^*| - K |\hat{\sigma}| - K_1 \frac{|\hat{\sigma}|}{|\hat{\sigma}|} \right) \\ \dot{V} &\leq |\sigma| (|\tilde{e}^*| - K |\hat{\sigma}| - K_1) \end{aligned} \quad (36)$$

As $K_1 > |\tilde{e}^*|$, K_1 and K are both positive constants, \dot{V} is negative definite. Thus, the sliding manifold is stable.

3 Application of Observer-Based DSMC Control for Position Control of DC Motor

The real-time implementation of the observer-based DSMC strategies discussed in this section has been done for position control of the Quanser DC motor setup. The block diagram of observer-based DSMC control of the DC motor is as in Fig. 3. It has potentiometer, encoder and tachometer. The potentiometer and encoder sensors measure the angular position of the load gear, and tachometer is used to measure its velocity. The different parameters of DC motor are considered as mentioned in Quanser setup manual. The constant parameters are $\lambda_0 = 15$, $\lambda_1 = 7$, $K = 1$ and $K_1 = 20$. Here, uniform measurement noise as a sensor noise of absolute magni-

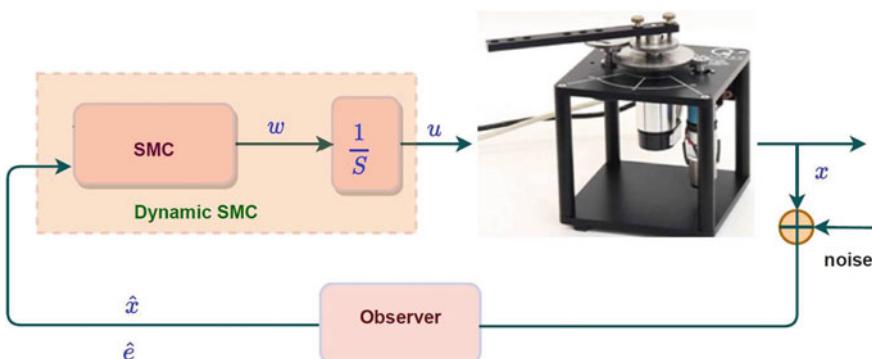


Fig. 3 Observer-based dynamic sliding mode control of DC motor with measurement noise

tude 0.01 is added. The observer poles are selected at $[-20 \ -25 \ -30]$. The initial conditions for states are considered as $[0 \ 0]^T$ and $[1 \ 0]^T$, respectively, for the consideration. The different parameters of DC servomotor are considered as mentioned in Quanser setup manual [13, 14]. Experimentation has been done under different considerations like with and without noise for both the observer-based DSMC strategies. The experimental results validate efficacy of the design for position control of DC motor as depicted in Figs. 4 and 5.

3.1 Results and Discussion

The results in Fig. 4 are taken from the encoder for output without any measuring noise. Figure 4a, b shows the states of the system. The second state is not smooth because it is affected by lumped uncertainty while the observer estimates the states precisely. Comparing the states of the observer and the system, one can observe that the observer has followed the model dynamics. Figure 4c shows the chattering in the auxiliary control w while the actual control to the system u is smooth control and not harmful to the actuator shown in Fig. 4d. The sliding surface σ is shown in Fig. 4e.

Next, in addition to actuator unmodeled dynamics due to friction in the system, here the noisy output is taken from the potentiometer; that is, the control is tested for the system with measurement noise. The results are shown in Fig. 5. As the output from the potentiometer is noisy, the second state calculated consists of chattering as shown in Fig. 5b and the observer state is smooth at the same time. Due to it, the auxiliary control is effected as shown in Fig. 5c but the actual control driving the system is smooth signal as shown in Fig. 5d. The sliding surface σ is shown in Fig. 5e.

4 Conclusion

In this paper, disturbance estimation-based DSMC strategy is proposed for position control of DC motor. The effectiveness of the proposed schemes has been validated in the presence of uncertainty due to Coulomb's friction, unmodeled dynamics and with measurement noise. DO-based DSMC estimates the system states as well as the uncertainty. Due to the simultaneous estimation of states and uncertainty, it reduces the number of sensors and robustify the control despite unknown disturbance and uncertainties in the system parameters or load changes. DSMC is made more robust by augmenting a generalized disturbance observer. The simulation and experimental results validate the robustness of the proposed DO-based DSMC technique. The essential boundness of lumped uncertainty estimation error and sliding manifold is proved by Lyapunov theory.

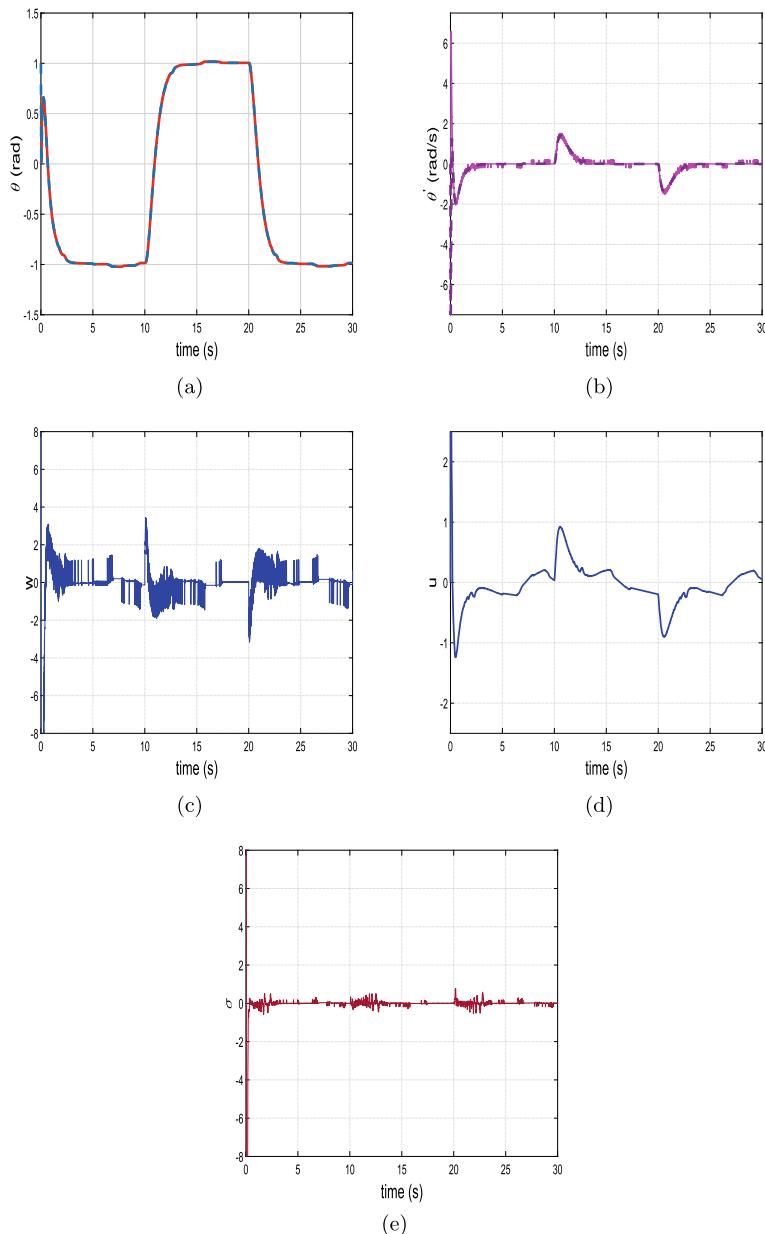


Fig. 4 DO-based DSMC control of DC motor without measuring noise. **a** Actual position of shaft and its estimation: state x_1 . **b** Actual speed of shaft and its estimation: state x_2 . **c** Auxillary control w . **d** Control u . **e** Sliding surface σ

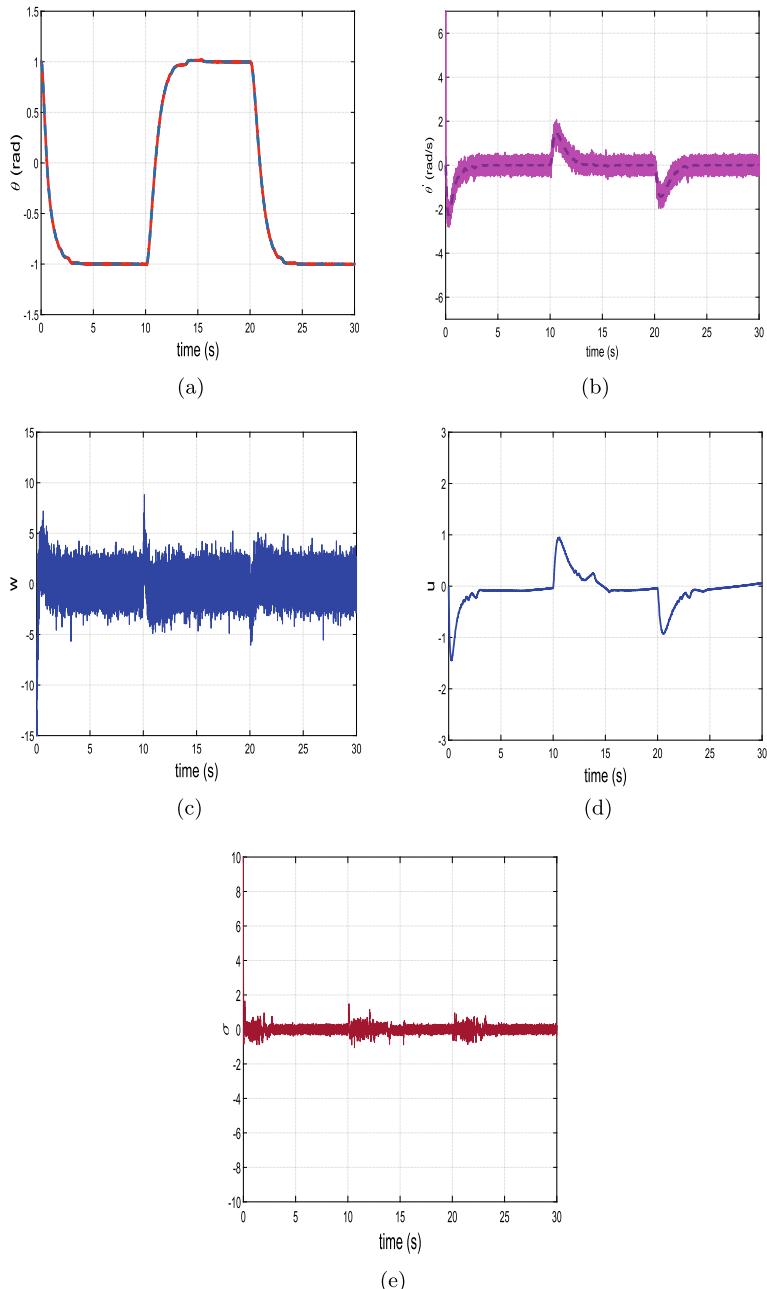


Fig. 5 DO-based DSMC control of DC motor with measuring noise. **a** Actual position of shaft and its estimation: state x_1 . **b** Actual speed of shaft and its estimation: state x_2 . **c** Auxillary control w . **d** Control u . **e** Sliding surface σ

References

1. Chen M-S, Hwang Y-R, Tomizuka M (2002) A state-dependent boundary layer design for sliding mode control. *IEEE Trans Autom Control* 47(10):1677–1681
2. Levant A (2003) Higher-order sliding modes, differentiation and output-feedback control. *Int J Control* 76(9–10):924–941
3. Koshkouei AJ, Burnham KJ, Zinober AS (2005) Dynamic sliding mode control design. *IEE Proc Control Theory Appl* 152(4):392–396
4. Shokohninia MR, Fateh MM, Gholipour R (2020) Design of an adaptive dynamic sliding mode control approach for robotic systems via uncertainty estimators with exponential convergence rate. *SN Appl Sci* 2(2):1–11
5. Chen M-S, Chen C-H, Yang F-Y (2007) An LTR-observer-based dynamic sliding mode control for chattering reduction. *Automatica* 43(6):1111–1116
6. Sira-Ramirez H (1993) On the dynamical sliding mode control of nonlinear systems. *Int J Control* 57(5):1039–1061
7. Sankeshwari SS, Chille RH (2020) Performance analysis of disturbance estimation techniques for robust position control of DC motor. *Int J Control Autom Syst* 18(2):486–494
8. Kadam A, Ray D, Shimjith S, Shendge P, Phadke S (2013) Time delay controller combined with sliding mode for DC motor position control: experimental validation on Quanser QET. In: 2013 international conference on power, energy and control (ICPEC). IEEE, pp 449–453
9. Nerkar S, Londhe P, Patre B (2021) Design of super twisting disturbance observer based control for autonomous underwater vehicle. *Int J Dyn Control* 1–17
10. Patil A, Ginoja D, Shendge P, Phadke S (2015) Uncertainty-estimation-based approach to antilock braking systems. *IEEE Trans Veh Technol* 65(3):1171–1185
11. Ginoja DL, Patel TR, Shendge P, Phadke S (2011) Design and hardware implementation of model following sliding mode control with inertial delay observer for uncertain systems. In: 2011 3rd international conference on electronics computer technology, vol 3. IEEE, pp 192–196
12. Pawar SN, Patre B et al (2021) Extended state observer based robust sliding mode control for fourth order nonlinear systems with experimental validation. *Int J Dyn Control* 1–12
13. Introduction to QUARC 2.0 & SRV 02 and instructor manual. Quanser Inc., Markham, ON, Canada
14. Rotary servo plant SRV 02 user manual, set up and configuration. Quanser Inc., Markham, Canada

Analysis on the Financial Performance of Technology Companies in Malaysia with VIKOR Model



Weng Siew Lam, Kah Fai Liew, Weng Hoe Lam,
and Mohd Abidin Bin Bakar

Abstract Technology companies play a crucial role in contributing to the development of economy in Malaysia. A systematic and rational approach is useful in measuring the technology companies' financial performance (FP). The aim of this paper is to analyze the FP and ranking of Malaysia's technology companies in terms of financial ratios with VIKOR model. The outcomes of the paper demonstrate that the top four outstanding technology companies consist of Malaysian Pacific Industries Berhad (MPI), Globetronics Technology Berhad (GTRONIC), Elsoft Research Berhad (ELSOFT) and ECS ICT Berhad (ECS). The study is significant to analyze the FP and ranking of Malaysia's technology companies with the proposed VIKOR model.

Keywords VIKOR · Technology company · Financial performance · Ranking

1 Introduction

Technology is an important sector that directs the investment and the economy of a country. In the context of the Fourth Industrial Revolution (4IR), economic growth and technological change in the drive for innovation have been greatly emphasized nowadays. Moreover, the financial performance (FP) of technology companies becomes an important indicator for the investors in decision making process. VlseKriterijuska Optimizacija I Komoromisno Resenje (VIKOR) is a decision tool which assists to tackle the multi-criteria decision making (MCDM) problems. The financial variables are crucial in measuring the FP of the companies [1]. Financial ratios have been studied in the MCDM problems to evaluate the FP of the companies. [2–4]. VIKOR model relies on aggregating function that represents closeness to the ideal [5]. In this study, VIKOR aims to identify the technology company that is longest to the worst ideal solution (WIS) and nearest to the best ideal solution (BIS) in terms of FP measurement. A number of applications from various disciplines

W. S. Lam · K. F. Liew · W. H. Lam (✉) · M. A. B. Bakar

Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Kampar Campus, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
e-mail: whlam@utar.edu.my

such as supplier selection [6, 7], academic performance [8, 9], facility location [10], personnel selection [11, 12] and water resources planning [13] have been carried out using the VIKOR model. According to past studies, the authors have analyzed the FP of the companies with VIKOR model [14, 15]. Lu et al. have used the VIKOR model to measure the FP and implement effective sustainable development strategies to enhance competitive advantages among the thin-film transistor liquid crystal display companies in Taiwan [14]. Zhao et al. have evaluated the FP of the banking companies in China by using VIKOR model. Performing analysis is important in order to resolve the problem by making an optimal solution [16–18]. Since the FP of the technology companies is very crucial to be investigated due to its significant contribution to the economy of a country and hence, an appropriate MCDM tool should be proposed to carry out the analysis. Therefore, the purpose of this paper is to analyze the FP of Malaysia's listed technology companies with VIKOR model.

2 Data and Methodology

In this study, the data is gathered from the companies' financial annual reports between the year 2015 and 2017 [19]. The conceptual framework is proposed and presented in Table 1 to measure the FP of Malaysia's technology companies with VIKOR model.

As shown in Table 1, the main purpose of this paper is to analyze the FP of Malaysia's technology companies. The financial data such as ROA, ROE, EPS, DAR, DER and CR are obtained for the listed technology companies in Malaysia Main Market which represent the overall performance of technology sector in Malaysia based on the period of study. FP is defined as the company's capability in controlling

Table 1 Proposed conceptual framework

Objective	Evaluation on the FP of technology companies
Decision criteria	Current ratio (CR)
	Debt to assets ratio (DAR)
	Debt to equity ratio (DER)
	Earnings per share (EPS)
	Return on asset (ROA)
	Return on equity (ROE)
Decision alternatives	ECS ICT Berhad (ECS), Excel Force MSC Berhad (EFORCE), Elsoft Research Berhad (ELSOFT), Grand-Flo Berhad (GRANFLO), Globetronics Technology Berhad (GTRONIC), Inari Amertron Berhad (INARI), JCY International Berhad (JCY), KESM Industries Berhad (KESM), Malaysian Pacific Industries Berhad (MPI), Theta Edge Berhad (THETA), Unisem (M) Berhad (UNISEM), ViTrox Corporation Berhad (VITROX), Willowglen MSC Berhad (WILLOW)

and managing its own resources. The financial variables such as ROA, ROE, EPS, DAR, DER and CR are the significant indicators in assessing the FP of the companies [2, 20]. In this study, the financial variables that seek minimization are DER and DAR. In contrast, CR, EPS, ROE and ROA are sought for maximization. Equal weight is assigned to each financial variable since these six financial variables are equally crucial in measuring the FP of technology companies [1, 2].

3 VIKOR Model

VIKOR model is utilized to identify the decision alternatives' ranking and find the compromise solution that is the nearest to the BIS and longest from the WIS. In this paper, the proposed VIKOR model is to resolve the problem with non-commensurable decision variables [5, 21]. The values of ROE and ROA are in percentage, whereas the rest of the financial variables are in ratio. The VIKOR model's steps are displayed below [12, 13].

Step 1: Develop a decision matrix [22–24].

$$D = \begin{bmatrix} x_{11} & \cdots & x_{1n} & \cdots & x_{1j} \\ \vdots & & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} & \cdots & x_{mj} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{in} & \cdots & x_{ij} \end{bmatrix} \quad (1)$$

where x_{mn} , $m = 1, \dots, i$, $n = 1, \dots, j$. Number of decision alternatives is denoted by i . Number of decision criteria is denoted by j .

Step 2: Identify the best value (f_n^*) and the worst value (f_n^-) of criterion n (financial variable), where $n = 1, 2, \dots, j$.

Step 3: Calculate the score for alternative m with criterion n for (S_{mn}) for $n = 1, \dots, j$, $m = 1, \dots, i$. The value of criterion n for alternative m is denoted by f_{mn} .

$$S_{mn} = \frac{w_n(f_n^* - f_{mn})}{(f_n^* - f_n^-)}, m = 1, \dots, i, n = 1, \dots, j \quad (2)$$

where the weight of criterion n is denoted by w_n .

Step 4: Compute the values of Utility, Regret and VIKOR indices, where Utility, Regret and VIKOR indices are denoted by S_m , R_m and Q_m , respectively, for $m = 1, \dots, i$. The strategy of maximum group utility's weight is denoted by v , whereas $1 - v$ is the individual regret's weight. These strategies could be compromised when v is set to be 0.5.

$$S_m = \sum_{n=1}^j \frac{w_n(f_n * - f_{mn})}{(f_n * - f_n^-)}, m = 1, \dots, i \quad (3)$$

$$R_m = \max \frac{w_n(f_n * - f_{mn})}{(f_n * - f_n^-)}, m = 1, \dots, i \quad (4)$$

$$Q_m = v \frac{(S_m - S^*)}{(S^- - S^*)} + (1 - v) \frac{(R_m - R^*)}{(R^- - R^*)} \quad (5)$$

where

$$R^* = \min(R_m, m = 1, \dots, i)$$

$$R^- = \max(R_m, m = 1, \dots, i)$$

$$S^* = \min(S_m, m = 1, \dots, i)$$

$$S^- = \max(S_m, m = 1, \dots, i)$$

R^- is the maximum value of R_m .

R^* is the minimum value of R_m .

S^- is the maximum value of S_m .

S^* is the minimum value of S_m .

Step 5: Determine the companies' ranking based on the values of Q [11–13]. Choose the best company based on Q value which is the smallest.

4 Empirical Results

Table 2 demonstrates the decision matrix.

Table 3 demonstrates the financial variables' worst f_n^- and the best f_n^* values.

According to the third step of VIKOR model, the normalized decision matrix is shown in Table 4.

Table 5 presents the values of Q_m , R_m , S_m and ranking of the companies. The values of S_m , R_m , Q_m are identified by using Eqs. (3), (4) and (5), respectively. Besides, the values of S^* , S^- , R^* and R^- have also been identified. The parameter to compare the FP of the companies is based on the values of Q for each company.

Based on Table 5, $S^- = 0.9890$, $S^* = 0.2188$, $R^- = 0.1667$ and $R^* = 0.1170$. According to past studies, v = strategy of maximum group utility's weight = 0.5 [25, 26]. In this research study, Malaysia's listed technology companies are analyzed and evaluated in terms of the important financial variables. According to VIKOR model, the company with the smallest Q value is identified as an ideal company. In this paper, MPI achieves the smallest Q value among the technology companies. Hence, MPI obtains the first ranking, followed by GTRONIC, ELSOFT, ECS, INARI, WILLOW, KESM, JCY, VITROX, GRANFLO, EFORCE, UNISEM and lastly THETA. In summary, MPI is classified as an ideal technology company in terms of FP.

Table 2 The decision matrix

Companies	CR	DAR	DER	EPS	ROA	ROE
ECS	37.1934	0.0090	0.0091	0.0733	11.0390	11.1389
EFORCE	8.0551	0.1458	0.1713	0.0316	12.2475	14.3816
ELSOFT	33.2180	0.0091	0.0092	0.0743	25.7115	25.9944
GRANFLO	9.8808	0.0577	0.0628	0.0204	9.7813	10.2015
GTRONIC	47.0061	0.0039	0.0039	0.2154	33.6186	33.7477
INARI	17.2810	0.0549	0.0585	0.1535	24.1869	25.5491
JCY	6.8193	0.0374	0.0389	0.0517	15.1802	15.7579
KESM	6.4405	0.0800	0.0874	0.4894	11.1674	12.1240
MPI	338.8841	0.0013	0.0013	0.4928	15.4754	15.4955
THETA	1.9147	0.1838	0.2425	0.0295	3.5391	4.1004
UNISEM	2.1277	0.1367	0.1586	0.1072	6.4622	7.4932
VITROX	4.6748	0.1927	0.2388	0.1088	34.2040	42.3630
WILLOW	62.8538	0.0067	0.0068	0.0319	14.1332	14.2288

Table 3 The financial variables' worst f_n^- and the best f_n^* values

Criteria	Worst (f_n^-)	Best (f_n^*)
CR	1.9147	338.8841
DAR	0.1927	0.0013
DER	0.2425	0.0013
EPS	0.0204	0.4928
ROA	3.5391	34.2040
ROE	4.1004	42.3630

Table 4 The normalized decision matrix

Companies	CR	DAR	DER	EPS	ROA	ROE
ECS	0.1492	0.0067	0.0054	0.1480	0.1259	0.1360
EFORCE	0.1636	0.1258	0.1175	0.1627	0.1193	0.1219
ELSOFT	0.1512	0.0068	0.0055	0.1476	0.0462	0.0713
GRANFLO	0.1627	0.0491	0.0425	0.1667	0.1327	0.1401
GTRONIC	0.1444	0.0023	0.0018	0.0979	0.0032	0.0375
INARI	0.1591	0.0467	0.0395	0.1197	0.0544	0.0732
JCY	0.1642	0.0314	0.0260	0.1556	0.1034	0.1159
KESM	0.1644	0.0686	0.0595	0.0012	0.1252	0.1317
MPI	0.0000	0.0000	0.0000	0.0000	0.1018	0.1170
THETA	0.1667	0.1589	0.1667	0.1634	0.1667	0.1667
UNISEM	0.1666	0.1179	0.1087	0.1360	0.1508	0.1519
VITROX	0.1653	0.1667	0.1641	0.1355	0.0000	0.0000
WILLOW	0.1365	0.0047	0.0038	0.1626	0.1091	0.1225

Table 5 Scores and ranking of technology companies

Companies	S_m	R_m	Q_m	Ranking
ECS	0.5712	0.1492	0.5530	4
EFORCE	0.8109	0.1636	0.8538	11
ELSOFT	0.4286	0.1512	0.4802	3
GRANFLO	0.6939	0.1667	0.8084	10
GTRONIC	0.2870	0.1444	0.3196	2
INARI	0.4926	0.1591	0.6012	5
JCY	0.5966	0.1642	0.7208	8
KESM	0.5507	0.1644	0.6929	7
MPI	0.2188	0.1170	0.0000	1
THETA	0.9890	0.1667	1.0000	13
UNISEM	0.8318	0.1666	0.8969	12
VITROX	0.6315	0.1667	0.7679	9
WILLOW	0.5393	0.1626	0.6670	6

5 Conclusion

In this paper, VIKOR model is proposed to analyze the FP of Malaysia's listed technology companies with respect to various imperative financial variables. This study shows that MPI is classified as an ideal technology company in terms of FP, followed by GTRONIC, ELSOFT, ECS, INARI, WILLOW, KESM, JCY, VITROX, GRANFLO, EFORCE, UNISEM and lastly THETA. This study makes a contribution by proposing a conceptual framework with VIKOR model to measure the technology companies' FP based on multiple financial variables. The results of this study provide insights on potential improvement for the technology companies based on the scores and ranking with VIKOR model.

References

1. Bulgurcu BK (2012) Application of TOPSIS technique for financial performance evaluation of technology firms in Istanbul Stock Exchange Market. *Procedia Soc Behav Sci* 62:1033–1040
2. Gundogdu A (2015) Measurement of financial performance using TOPSIS method for foreign banks of established in Turkey between 2003–2013 years. *Int J Bus Soc Sci* 6(1):139–151
3. Lam WS, Liew KF, Lam WH (2018) Investigation on the efficiency of financial companies in Malaysia with data envelopment analysis model. *J Phys: Conf Ser* 995(1):012021
4. Liew KF, Lam WS, Lam WH (2017) An empirical evaluation on the efficiency of the companies in Malaysia with data envelopment analysis model. *Adv Sci Lett* 23(9):8264–8267
5. Opricovic S, Tzeng GH (2004) Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *Eur J Oper Res* 156(2):445–455
6. Mirahmadi N, Teimoury E (2012) A fuzzy VIKOR model for supplier selection and evaluation: case of EMERSUN Company. *J Basic Appl Sci Res* 2(5):5272–5287

7. Shemshadi A, Shirazi H, Toreihi M, Tarokh MJ (2011) A fuzzy VIKOR method for supplier selection based on entropy measure for objective weighting. *Expert Syst Appl* 38(10):12160–12167
8. Ummaheswari A, Kumari P (2014) Fuzzy TOPSIS and fuzzy VIKOR methods using the triangular fuzzy hesitant sets. *Int J Comput Sci Eng Inf Technol Res* 4(3):15–24
9. Musani S, Jemain AA (2015) Ranking schools' academic performance using a fuzzy VIKOR. *J Phys: Conf Ser* 622(1):1–10
10. Guzel D, Erdal H (2015) A comparative assessment of facility location problem via fuzzy TOPSIS and fuzzy VIKOR: a case study on security. *Int J Bus Soc Res* 5(5):49–61
11. Alguliyev RM, Alguliyev RM, Mahmudova RS (2015) Multicriteria personnel selection by the modified fuzzy VIKOR method. *Sci World J*, pp 1–16
12. Salehi K (2016) An integrated approach of fuzzy AHP and fuzzy VIKOR for personnel selection problem. *Glob J Manag Stud Res* 3(3):89–95
13. Opricovic S (2011) Fuzzy VIKOR with an application to water resources planning. *Expert Syst Appl* 38(10):12983–12990
14. Lu IY, Kuo T, Lin TS, Tzeng GH, Huang SL (2016) Multicriteria decision analysis to develop effective sustainable development strategies for enhancing competitive advantages: case of the TFT-LCD industry in Taiwan. *Sustainability* 8(7):646–676
15. Zhao Q, Tsai PH, Wang JL (2019) Improving financial service innovation strategies for enhancing China's banking industry competitive advantage during the Fintech revolution: a hybrid MCDM model. *Sustainability* 11(5):1419–1447
16. Kumar S, Mathur YP (2022) Optimal location of pumping station to minimize the maximum cover depth of sewerage system. In *Soft Comput: Theo Appl* 425:69–79
17. Garg P, Chauhan Gonder SS, Singh D (2022) Hybrid crossover operator in genetic algorithm for solving N-Queens problem. In *Soft Comput: Theo Appl* 425:91–99
18. Sreelakshmy K, Gupta H, Ansari IA, Sharma S, Goyal KK, Verma OP (2022) Metaheuristic optimization for three dimensional path planning of UAV. In *Soft Comput: Theo Appl* 425:791–802
19. Bursa Malaysia. Company Announcements | Bursa Malaysia Market. Retrieved from <http://www.bursamalaysia.com/market/listed-companies/company-announcements/#/?category=all> (n.d.)
20. Akkoc S, Vatansever K (2013) Fuzzy performance evaluation with AHP and TOPSIS methods: Evidence from Turkish banking sector after the global financial crisis. *Eurasian J Bus Econ* 6(11):53–74
21. Opricovic S (1998) Multi-criteria optimization of civil engineering systems. Faculty of Civil Engineering, Belgrade
22. Percin S, Aldalou E (2018) Financial performance evaluation of Turkish airline companies using integrated fuzzy AHP fuzzy TOPSIS model. *Int J Econ Admin Stud*, 583–598
23. Kuo MS, Tzeng GH, Huang WC (2007) Group decision-making based on concepts of ideal and anti-ideal points in a fuzzy environment. *Math Comput Model* 45(3–4):324–339
24. Sun CC (2010) A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods. *Expert Syst Appl* 37(12):7745–7754
25. Arabameri A, Lee S, Tiefenbacher JP, Ngo PTT (2020) Novel ensemble of MCDM-artificial intelligence techniques for groundwater-potential mapping in arid and semi-arid regions (Iran). *Remote Sens* 12(3):490–516
26. Sałabun W, Wątrowski J, Shekhovtsov A (2020) Are MCDA methods benchmarkable? A comparative study of TOPSIS, VIKOR, COPRAS, and PROMETHEE II methods. *Symmetry* 12(9):1549–1604

IoT-Based Online Condition Monitoring and Fault Analysis of Bearings of a Rotating Machinery



Sudarsan Sahoo, Chokka Upendra, Krishnananda Sahu, Nabajit Bharali, and Suresh Nuthalapati

Abstract In this era of Internet of Things (IoT), the world has become a futuristic timeline where objects are connected and communicating with each other across a global communication. The devices generate enormous amount of data and Internet makes the possibilities for interaction between devices is unlimited. The proposed work is to monitor the condition of a bearing of a rotating machine online by using IoT and to diagnose the faults by using signal processing techniques. The proposed work using IoT is designed for monitoring the motor condition at a particular place and make the information visible anywhere on the earth. Here the things are referred to be the sensors and instruments which interacts with the IoT. The proposed method deals with monitoring the bearing conditions by monitoring the measuring parameters that is vibration, sound and RPM by using sensors and send the information to the web server and then plot the sensors data as graphical statistics and also reacting on Twitter accordingly. The data updated from the implemented system can be accessible in the Internet anywhere on the earth. The data acquired from the web server is analyzed by using the frequency (FFT) and time–frequency (wavelet) analysis in MATLAB tool to diagnose the defect in the bearing. In the analysis stage the vibration data acquired from the defective bearing is compared with the vibration data of the healthy bearing to detect the defect. The proposed work is a less expensive solution and the result shows that it is a reliable solution in providing online monitoring of bearing conditions.

Keywords Condition monitoring · Bearing · Vibration · FFT · Wavelet transform · IoT

S. Sahoo (✉) · C. Upendra · K. Sahu · N. Bharali
NIT Silchar, Silchar, Assam, India
e-mail: sudarsan_iisc@yahoo.in

S. Nuthalapati
Technische University at Dresden, 01062 Dresden, Germany

1 Introduction

In online condition monitoring of bearing of a rotating machinery, the measurements taken includes acceleration, RPM and sounds [1, 2]. The vibration of the rotating bearing is measured by using the accelerometer and sound is measured using microphone sensor. Vibration signal processing is one tool that has been used for fault diagnosis [3]. In some literatures time analysis of the vibration data is used for the fault diagnosis in rotating machinery [4]. Sukhjeet Singh et al. shows the use of current signature analysis to detect the fault [5]. The wavelet analysis of vibration signal is used for the fault diagnosis and can be used to detect the fault in rolling element bearings [6]. Kankar PK et al. shows the use of continuous wavelet transform (CWT) to diagnose the fault in the ball bearings [7]. Many past works show the use of IoT for online monitoring of machines and physiological parameters [8–10]. The condition monitoring of other machinery can be found in many literatures [11, 12].

The present work is done in two stages. In first stage the vibration and sound generated by the bearing under test and the RPM of the bearing under test is sensed by the sensors which are attached in the vicinity of the bearing under test and the acquired data is sent to the ESP8266 Wi-Fi module which processes the data and sent it to the cloud server for monitoring the values of vibration and sound. In second stage by using the vibration data, the analysis is done using the frequency and time–frequency analysis to diagnose the faults in the bearings.

2 Theoretical Framework and Technologies Used

Given that the central focus of the proposed work will be to adapt an activity to be carried out in a different way, it will be necessary to propose several parameters that serve as conceptual axes on which to support the interpretive reading of the corpus. Although there are many new technologies that provide us with never before accessible information, the elements of rotating machinery should have some tools which require less human intervention.

2.1 *Internet of Things*

The electronics and physical devices can be connected to each other through IoT. The IoT is the network of physical devices and other items embedded with electronics, software, sensors, actuators and network connectivity which enable these objects to connect and exchange data.

2.2 *ESP8266 Wi-Fi Module*

The ESP8266 is a Wi-Fi chip compatible with the IoT environment and is used along with the sensor and processor to send data to the cloud server.

2.3 *ThingSpeak*

ThingSpeak is an Application Programming Interface which is used in IoT system for storing and retrieving data. The measuring data and information can be stored and retrieved whenever necessary. The ThingSpeak can be used with MATLAB to plot the graph from the retrieved data.

2.4 *Sensors*

The system consists of IR sensor (RPM monitoring), sound sensor (Sound monitoring) and accelerometer (Vibration monitoring). These sensors will measure the primary factors, respectively. Each sensor generates the analog output voltage with respect to the physical measuring signal given at its input.

2.5 *Types of Faults in Bearing*

The rolling element bearing consists of three main parts. These are the outer race, the inner race and the balls. The different faults that may occur in bearings can be:

- The outer race defect
- The inner race defect
- The ball defect.

In the present experimental work, the bearing having outer race defect has been taken into consideration to detect and diagnose the faults.

2.6 *Fault Analysis Techniques*

The condition monitoring of the rolling element bearing means to monitor the measuring parameters like vibration, current and sound generated by the bearings. The following signal analysis tools can be used to detect the faults in the bearings:

- Frequency Domain Analysis
- Time–Frequency Domain Analysis.

2.7 Twitter API

The Twitter API stands for ‘Twitter Application Programming Interface’ where real world is communicated to visual world using Twitter to react. Here ThingTweet app is used to link a Twitter account to the ThingSpeak account. The system can then send alerts via Twitter by using the Tweet control application.

The condition has been created that if vibrations are more than the maximum threshold value then ‘Tweet’ will be generated accordingly to react in Twitter Social platform. This application is used to share the information to the external world instantaneously so that accordingly action can be performed.

3 Experimental Set-Up and Procedure

See Figs. 1 and 2.

The experimental set-up is shown in Fig. 3 and the bearing used in this experimental work is shown in Fig. 4. The experimental set-up consists of a motor, the bearing under test and a mechanical alignment. The motor is used to drive the bearing. The ideal location for keeping the device is the top of the bearing where it should be mounted. The sensors can provide a very good reading as they are attached to the body of the host. Therefore, the sensors and the IoT device are installed on the top of the set-up to acquire the measuring data. The acquired measuring data that is vibration, sound and RPM are sent to the cloud server through the Wi-Fi device.

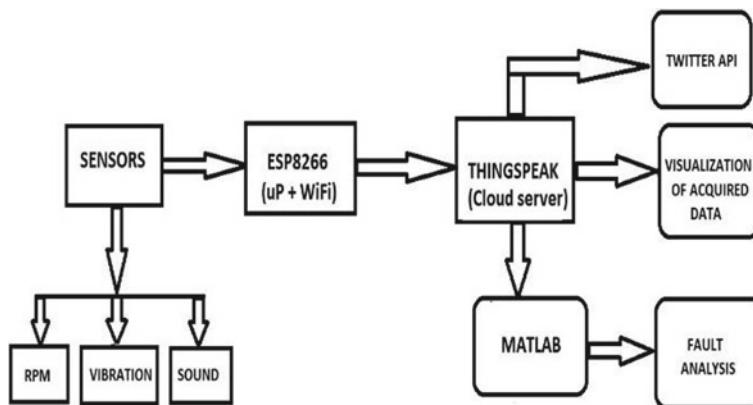


Fig. 1 Block diagram of the proposed model

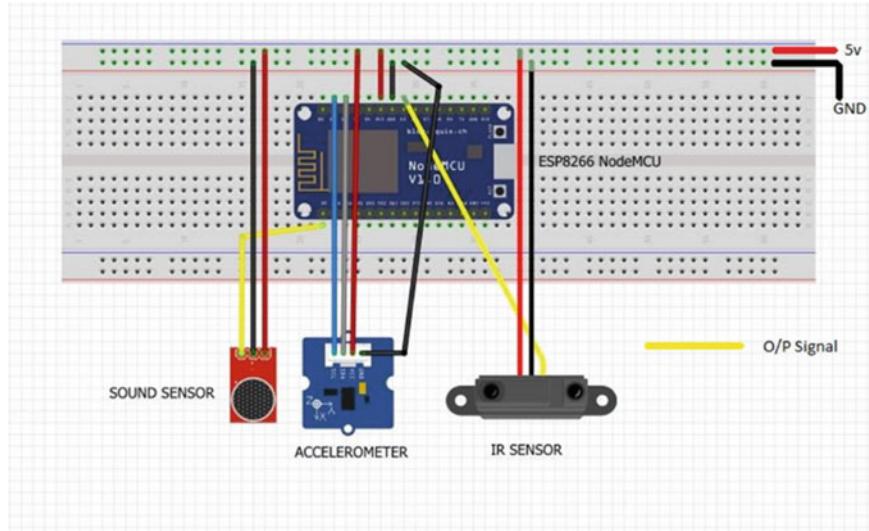


Fig. 2 Schematic diagram of the circuit connection on breadboard

Then the measuring data from the cloud is collected and plotted using MATLAB. In this work, only the vibration data is analyzed to diagnose the bearing faults using FFT (frequency) and CWT (time–frequency).

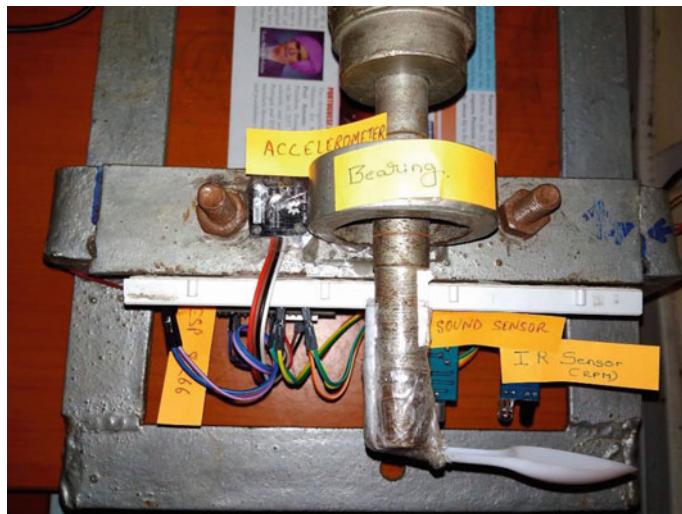


Fig. 3 Experimental set-up

Fig. 4 The defective bearing used in the experiment



4 Results and Discussion

4.1 Visualization of Condition Parameters

The measured parameters namely vibration, sound and speed has been visualized in real-time in ThingSpeak Cloud Server. The graphs are plotted and are shown in Fig. 5.

4.2 Fault Analysis

Frequency Analysis

The Fast Fourier Transform (FFT) is used in the frequency analysis technique for fault diagnosis in the bearing. The FFT of both the healthy bearing and defective bearing is plotted and the comparison of both the spectrum is done to diagnose the faults. The FFT of the healthy bearing and defective bearing is shown in Figs. 6 and 7 respectively. The comparison of FFT is made at Outer race fault frequency (ORDF) which is shown in Fig. 8. The ORDF is computed from the geometrical configuration of the bearing. The comparison at ORDF clearly indicates the faults in the bearing.

Time-Frequency Analysis

The Time-frequency analysis is the better diagnostic tool of fault analysis as it provides the information in time-frequency scale. Wavelet analysis is used for this purpose. In this work continuous wavelet transform (CWT) is used for wavelet analysis. The CWT comparison is shown in Fig. 9. From this graph, 'Red' color indicates the defective bearing vibration signal and 'Blue' color indicates the healthy bearing vibration signal where it is clearly shown in the cwt graph that defective bearing peaks are higher as compared to the healthy bearing peaks.

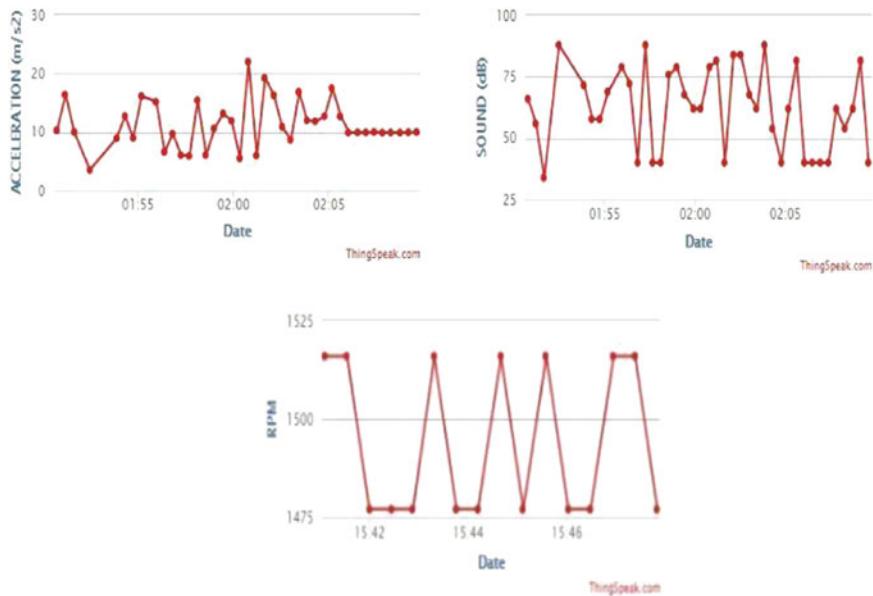
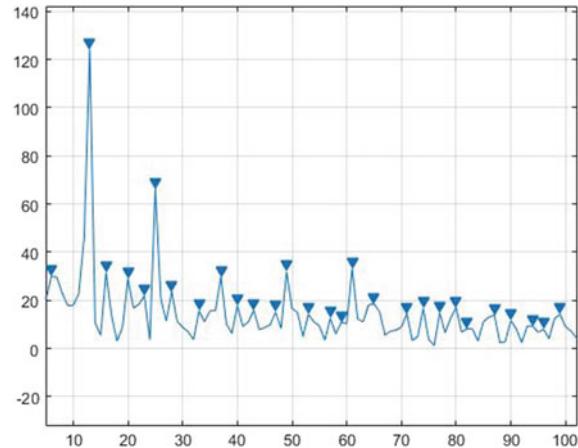


Fig. 5 Monitoring of vibration, sound and RPM in ThingSpeak

Fig. 6 FFT spectrum of vibration signal from healthy bearing



5 Conclusion

The proposed model is successfully built and tested. The proposed model has updated the conditions of bearing successfully and efficiently. It is a less expensive model. The Embedded controller sensor networks has proved that it is a reliable solution of the online remote monitoring of the condition of the rolling element bearing. The

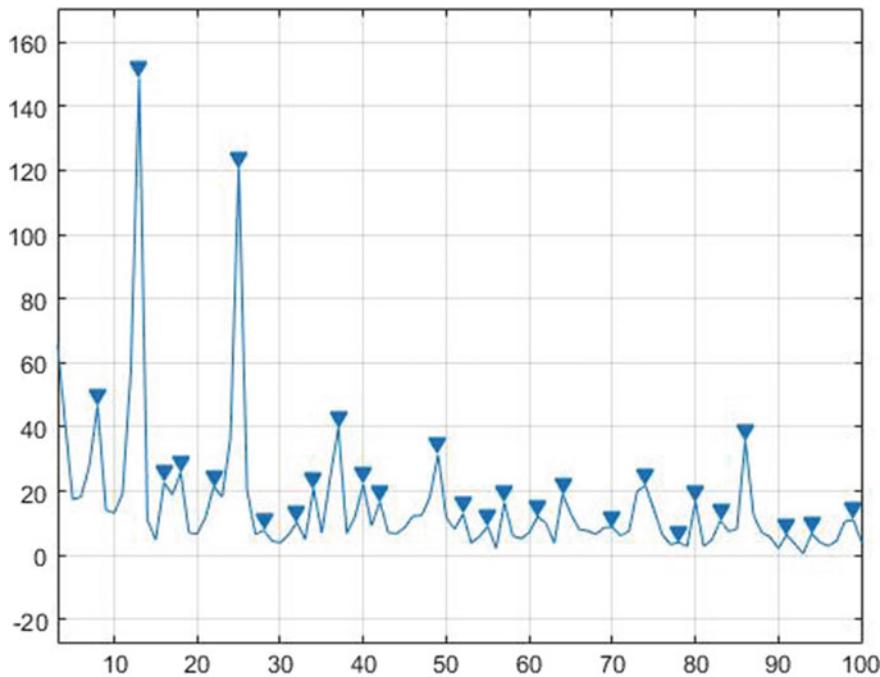


Fig. 7 FFT spectrum of vibration signal from defective bearing

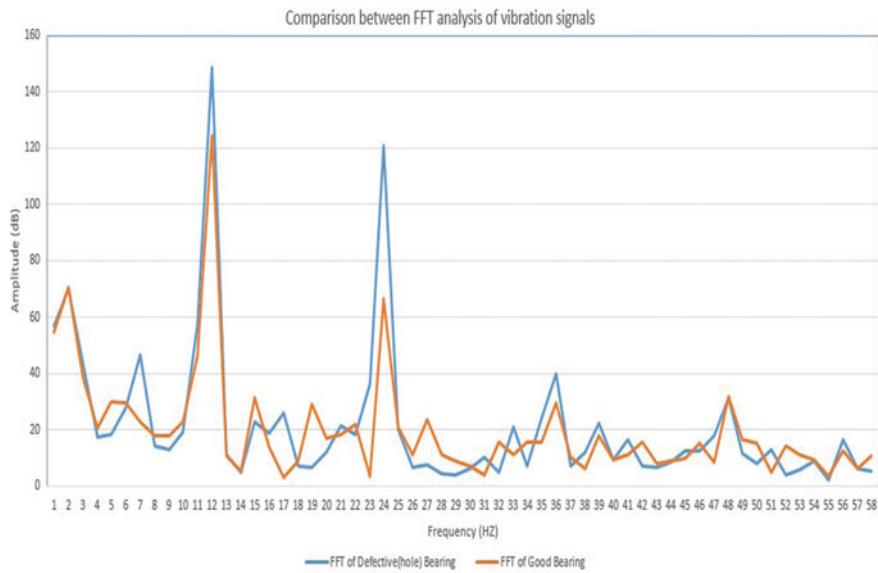


Fig. 8 Comparison of FFT of vibration signal from healthy and defective bearings

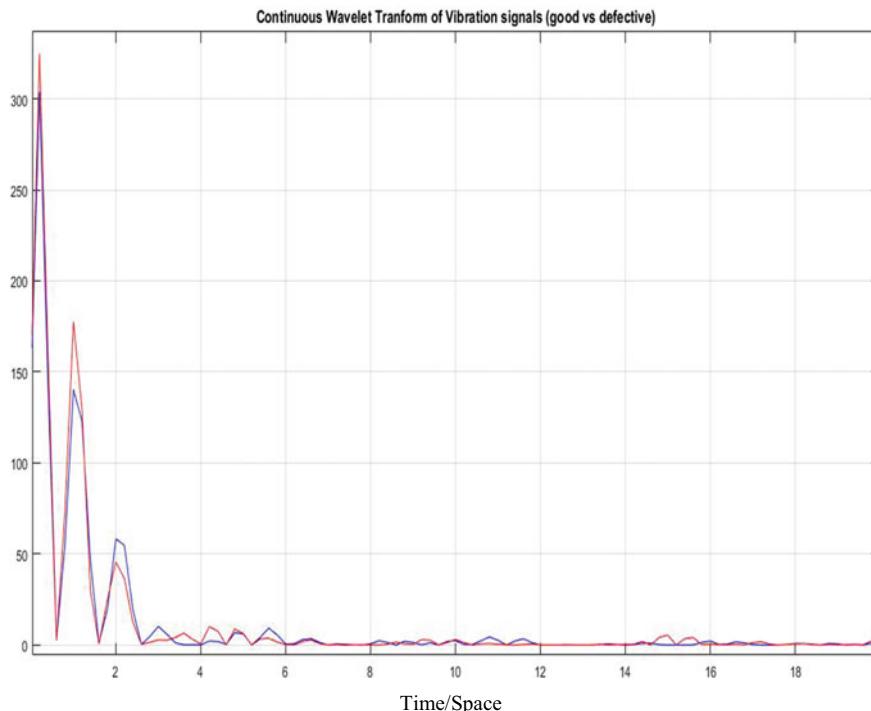


Fig. 9 Continuous wavelet transforms (CWT) of vibration signal (healthy vs. defective)

condition has been created that if vibrations are more than the maximum threshold value then ‘Tweet’ will be generated accordingly to react in Twitter social platform. Using the data acquired from the cloud server it could even predicted the condition of bearing using the site-specific ground data. Fault diagnosis has been carried out well with the help of MATLAB using FFT and CWT analysis. The proposed work can be used for the fault diagnosis of other elements of rotating machinery. In the proposed work only one defective bearing is used. The proposed work may be improved by using other defective bearings and classify the faults using the machine learning techniques. The proposed work at the current stage is having the drawback of delay in response. This limitation can be improved by using a better quality processor and Wi-Fi device.

References

1. Peng ZK, Chu FL (2004) Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography. *Mech Syst Signal Process* 18:199–221
2. El Hachemi Benbouzid M (2000) A review of induction motors signature analysis as a medium for faults detection. *IEEE Trans Ind Electron* 47(5):984–993

3. Nandi S, Toliat HA (1999) Condition monitoring and fault diagnosis of electrical machines—a review. In: Industry applications conference, 1999, thirty-fourth IAS annual meeting. conference record of the 1999, vol 1. IEEE, pp 197–204
4. Deore KS, Khandekar MA (2014) Bearing fault detection in induction motor using time domain analysis. *Int J Adv Res Electr, Electron Instrum Eng* 3(7)
5. Singh S, Kumar A, Kumar N (2014) Detection of bearing faults in mechanical systems using motor current signature and acoustic signatures. In: ICSV21, Beijing, China, 13–17 July 2014
6. Kulkarni PG, Sahasrabudhe AD (2013) Application of wavelet transform for fault diagnosis of rolling element bearings. *Int J Sci Technol Res* 2(4)
7. Kankar PK, Sharma SC, Harsha SP (2011) Fault diagnosis of ball bearings using continuous wavelet transform. *Appl Soft Comput* 11:2300-12
8. Tran M-Q, Elsisi M, Mahmoud K, Liu M-K, Lehtonen M, Darwish MMF (2021) Experimental setup for online fault diagnosis of induction machines via promising IoT and machine learning: towards Industry 4.0 empowerment. *IEEE Access* 9:115429–115441. <https://doi.org/10.1109/ACCESS.2021.3105297>
9. Pesch AH, Scavelli PN (2019) Condition monitoring of active magnetic bearings on the internet of things. *Actuators* 8(1):17
10. Sahoo S, Borthakur P, Baruah N, Chutia BP (2021) IoT and machine learning based health monitoring and heart attack prediction system. *J Phys* 1950(1):012056. <https://doi.org/10.1088/1742-6596/1950/1/012056>
11. Yaseen M, Swathi D, Kumar TA (2017) IoT based condition monitoring of generators and predictive maintenance. In: 2017 2nd International conference on communication and electronics systems (ICCES), 2017, pp 725–729. <https://doi.org/10.1109/CESYS.2017.8321176>
12. Choudhary A, Jamwal S, Goyal D, Dang RK, Sehgal S (2020) Condition monitoring of induction motor using Internet of Things (IoT). In: Kumar H, Jain P (eds) Recent advances in mechanical engineering. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-1071-7_30

Influence of Time Delay on Predator-Prey Model Having Herd Behaviour and Hunting Cooperation



Shivam, Teekam Singh, and Mukesh Kumar

Abstract This article investigates a predator-prey system in which predator species cooperate in hunting and prey species herd. We obtain a delayed predator-prey model by taking into account the fact that there is always a time delay in the conversion of biomass from prey to predator in this system. We primarily investigate the stability of positive steady state and the existence of Hopf-bifurcation in this system by using the discrete-time delay as the bifurcation parameter. The findings of this study will help us understand how realistic models of ecological systems behave in terms of dynamics.

Keywords Predator-prey system · Hunting cooperation · Herd behaviour · Time delay

1 Introduction

The relationship between different species and their living environment is one of the most important features of ecological systems, so modelling predator-prey interplay is essential in computational ecology. The Lotka–Volterra formalisation is central to the classic study of predator-prey interplay [8, 15]. Predation is a predator-prey interplay that influences the nonlinear dynamics of populations. Species are eradicated from the cluster as a result of predation. It has a significant impact on the population of prey. Consequently, all prey species modify their real environment to protect themselves from predators, moving towards low-risk from high-risk atmospheres and constructing clusters.

Shivam · M. Kumar

Department of Mathematics, Graphic Era (Deemed to be) University, Dehradun,
Uttarakhand 248002, India

T. Singh (✉)

Department of Computer Science and Engineering, Graphic Era Deemed to be University,
Dehradun, Uttarakhand 248002, India
e-mail: tsingh@ma.iitr.ac.in

Foraging, group defence, territorial defence, and cooperative hunting are all examples of cooperative behaviour in the natural world. Numerous organisms use collective defence when confronted by a predator [4]. Many animals, on the other hand, hunt in groups to increase predation efficiency. Some animals, such as wolves, African wild dogs, lions, and chimps, hunt prey in groups. Aquatic organisms, spiders, birds, and ants work together to find, attack, and transport their prey [2, 5, 6, 9, 14].

The study of time delays in mathematical models describing biological, chemical, and physical processes is gaining popularity at present. In ecological modelling, the time lag is an inescapable tool. Every biological process, such as predation, maturation, the reproductive cycle, and the conversion of prey biomass to the predator, takes time; thus, we cannot argue with facts. As a consequence, research on time-delayed systems has received more interest in recent years [1, 7, 10, 16]. It is evident by the survey that influence of time delay in predator-prey system is not studied [3, 11–13, 17, 18]. So, in this article, we investigated the effect of time delay on the conversion of prey biomass to predator biomass in the predator-prey model with both herd behaviour and cooperative hunting.

The article is presented as: In Sect. 2, the mathematical model is discussed which has discrete-time delay. In Sect. 3, the stability analysis around coexistence steady state with time delay is discussed. The Hopf-bifurcation around biological feasible steady state is discussed in Sect. 4. In Sect. 5, we validate the analytical results by numerical simulations and the article ends with conclusion in Sect. 6.

2 The Mathematical Model

The mathematical model in which both herd behaviour and hunting cooperation is defined by [12]

$$\begin{aligned} \frac{du}{dt} &= ru \left(1 - \frac{u}{c}\right) - (1 + av)\sqrt{uv}, \\ \frac{dv}{dt} &= (1 + av)\sqrt{uv} - (1 + sv)v, \end{aligned} \quad (1)$$

where population density of prey and predator are defined by u and v , respectively, at time t . The other parameters biological description: r defines the intrinsic growth of prey, c is the holding capacity, a is the hunting cooperation by predator species, m is the natural mortality of predator, and s represents the death of predator due to interspecific competition, and all the parameters in system (1) are assumed to be non-negative.

With incorporation of time delay to the system (1), then the modified form is

$$\begin{aligned}\frac{du}{dt} &= ru \left(1 - \frac{u}{c}\right) - (1 + av)\sqrt{uv}, \\ \frac{dv}{dt} &= (1 + av)\sqrt{u(t - \tau)}v - (1 + sv)v,\end{aligned}\tag{2}$$

where τ is the time delay which means the growth rate of predator population depends on the prey population τ unit of time earlier.

The initial condition for system (2) of the form

$$u_0(\theta) = \psi(\theta) \geq 0, \quad v_0(\theta) = \phi(\theta) \geq 0, \quad \theta \in [-\tau, 0], \quad u_0(0) \geq 0, \quad v_0(0) \geq 0,\tag{3}$$

where $(\psi(\theta), \phi(\theta)) \in C([-\tau, 0], R_+^2)$, $R_+^2 = \{(a_1, a_2) : a_1 \geq 0, a_2 \geq 0\}$.

It is trivial to calculate that the system (2) has following steady state:

1. $E_0 = (0, 0)$, trivial steady state.
2. $E_{\text{axial}} = (c, 0)$, axial steady state.
3. $E^* = (u^*, v^*)$, biological feasible steady state, i.e. (both species coexists).

Where the value of $u^* = ((1 + sv)/1 + av)^2$, and v^* is the solution of

$$b_1v^5 + b_2v^4 + b_3v^3 + b_4v^2 + b_5v + b_6 = 0,$$

where

$$\begin{aligned}b_1 &= a^4c, \\ b_2 &= 4a^3c, \\ b_3 &= r(s^3 - a^2sc) + 6a^2c, \\ b_4 &= r(3s^2 - 2asC - a^2c) + 4ac, \\ b_5 &= c - r(s(c - 3) + 2ac), \\ b_6 &= r(1 - c).\end{aligned}$$

For the existence of biological feasible steady state, the system must satisfy $c < 1$ and $s^2 > a^2c$.

3 Stability Analysis

Because the study must be effective in environments where both prey and predator coexist, we only use biologically feasible steady states. We now investigate the stability of the positive steady state and the effect of time delay on the dynamics of the system, using τ as the bifurcation parameter and analysing the corresponding linearized system (2).

For stability of the system with delay, we first linearizing the system (2), by substituting $u_1 = u - u^*$ and $v_1 = v - v^*$. Then, we get

$$\begin{aligned}\frac{du_1}{dt} &= b_{11}u_1 + b_{12}v_1, \\ \frac{dv_1}{dt} &= c_{21}u_1(t - \tau) + b_{22}v_1\end{aligned}\tag{4}$$

where

$$\begin{aligned}b_{11} &= -\frac{v^*(av^* + 1)}{2\sqrt{u^*}} - \frac{2ru^*}{c} + r, \quad b_{12} = -\sqrt{u^*}(2av^* + 1), \\ b_{22} &= \sqrt{u^*}(2av^* + 1) - 2sv^* - 1, \quad c_{21} = \frac{v^*(av^* + 1)}{2\sqrt{u^*}}.\end{aligned}$$

The Jacobian matrix J at $E^* = (u^*, v^*)$ is

$$J = \begin{bmatrix} b_{11} - \lambda & b_{12} \\ c_{21}e^{-\lambda\tau} & b_{22} - \lambda \end{bmatrix}\tag{5}$$

and its characteristics equation is

$$\lambda^2 - (b_{11} + b_{22})\lambda + b_{11}b_{22} - b_{12}c_{21}e^{-\lambda\tau} = 0.\tag{6}$$

Now, we will examine the system's response to delay.

Lemma 1 *If $c < 1$ and $\text{Tr}[J] < 0$, the eigenvalues of characteristics equation (6) with condition $\tau = 0$ have always (real part is $-ve$) its means the model is asymptotically stable.*

Proof For $\tau = 0$, Eq. (6) reduces to

$$\lambda^2 - (b_{11} + b_{22})\lambda + b_{11}b_{22} - b_{12}c_{21} = 0.\tag{7}$$

From Routh–Hurwitz criteria, the system has no delay is stable if the real part of eigenvalues must have negative real parts. The necessary and sufficient conditions which ensure the stability of system are $\text{Tr}[J] < 0$ and $\text{Det}[J] > 0$, i.e.

$$\begin{aligned}\text{Tr}[J] &= b_{11} + b_{22} < 0, \\ \text{Det}[J] &= b_{11}b_{22} - b_{12}c_{21} > 0,\end{aligned}\tag{8}$$

and the corresponding values of λ are $\lambda_i = \text{Tr}[J] \pm \sqrt{\text{Tr}[J]^2 - 4\text{Det}[J]}/2$, ($i = 1, 2$). Hence proved. \square

Now, for $\tau > 0$, i.e. case of positive delay.

Lemma 2 *The delayed system (2), the transcendental equation (6) has one pure imaginary root.*

Proof As Eq. (6) is the transcendental equation implies it has infinitely many solutions at $E^* = (u^*, v^*)$. Let $\lambda = a + i\omega$ be the general form of eigenvalue of Eq. (6). But, we are intrusted to find the periodic solutions of the system, the eigenvalue of Eq. (6) must be purely imaginary.

Assuming $a = 0$, then $\lambda = i\omega$, where $\omega > 0$. On substituting $\lambda = i\omega$ into the characteristic equation (6), yields

$$-\omega^2 - (b_{11} + b_{22})\omega i + b_{11}b_{22} - b_{12}c_{21}(\cos \omega\tau - i \sin \omega\tau) = 0. \quad (9)$$

On separating the real and imaginary parts, we have

$$\begin{aligned} b_{12}c_{21} \cos \omega\tau &= -\omega^2 + b_{11}b_{22}, \\ b_{12}c_{21} \sin \omega\tau &= (b_{11} + b_{22})\omega, \end{aligned} \quad (10)$$

which gives

$$\omega^4 + (b_{11}^2 + b_{22}^2)\omega^2 + (b_{11}^2b_{22}^2 - b_{12}^2c_{21}^2) = 0. \quad (11)$$

From above Eq. (11), it is clear that $b_{11}b_{22} - b_{12}c_{21} > 0$, we has no positive roots if $b_{11}b_{22} - b_{12}c_{21} \geq 0$, and has positive root ω^+ if $b_{11}b_{22} - b_{12}c_{21} < 0$, where

$$\omega^+ = \sqrt{\frac{-(b_{11}^2 + b_{22}^2) + \sqrt{(b_{11}^2 + b_{22}^2)^2 + 4(b_{11}^2b_{22}^2 - b_{12}^2c_{21}^2)}}{2}} \quad (12)$$

Substituting the value of ω^+ into Eq. (10), we have

$$\tan(\omega^+\tau) = \frac{(b_{11} + b_{22})\omega^+}{b_{11}b_{22} - (\omega^+)^2}. \quad (13)$$

Then,

$$\tau_j = \frac{1}{\omega^+} \arctan \left[\frac{(b_{11} + b_{22})\omega^+}{b_{11}b_{22} - (\omega^+)^2} \right] + \frac{2j\pi}{\omega^+}, \quad (j = 0, 1, 2 \dots) \quad (14)$$

As a result of Butler's Lemma, at $E^* = (u^*, v^*)$, the delayed system (2) stays stable for $\tau < \tau_j$ at $k = 0$. \square

4 Hopf-Bifurcation

In this section, we show the transversality condition hold and system (2) will experience Hopf-bifurcation at $\tau = \tau_j$.

Lemma 3 Let $c < 1$, and $b_{11} + b_{22} < 0$, then the transversality condition must satisfied, i.e.

$$\left[\frac{d\operatorname{Re}(\lambda)}{d\tau} \right]_{\tau=\tau_j} > 0, \quad j = 1, 2, \dots \quad (15)$$

Proof As $\lambda = a(\tau) + i\omega(\tau)$ be the root of characteristic equation (6). Then, at $\tau = \tau_j$, $a(\tau_j) = 0$, $\omega(\tau_j) = \omega^+$, ($j = 1, 2, \dots$). On differentiating Eq. (6) with respect to τ , we have

$$2\lambda \frac{d\lambda}{d\tau} - (b_{11} + b_{22}) \frac{d\lambda}{d\tau} + b_{12}c_{21}e^{-\lambda\tau} \left(\lambda + \tau \frac{d\lambda}{d\tau} \right) = 0, \quad (16)$$

which gives

$$\begin{aligned} \left(\frac{d\lambda}{d\tau} \right)^{-1} &= \frac{2\lambda - b_{11} - b_{22}}{-b_{12}c_{21}\lambda} e^{\lambda\tau} - \frac{\tau}{\lambda} \\ \operatorname{sign} \left[\operatorname{Re} \left(\frac{d\lambda}{d\tau} \right)^{-1} \right]_{\tau=\tau_j} &= \operatorname{sign} \left\{ \operatorname{Re} \left[\frac{2\lambda - b_{11} - b_{22}}{-b_{12}c_{21}\lambda} e^{\lambda\tau} - \frac{\tau}{\lambda} \right] \right\}_{\tau=\tau_j} \\ \operatorname{sign} \left[\operatorname{Re} \left(\frac{d\lambda}{d\tau} \right)^{-1} \right]_{\tau=\tau_j} &= \operatorname{sign} \left[\operatorname{Re} \left[\frac{b_{11} + b_{22} - 2\lambda}{b_{12}c_{21}\lambda} e^{\lambda\tau} - \frac{\tau}{\lambda} \right] \right\}_{\tau=\tau_j} \\ \operatorname{sign} \left[\operatorname{Re} \left(\frac{d\lambda}{d\tau} \right)^{-1} \right]_{\tau=\tau_j} &= \operatorname{sign} \left\{ \operatorname{Re} \left[\frac{b_{11} + b_{22} - 2i\omega^+}{b_{12}c_{21}i\omega^+} (\cos \omega^+ \tau_j + i \sin \omega^+ \tau_j) \right] \right\} \\ \operatorname{sign} \left[\operatorname{Re} \left(\frac{d\lambda}{d\tau} \right)^{-1} \right]_{\tau=\tau_j} &= \operatorname{sign} \left\{ \operatorname{Re} \left[\frac{2((\omega^+)^2 - b_{11}b_{22})}{b_{12}^2c_{21}^2} + \frac{(b_{11} + b_{22})^2}{b_{12}^2c_{21}^2} \right] \right\} \\ \operatorname{sign} \left[\operatorname{Re} \left(\frac{d\lambda}{d\tau} \right)^{-1} \right]_{\tau=\tau_j} &= \operatorname{sign} \left\{ \operatorname{Re} \left[\frac{2(\omega^+)^2 + b_{11}^2 + b_{22}^2}{b_{12}^2c_{21}^2} \right] \right\} > 0. \end{aligned}$$

□

From all above three lemmas, we have a theorem.

Theorem 1 If $c < 1$, and $b_{11} + b_{22} < 0$, the system (2) the following conditions hold

1. At $E^* = (u^*, v^*)$, the system is locally asymptotically stable when $0 \leq \tau < \tau_0$ and unstable when $\tau > \tau_0$.
2. The system will undergoes Hopf-bifurcation at $E^* = (u^*, v^*)$ when $\tau = \tau_0$.

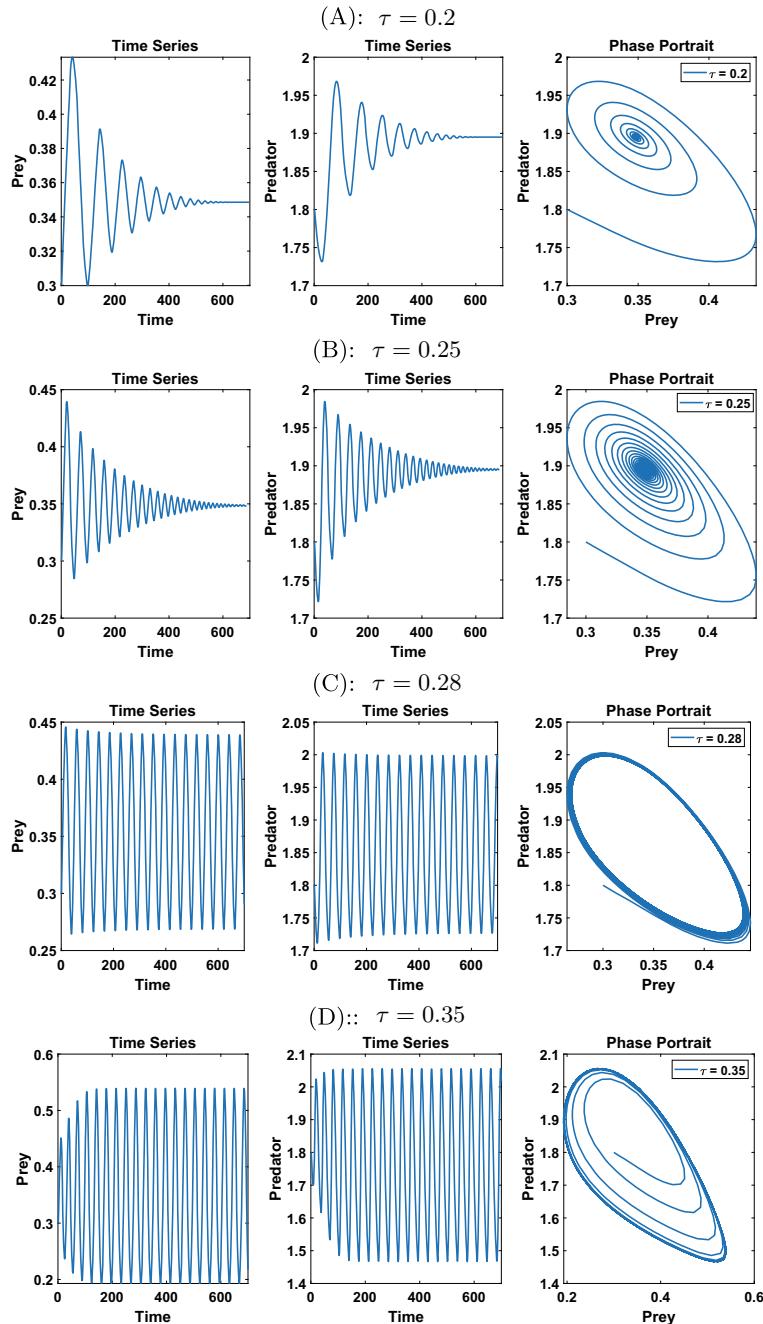


Fig. 1 Time evolution and phase portrait of system (2)

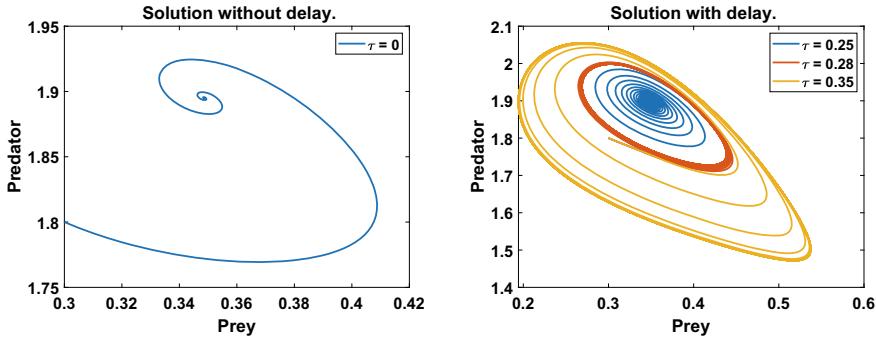


Fig. 2 Influence of time delay on system (2) with different τ

5 Numerical Simulations

For fixed parameter values, we present numerical simulations in order to validate our analytical result. The arbitrary values of parameters are:

$$a = 0.4, \quad s = 0.02, \quad c = 0.8, \quad \text{and} \quad r = 10.$$

For the fixed value of parameters the system has following steady state: $E_0 = (0, 0)$, $E_a = (0.8, 0)$, and two coexistence steady state E^* , i.e. $(0.3485, 1.8951)$ and $(0.7581, 0.3938)$.

At $(0.7581, 0.3938)$, the system has eigenvalues -9.1823 and 0.0971 . Hence, the system is saddle at $(0.7581, 0.3938)$. Whereas, at $(0.3485, 1.8951)$, the calculated eigenvalues are $-0.5628 + 1.80157i$, and $-0.5628 - 1.80157i$. Here the real part of the characteristics-values are negative, therefore the model is stable at this coexistence point and used in further analysis.

From the analytical results, we compute the value of $\tau_0 = 0.28$. By Theorem 1, we choose four values of τ , i.e. $\tau = 0.2, 0.25, 0.28, 0.35$. The computational simulations are depicted in Fig. 1 and the outcomes support the results of Theorem 1. When the value of $\tau < \tau_0$ the system shows spiral which means the system is stable (cf. Fig. 1a, b), when $\tau > \tau_0$ the system is unstable (cf. Fig. 1d), and when $\tau = \tau_0$ the system shows Hopf-bifurcation (cf. Fig. 1c). Figure 2 shows the influence of time delay on predator-prey system with comparison to system having no delay.

6 Conclusion

In this paper, we look at a delayed predator-prey system in which predators cooperate in hunting and prey species exhibit herd behaviour. We determined the system stability around the biologically feasible steady state with the addition of τ to investi-

gate the effect of a discrete-time delay on the predator-prey system. We calculate the discrete-time delay ($\tau = \tau_0 = 0.28$). From the summarised Theorem 1, it is clear that the predator-prey system is stable if $\tau < \tau_0$, unstable $\tau > \tau_0$ and undergoes Hopf-bifurcation if $\tau = \tau_0$.

From the finding, it is clear that the dynamical behaviour of the system (2) corresponding to a non-delayed system is robust with respect to the time delay (cf. Fig. 2). So, the biological factor, i.e. time delay, can influence the system's stability. The finding may enhance the dynamics of predator-prey systems to improve our understanding of the interplay between predators and prey in natural ecosystems.

References

1. Arditi R, Abillon JM, da Silva JV (1977) The effect of a time-delay in a predator-prey model. *Math Biosci* 33(1–2):107–120
2. Bshary R, Hohner A, Ait-el Djoudi K, Fricke H (2006) Interspecific communicative and coordinated hunting between groupers and giant moray eels in the Red Sea. *PLoS Biol* 4(12):e431
3. Du Y, Niu B, Wei J (2022) A predator-prey model with cooperative hunting in the predator and group defense in the prey. *Discrete Contin Dyn Syst B* 27(10):5845
4. Dugatkin LA (1997) Cooperation among animals: an evolutionary perspective. Oxford University Press on Demand
5. Hector DP (1986) Cooperative hunting and its relationship to foraging success and prey size in an avian predator. *Ethology* 73(3):247–257
6. Holmes J (1972) Modification of intermediate host behaviour by parasites. Behavioural aspects of parasite transmission
7. Kar TK (2004) Stability analysis of a prey-predator model with delay and harvesting. *J Biol Syst* 12(01):61–71
8. Lotka AJ (1956) Elements of mathematical biology. Dover Publications
9. Moffett MW (1988) Foraging dynamics in the group-hunting myrmicine ant, *Pheidologeton diversus*. *J Insect Behav* 1(3):309–331
10. Mukhopadhyay B, Bhattacharyya R (2005) Dynamics of a delay-diffusion prey-predator model with disease in the prey. *J Appl Math Comput* 17(1):361–377
11. Shivam, Kumar M, Singh T, Chauhan S (2022) Positive effect of predator's mortality in predator-prey system via turing patterns. *Braz J Phys* 52(5):1–11
12. Shivam, Singh K, Kumar M, Dubey R, Singh T (2022) Untangling role of cooperative hunting among predators and herd behavior in prey with a dynamical systems approach. *Chaos Solitons Fract* 162:112420
13. Singh T, Shivam, Kumar M, Vimal V (2021) Pattern dynamics of prey–predator model with swarm behavior via turing instability and amplitude equation. In: Mathematical modeling, computational intelligence techniques and renewable energy. Springer, pp 275–285
14. Uetz GW (1992) Foraging strategies of spiders. *Trends Ecol Evol* 7(5):155–159
15. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Società anonima tipografica “Leonardo da Vinci”
16. Wangersky PJ, Cunningham W (1957) Time lag in prey-predator population models. *Ecology* 38(1):136–139
17. Yuan S, Xu C, Zhang T (2013) Spatial dynamics in a predator-prey model with herd behavior. *Chaos Interdiscipl J Nonlinear Sci* 23(3):033102
18. Zhu Z, Wu R, Lai L, Yu X (2020) The influence of fear effect to the Lotka–Volterra predator-prey system with predator has other food resource. *Adv Differ Equ* 2020(1):1–13

An Improved Jaya Algorithm (IJAYA) for Optimization



Sonal Deshwal, Pravesh Kumar, and Sandeep Mogha

Abstract Jaya algorithm is a newly developed algorithm with a very straightforward structure and requires only a few number of control parameters for optimization. However, it needs to be improved in terms of exploration and accuracy of outcomes with a fast convergence speed. This work proposed an improved version “IJAYA”, of the Jaya algorithm by introducing an average number-based equation of best and worst position of vectors in search space. The proposed method approaches worst and best positions to each other and resulting in increase in the search power and exploration index. IJAYA tested on six benchmark problems picked from the literature. Finally, the comparison and result demonstrate the effectiveness of our suggested approach.

Keywords Optimization · Population-based algorithms · Jaya algorithm

1 Introduction

The process of determining the optimum solutions for a given system from all available values in order to maximize or minimize the output is referred to as optimization. Many real-world issues have a significant number of nonlinear constraint-based solution spaces. In addition to being non-convex and difficult, these issues have a significant computational cost. As a result, addressing such issues with a high number of variables and constraints is a difficult process. Furthermore, local optimal solutions derived from a variety of traditional methodologies may not always imply the best option. The researchers suggest a number of metaheuristic optimization

S. Deshwal (✉) · S. Mogha

Department of Mathematics, Chandigarh University, Mohali, Punjab, India

e-mail: sonaldeswal1995@gmail.com

S. Mogha

e-mail: moghadma@gmail.com

P. Kumar

Department of Mathematics, Rajkiya Engineering College Bijnor, Bijnor, India

e-mail: pkumarrecb@gmail.com

techniques [1] to address these challenges, which have been proven to be highly efficient in handling exceedingly complicated problems. Researchers have placed a greater emphasis on the development of metaheuristic algorithms that are computationally cheap, versatile, and simple [2]. Evolutionary algorithms (EA) and swarm intelligence (SI)-based algorithms are the two major types of population-based algorithms. Distinct EA are inspired by natural processes, which encompass reproduction, mutation, recombination, and selection. All of these EA are based on the candidate's survival fitness in a population (i.e., a set of solutions). Genetic algorithm (GA) [3], differential evolution (DE) [4], bacterial foraging optimization (BFO) [5], and others are some of well-known evolutionary algorithms. Swarm intelligence is typically mimicked by biologically inspired algorithms. Such intelligence can be adopted by flocks of birds, ants, and other animals. Particle swarm optimization (PSO) [6], shuffled frog leaping (SFL) [7], ant colony optimization (ACO) [8], artificial bee colony (ABC) [9], fire fly (FF) method [10], gray wolf optimizer (GWO) [11], and others are examples of well-known swarm intelligence-based algorithms. Aside from evolutionary and swarm intelligence-based algorithms, some additional algorithms are based on the fundamentals of various natural occurrences. Harmony search (HS) method [12], biogeography-based optimization (BBO) [13], integrated radiation algorithm (IRA) [14], charged system search algorithm (CSSA) [15], artificial physics algorithm (APA), gravitational field algorithm (GFA) [16], multi-area economic dispatch using evolutionary algorithms [17], GAMS environment-based solution methodologies for ramp rate constrained profit-based unit commitment problem [18], and others are among them.

It is worth noting that, depending on the particular optimisation problem, the majority of metaheuristic algorithms depend on a few (usually fine-tuned) factors whose impact might significantly alter the algorithmic performance. On the other hand, it would be preferable from the perspective of a user to have an algorithm that needs the fewest number of parameters in order to generalize its use without requiring additional work to alter the parameter value for each unique situation. Because of this, a current direction in metaheuristic optimization study is to develop self-adaptive methods that can be used to solve any optimization issue with no adjustment. The teaching–learning-based optimization (TLBO) [19] method is one example that demonstrates this trend. It just requires two parameters, the population size and the number of generations, both of which are apparently necessary for all swarm intelligence algorithms. Many algorithms built on similar concepts have been inspired by TLBO in recent years, including the Jaya algorithm [20].

Rao [20] introduced the Jaya algorithm, a contemporary metaheuristic-based algorithm that amalgamate the qualities of EA with regard to survivability for the fittest rule, along with SI, wherein the swarm follows the captain throughout the look for the best solution. Use of it is easy and does not require any parameters. It is adjustable, versatile, and robust. This is why it has quickly gained popularity in tackling many real-world optimization issues. As a result, the Jaya algorithm has been extensively used for a variety of optimization problems in various domains, including optimal power flow, feature selection, parameter extraction of solar cells,

etc. A number of researchers have updated the Jaya algorithm to improve its convergence behavior in response to the complexity of specific optimization problems and their harsh aspects in their search space. Some researchers combine the Jaya algorithm with other optimization techniques. Consequently, several Jaya algorithm versions, including binary Jaya, chaotic Jaya, hybridization with swarm intelligence algorithms, hybridization with evolutionary algorithms and hybridization with other components. Several newly created Jaya algorithm versions and their uses are provided in [21–31]. However, it can be improved in terms of exploration and accuracy of outcomes as well as reaching a better solution more quickly. This work proposes IJAYA, which is an improved version of the Jaya algorithm.

The remaining sections are structured as follows: Sect. 2 gives a brief description of Jaya. Section 3 presents a brief detail of proposed IJAYA algorithm. Results and comparison are reported in Sect. 4, and finally, the conclusions derived from the present study are drawn in Sect. 5.

2 Jaya Algorithm

Rao introduced the Jaya algorithm, a population-based metaheuristic method for tackling constrained and unconstrained optimization problems [20]. The title Jaya comes from Sanskrit and signifies “victory”. Its foundation is the natural behavior of “survival of the fittest” idea. This indicates that among the Jaya population, the finest global solutions are sought for while the poorest concepts are disregarded. To put it differently, the search method used by the Jaya algorithm aims to prevent failure by eliminating the worst solutions as well as works to find the global best answers. Therefore, each iteration updates all Jaya candidates, all iteration solutions are better than the prior worst solution. Every solution in Jaya is referred to as a particle. In the search region, each particle seeks the best solution and avoids the worst solution of cost function or objective.

Let $f(x)$ be the minimized (or maximized) objective function. Assume that population of size N (i.e., $i = 1, 2, 3, \dots, N$). If $X_{\text{best}}(k)$ and $X_{\text{worst}}(k)$ are the global optimal and global worst vectors in the search space for any generation k , then Eq. 1 generates the new location for any vector $X_i(k)$ for the following generation.

$$X_i(k+1) = X_i(k) + \text{rand}_1 * (X_{\text{best}}(k) - |X_i(k)|) - \text{rand}_2 * (X_{\text{worst}}(k) - |X_i|) \quad (1)$$

where rand_1 and rand_2 are uniform random values ranging from 0 to 1. $X_i(k+1)$ represents $X_i(k)$'s updated value. If $X_i(k+1)$ delivers the better fitness value when compared to $X_i(k)$, it is preserved. If not, $X_i(k)$ will be kept for the next generation. When an iteration is complete, the values of the objective function are stored, and the following iteration uses these values as input for calculation until the optimal solution is found.

3 IJAYA Algorithm

This section introduces the proposed algorithm before moving on to mathematical modeling. The new improved IJAYA optimization technique is more versatile and achieves a fair mix of exploration and exploitation. It is worth noting that this approach has no effect on the original algorithm's computational complexity. The improved Jaya algorithm is conditioned by $\text{rand} < P$, where P is a fixed parameter. Before performing the algorithm, a value between 0 and 1 is supplied to P , and the algorithm is then executed using this fixed value.

If $\text{rand} < P$ Eq. 1 will execute the objective; otherwise, an average number-based equation will be performed. Here, the average value is used to represent the average of the best and worst positions. With this approach, the worst and best population members approaching one another and hence, this method helps to increases the search power and exploration index. The IJAYA flowchart is shown in Fig. 1.

The planned IJAYA is described in detail below.

Let $f(x)$ be the minimized (or maximized) objective function.

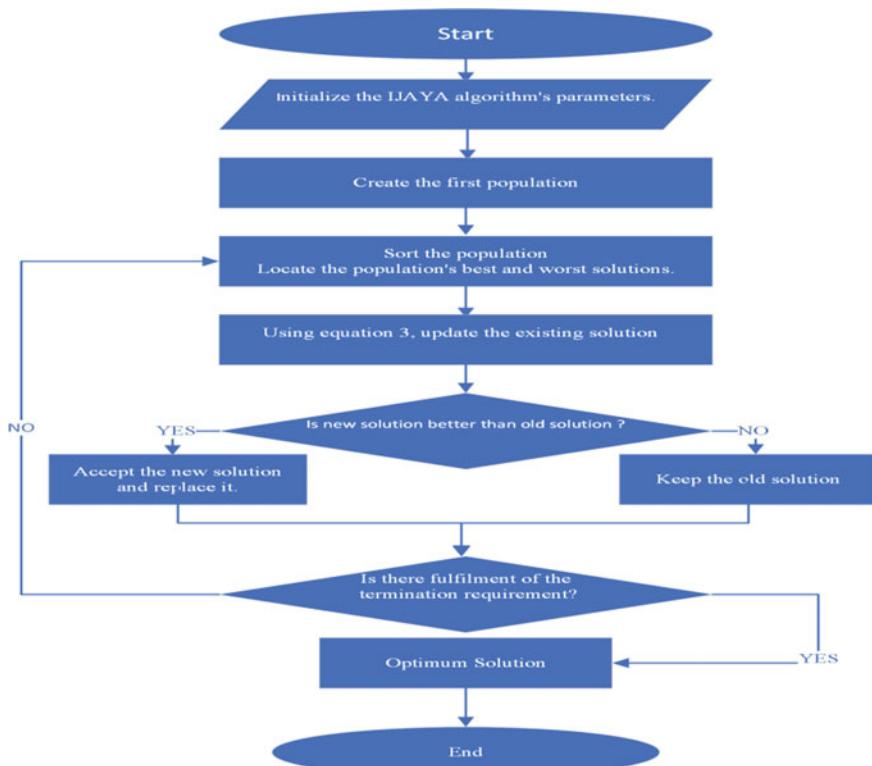


Fig. 1 The IJAYA flowchart

- Step 1: In the initial step of the run, the parameters of the IJAYA algorithm are established. The population size N , iteration numbers k , and P are the sole algorithmic parameters.
- Step 2: IJAYA algorithm's initial population are built and stored in an augmented matrix of size $N \times D$, where N represents the number of candidate solutions (i.e., population size) and D is the solution dimension, as shown in Eq. 2. Traditionally, solutions are built at random

$$x_{ij} = l_j + (u_j - l_j) * \text{rand}, \text{ where } i = 1, 2, \dots, N \& j = 1, 2, \dots, D$$

Random numbers between 0 and 1 are generated using the uniform function rand . l_j and u_j are, respectively, lower bounds and upper bounds of search domain.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \quad (2)$$

For each solution, evaluates the objective function $f(x)$, and also the X solutions are sorted in ascending order by their objective function values. As a result, for each generation k , $X_{\text{best},j}^k$ is the best value in the search space, whereas $X_{\text{worst},j}^k$ is the worst value.

- Step 3: Process of IJAYA evolution. Iteratively, using the IJAYA operator described in Eq. 3, all decision variables of all solutions in the X are changed.

$$X_{i,j}^{k+1} = \begin{cases} X_{i,j}^k + \text{rand}_1 * (X_{\text{best},j}^k - |X_{i,j}^k|) - \text{rand}_2 * (X_{\text{worst},j}^k - |X_{i,j}^k|) & \text{if } (\text{rand}_1 \text{rand}_2 < \text{Pr}) \\ \text{rand}_3 * ((X_{\text{worst},j}^k + X_{\text{best},j}^k)/2) - X_{i,j}^k & \text{else} \end{cases} \quad (3)$$

where Pr is a parameter that has a value between 0 & 1, $X_{i,j}^{k+1}$ represents the modified solution, and $X_{i,j}^k$ represents the current solution. $X_{i,j}^{k+1}$ represents the decision value's $(X_{i,j}^k)$ modified value. The uniform functions rand_1 , rand_2 and rand_3 generate random numbers between 0 and 1. The decision variable j 's value for the best candidate, $X_{\text{best},j}^k$ and the worst candidate, $X_{\text{worst},j}^k$ are represented by the decision variable j 's values.

- Step 4: Updated M . At each iteration k , the X solutions will be updated. If the new created vector $X_{i,j}^{k+1}$ has a better fitness value than $X_{i,j}^k$, it will be replaced; otherwise, $X_{i,j}^k$ will be preserved for the next generation, as shown in Eq. 4.

$$X_{i,j}^{k+1} = \begin{cases} X_{i,j}^{k+1}, \text{ if } f(X_{i,j}^{k+1}) < f(X_{i,j}^k) \\ X_{i,j}^k & \text{else} \end{cases} \quad (4)$$

Table 1 Benchmark functions

Function	Name	Boundary	Dimension	Global value	Pr
F1	Ackley	[-32, 32]	30	0	0.1
F2	Bohachevsky 1	[-100, 100]	2	0	0.1
F3	GoldStein-Price	[-2, 2]	2	0	0.6
F4	Michalewicz 2	[0, Π]	2	0	0.9
F5	Perm	[-2, 2]	2	0	0.9
F6	Himmelblau	[-6, 6]	2	0	0.1

This procedure will be repeated N times.

Step 5: Until the IJAYA algorithm hits the terminating rule, which is frequently the maximum number of iterations k , steps 3 and 4 are repeated.

4 Conclusions and Discussion

The importance of the suggested IJAYA algorithm has been tested through a number of tests. In this part, experimental results analysis and a detailed description of test functions, parameter settings, and comparison criteria are covered.

4.1 Benchmark Functions

Six benchmark functions have been used in the research. All benchmark functions are listed in Table 1 and have been obtained from the literature [32]. By solving six standard benchmark functions, the numerical efficiency of algorithm was assessed.

4.2 Parameter Setting

Table 2 contains the parameter settings for the test. For every function in Table 2's optimization problems to be fairly compared, the population size was set to 50 and the number of iterations to 500. MATLAB version R2016a was used to implement the proposed technique on a laptop with these specs: 4 GB RAM, Intel(R) Core(TM) i3-7100U CPU @ 2.40 GHz, 64-bit, and Windows 10.

Table 2 Parameter settings

S. No.	Parameter name	Parameter setting
1	Population size (N)	50
2	Maximum iteration	500
3	Total run	30
4	Rand_1 , Rand_2 , & Rand_3	[0, 1]
5	Software used	MATLAB version R2016a

4.3 Results and Comparison

This section evaluates the performance of the proposed IJAYA by comparing it with existing optimization methods as BA, DE, PSO, GWO, GA, and CS. For each benchmark function, appropriate population sizes were selected to run the suggested IJAYA 30 times. The average outcomes are compared with those of other algorithms after the same number of runs. Here, it is clear that the suggested IJAYA achieved the desired outcomes for all functions when compared to alternatives. According to Table 3's outcomes, IJAYA performs significantly better than BA, DE, PSO, GWO, GA, CS, and Jaya in all statistical values except for the best value for F3. All algorithms except GA and GWO work similarly for function F1. Except for GA, GWO, and CS, all algorithms for function F4 perform similarly. For the functions F2 and F5, IJAYA outperforms all other methods. Jaya and IJAYA perform better than any other algorithms for function F6 (Table 3).

The Wilcoxon test indicates that, for function F1, IJAYA outperforms GWO and GA while performing similarly to DE, PSO, CS, BE, and Jaya. Its performance for function F2 is comparable to GWO and superior to GA, DE, PSO, CS, BE, and Jaya, similar to DE, PSO, CS, BE, Jaya, GWO, and GA in terms of performance for function F3. For function F4, IJAYA beats GWO, GA, and CS while performing similarly to DE, PSO, BE, and Jaya. It demonstrates that for function F5, IJAYA outperforms all alternative techniques. And for function F6, IJAYA surpasses all other algorithms and performs comparable to Jaya.

5 Conclusion

In this paper, we have suggested an improved Jaya algorithm (IJAYA) by introducing an average equation of best and worst values and implemented it with Jaya algorithm by a probability Pr . A set of six standard benchmark problems were used to test the proposed algorithm. The numerical results proved that the proposed adjustments assist improve Jaya's performance in terms of convergence rate without sacrificing solution quality. The proposed method is simple and enhances the working of fundamental Jaya, performing better or equal with some other population based search algorithms.

Table 3 Performance comparisons between several benchmark functions

Functions	Statistical values	DE	BA	PSO	GWO	CS	GA	Jaya	IIAYA
F1	Best	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$1.003e^{-13}$	$0.000e^{+00}$	$1.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$
	Mean	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$1.061e^{-13}$	$0.000e^{+00}$	$1.456e^{+01}$	$0.000e^{+00}$	$0.000e^{+00}$
	SD	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$1.802e^{-14}$	$0.000e^{+00}$	$1.416e^{-06}$	$0.000e^{+00}$	$0.000e^{+00}$
F2	Best	$9.724e^{-06}$	$8.584e^{-06}$	$8.752e^{-06}$	$0.000e^{+00}$	$9.518e^{-06}$	$8.639e^{-06}$	$8.447e^{-06}$	$0.000e^{+00}$
	Mean	$9.724e^{-06}$	$8.584e^{-06}$	$8.775e^{-06}$	$0.000e^{+00}$	$9.114e^{-06}$	$8.607e^{-06}$	$8.447e^{-06}$	$0.000e^{+00}$
	SD	$3.950e^{-09}$	$3.807e^{-08}$	$3.828e^{-09}$	$0.000e^{+00}$	$3.978e^{-08}$	$3.836e^{-08}$	$3.798e^{-08}$	$0.000e^{+00}$
F3	Best	$8.384e^{-05}$	$7.660e^{-05}$	$8.396e^{-05}$	$1.720e^{-08}$	$9.162e^{-05}$	$7.392e^{-05}$	$6.549e^{-05}$	$4.236e^{-05}$
	Mean	$8.384e^{-05}$	$7.660e^{-05}$	$8.400e^{-05}$	$1.240e^{-05}$	$9.162e^{-05}$	$5.870e^{-00}$	$6.549e^{-05}$	$4.236e^{-05}$
	SD	$0.000e^{+00}$	$8.482e^{-07}$	$9.200e^{-07}$	$1.304e^{-05}$	$9.420e^{-07}$	$1.000e^{+00}$	$8.750e^{-07}$	$3.052e^{-05}$
F4	Best	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$-1.801e^{+00}$	$7.220e^{-07}$	$-1.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$
	Mean	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$-1.801e^{+00}$	$7.220e^{-07}$	$-1.800e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$
	SD	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$	$-4.562e^{-16}$	$4.590e^{-08}$	$0.000e^{+00}$	$0.000e^{+00}$	$0.000e^{+00}$
F5	Best	$6.620e^{-06}$	$6.486e^{-06}$	$6.630e^{-06}$	$2.768e^{-13}$	$7.299e^{-06}$	$2.768e^{-13}$	$6.419e^{-06}$	$0.000e^{+00}$
	Mean	$6.620e^{-06}$	$6.486e^{-06}$	$6.620e^{-06}$	$1.126e^{-09}$	$7.712e^{-06}$	$1.126e^{-09}$	$6.419e^{-06}$	$0.000e^{+00}$
	SD	$0.790e^{-07}$	$0.389e^{-07}$	$0.399e^{-07}$	$2.665e^{-09}$	$0.492e^{-07}$	$1.932e^{-01}$	$0.349e^{-07}$	$0.000e^{+00}$
F6	Best	$6.289e^{-05}$	$1.059e^{-05}$	$7.439e^{-06}$	$4.143e^{-07}$	$0.000e^{+00}$	$1.260e^{-14}$	$0.000e^{+00}$	$0.000e^{+00}$
	Mean	$2.390e^{-05}$	$7.452e^{-02}$	$4.650e^{-07}$	$6.492e^{-06}$	$2.602e^{-19}$	$1.170e^{-13}$	$0.000e^{+00}$	$0.000e^{+00}$
	SD	$1.209e^{-05}$	$7.452e^{-02}$	$1.630e^{-08}$	$8.519e^{-06}$	$5.350e^{-19}$	$1.930e^{-01}$	$0.000e^{+00}$	$0.000e^{+00}$

Table 4 Results on Wilcoxon test

Functions	Wilcoxon test						
	8/1	8/2	8/3	8/4	8/5	8/6	8/7
F1	=	=	=	+	=	+	=
F2	+	+	+	=	+	+	+
F3	=	=	=	=	=	=	=
F4	=	=	=	+	+	+	=
F5	+	+	+	+	+	+	+
F6	+	+	+	+	+	+	=

References

1. Talbi E-G (2009) Metaheuristics: from design to implementation. John Wiley & Sons, Hoboken, New Jersey
2. Kaveh A (2016) Advances in metaheuristic algorithms for optimal design of structures, 2nd edn. Springer Nature, Cham, Switzerland
3. Goldberg D (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, New York
4. Basturk B, Karaboga D (2006) An artificial bee colony (ABC) algorithm for numeric function optimization. In: IEEE swarm intelligence symposium, 2006
5. Passino KM (2010) Bacterial foraging optimization. *Int J Swarm Intell Res* 1(1):1–16
6. Okwu MO, Tartibu LK (2021) Particle swarm optimisation. *Stud Comput Intell* 927:5–13
7. Eusuff M, Lansey K, Pasha F (2006) Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. *Eng Optim* 38(2):129–154
8. Blum C (2005) Ant colony optimization: introduction and recent trends. *Phys Life Rev* 2(4):353–373
9. Karaboga D, Akay B (2009) A comparative study of Artificial Bee Colony algorithm. *Appl Math Comput* 214(1):108–132
10. Yang XS, He X (2013) Firefly algorithm: recent advances and applications. *Int J Swarm Intell* 1(1):36
11. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. *Adv Eng Softw* 69:46–61
12. Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. *SIMULATION* 76(2):60–68
13. Simon D, Member S (2008) Biogeography-based optimization 12(6):702–713
14. Chuang CL, Jiang JA (2007) Integrated radiation optimization: inspired by the gravitational radiation in the curvature of space-time. 2007 IEEE Congr Evol Comput CEC 2007 3157:3157–3164
15. Kaveh A, Talatahari S (2010) A novel heuristic optimization method: charged system search. *Acta Mech* 213(3–4):267–289
16. Zheng M, Liu GX, Zhou CG, Liang YC, Wang Y (2010) Gravitation field algorithm and its application in gene cluster. *Algorithms Mol Biol* 5(1):1–11
17. Kumar S, Kumar V, Katal N, Singh SK, Sharma S, Singh P (2021) Multiarea economic dispatch using evolutionary algorithms. *Math Prob Eng*
18. Kumar V, Naresh R, Sharma V (2021) GAMS environment based solution methodologies for ramp rate constrained profit based unit commitment problem. *Iran J Sci Technol, Trans Electr Eng* 45(4):1325–1342
19. Rao RV, Savsani VJ, Vakharia DP (2011) Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. *CAD Comput Aided Des* 43(3):303–315

20. VenkataRao R (2016) Jaya: a simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int J Ind Eng Comput* 7(1):19–34
21. Kumar P, Sharma A (2022) Mrl-Jaya: a fusion of Mrlde and Jaya algorithm. *Palest J Math* 11:54–74
22. Mishra S, Ray PK (2016) Power quality improvement using photovoltaic fed DSTATCOM based on Jaya optimization. *IEEE Trans Sust Energy* 99:1–9
23. Gong C (2017) An enhanced Jaya algorithm with a two group adaption. *Int J Comput Intell Syst* 10:1102–1115
24. Yu K, Liang J, Qu B, Chen X, Wang H (2017) Parameters identification of photovoltaic models using an improved JAYA optimization algorithm. *Energy Convers Manage* 150:742–753
25. Gao K, Zhang Y, Sadollah A, Lentzakis A, Su R (2017) Jaya harmony search and water cycle algorithms for solving large-scale real-life urban traffic light scheduling problem. *Swarm Evol Comput* 37:58–72
26. Rao RV, More KC (2017) Design optimization and analysis of selected thermal devices using self-adaptive Jaya algorithm. *Energy Convers Manage* 140:24–35
27. Singh SP, Prakash T, Singh V, Babu MG (2017) Analytic hierarchy process based automatic generation control of multi-area interconnected power system using Jaya algorithm. *Eng Appl Artif Intell* 60:35–44
28. Rao RV, Saroj A (2017) Economic optimization of shell-and-tube heat exchanger using Jaya algorithm with maintenance consideration. *Swarm Evol Comput* 116:473–487
29. Venkata Rao R, Saroj A (2017) A self-adaptive multi-population based Jaya algorithm for engineering optimization. *Swarm and Evol Comput* 37:1–37
30. Rao RV, Saroj A (2018) Multi-objective design optimization of heat exchangers using elitist-Jaya algorithm. *Energy Syst* 9:305–341
31. Yu J-T, Kim C-H, Wadood A, Khurshaid T, Rhee S-B (2019) Jaya algorithm with self-adaptive multi-population and lévy flights for solving economic load dispatch problems. *IEEE Access* 7:21372–21384
32. Houssein EH, Gad AG, Wazery YM (2021) Jaya algorithm and applications: a comprehensive review. *Metaheuristics Optim Comput Electr Eng* 3–24

An Optimized Approach for Emotion Detection-Based Music Recommendation System



Manoj K. Sabnis and Bhavesh Bhatia

Abstract In the current world of cut throat competition, every company aims to get the best from their employees. For this an important factor on which the human working capability depends is his emotions, i.e. positive emotion aims for a higher output. The emotions of the person thus have to be detected. Various mechanisms have to be further applied to enhance this emotion from negative to positive, i.e. from sad or angry to happy. Thus, a system is required, which can detect the mood of the person by using the state of the art technology. On basis of this, detected mood can suggest solution so as to enhance the persons mood. This paper thus defines a system which uses face recognition for mood detection and as per the detected mood, a song list will be recommended which will further enhance the persons mood. The system is also evaluated for its performance on comparative basis and the results are evaluated on qualitative and quantitative basis.

Keywords Face recognition · Emotion · Recommendation

1 Introduction

The face of a person is the expression of his feeling, i.e. whether he is sad, happy, angry and so on. The face has eyes, colour, mouth, which convey the feelings. Certain parameters of this part of the face, if captured and analysed then the mood of the person whether angry sad or happy can be detected and can then be used further as per the need [1, 2].

This paper proposes a system design approach for product development. The first subsystem, which can capture the face of the person, from its face parts recognize and capture the required features. These features can be used to detect the persons mood. This can then act as the input to the next subsystem which will suggest songs from its database or from the Internet. This will help to enhance the persons mood from sad, angry to happy [3, 4].

M. K. Sabnis (✉) · B. Bhatia
VESIT, Collector Colony, Chembur, Mumbai 400007, India
e-mail: manoj.sabnis@ves.ac.in

The technologies used in the system are image processing for face detection, feature extraction [5], machine learning for training the system and testing and comparing for results [6].

The system also has an evaluation module which evaluates the available results qualitatively and quantitatively for its performance measure [7].

2 System Design

This is the proposed system having the input module, the output module, evaluation module and the database module. The main module is the central module which has the preprocessing and the processing module. The diagram of the proposed system is as per Fig. 1 and the details of each module are as follows.

2.1 Input Module

The input module can accept the data in two ways, namely, the user can either send his live photo or the user can also select the emoji to indicate his mood [8, 9].

Figure 2 indicates the GUI of the input screen which will take the face snap shot of the person. Figure 3 indicates the process of capturing the image. The face is set and the window is made to rest on the face so as to cover all the outer boundaries of the face. Then the image of the face is captured and sent for further processing. Figure 4 gives the second option of input capture where there are a set of emojis representing various human moods. The person can also select any one of these, if there is some camera problem.

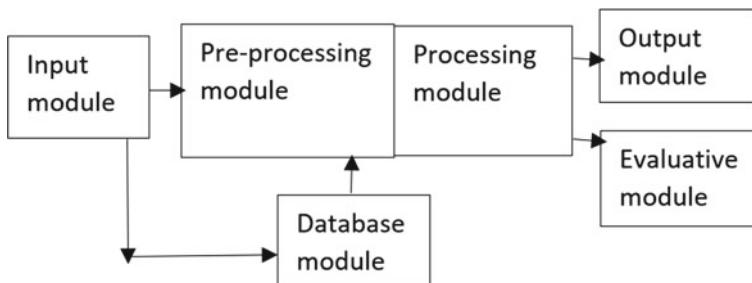
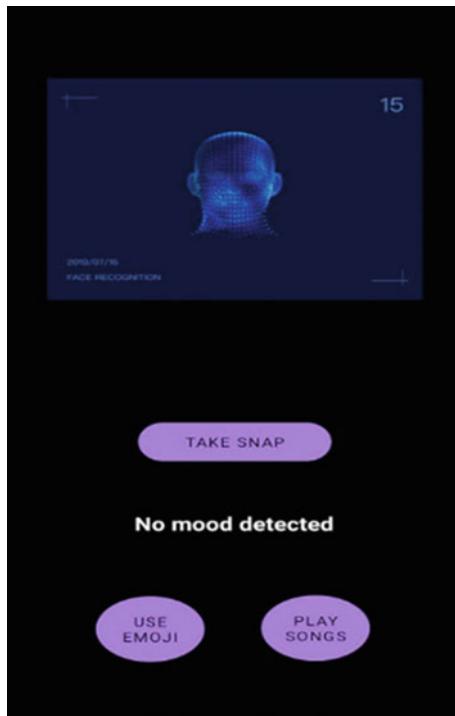


Fig. 1 Emotion-based music recommendation system

Fig. 2 Input screen

2.2 Preprocessing Module

The images captured by the webcam or the images of the selected emojis are then sent to the preprocessing unit. Here, the colour images are first cleaned for any noise by adding blurriness and then converted into grey scale images and rescaled to 28×28 pixel [10].

2.3 Processing Module

The processing module consist of three parts, the detection unit, the training unit and the comparison unit.

There are many models for face detection, these are used as a reference to select the right and the appropriate model. Local binary pattern model based on texture classification was used for facial recognition. This model can be used only for static images [11]. The tracking learning model was able to track the face and detect, but the tracking time and computational complexity was found to be high [12]. The extended version of this model, the learning-based model histogram (LBPH) recognized the face by comparing test images with the trained images. This lead to the survey of

Fig. 3 Input detected screen

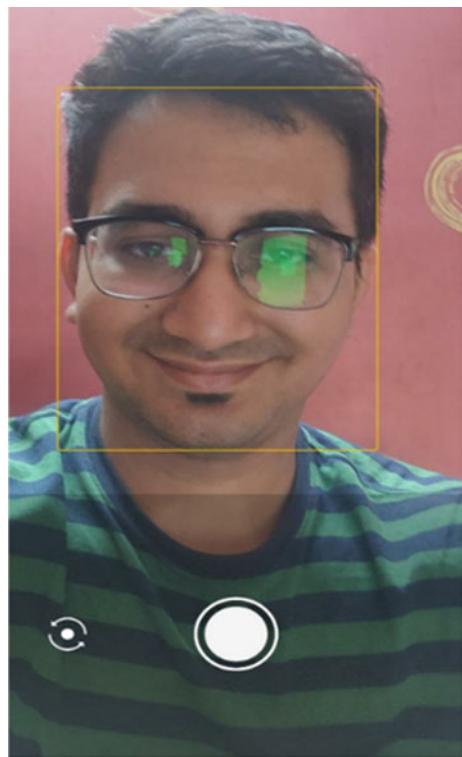
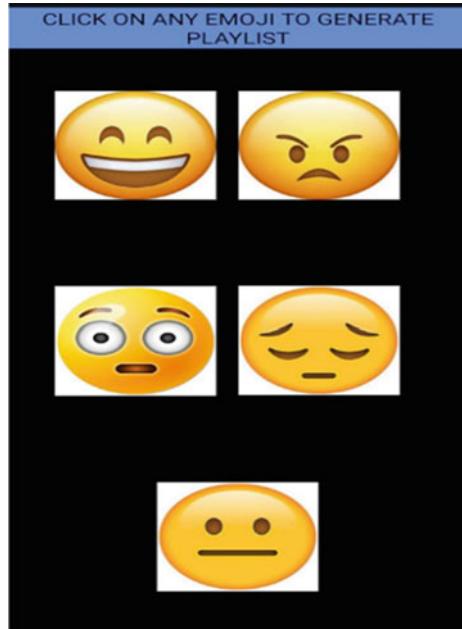


Fig. 4 Emoji selection



support vector machine (SVM) model of face recognition in the domain of machine learning. SVM reduces the structural risk where an optimal function is built for classification. This in turn reduces the prediction error. However, in case of training the SVM, the entire photo is used instead of the features detected. This increases the calculation time when large samples are considered.

Thus, to satisfy the requirements of large data set, the domain of deep learning was explored. On these lines, the first model that was proposed by Nguyen et al. [13]. This model represented three types of emotions positive, negative and normal. A single approach did not give accurate results. So a multimodal strategy was required which can detect and map the human emotions more accurately for supervised and unsupervised learning models [14].

Thus, for these reasons convolutional neural network (CNN)-based approach was selected for this work which could analyse multimodal emotional information. This could capture information from facial movements and hence intensify the decision of the system on recognizing the emotions in real time [15].

Requirements: As the system is implemented as client server model, the requirements are from the server side and the client side. The client will be the user side, thus the user should have a smart mobile phone with camera to capture the face and speaker to hear the recommended song list. The mobile phone should have operating system, Android 6.0 or above.

The server side is the main processing side having required hardware. The softwares required are, Android studio for Google android operating system, python as the programming language, firebase platform for creating web and mobile application.

Proposed algorithm: The input images acquired from the camera, after preprocessing enter the processing unit. The steps of the proposed algorithms are as follows:

Face detection: This is done in two parts, first part trains the cascaded function and the second part detects the features. Open CV is used for training and detection.

The training of the cascaded function is done with positive and negative images. Positive images are those with faces and negative ones are those without faces. After training the cascaded function is ready to detect the required objects from other images.

cv::CascadeClassifier: This creates the classifier.

CascadeClassifierface_cascade: The individual classifiers for eyes, nose, can be defined.

For detection, the clean and pre-processed image is then said to be ready as the input. From the 48×48 grey scale value rescaled image, the face is detected by using harr-based cascade classifier. They are convolutional kernel where each feature is a single value.

cv::CascadeClassifier::load: This is the method to the data file for detection.

cv::CascadeClassifier::detectMultiScale: This method will return rectangular boundaries for detected eyes, face, etc.

Feature extraction: After detection of the face, the next stage is the detection of features of the face. These features after detection can be put together to be classified

as angry, happy, sad, surprise and normal. This represents the emotions and hence the persons mood. For detecting the features, first the local binary features are marked as eyes, outer edges of the face, inner edges of the face and so on. Then pairs of landmark points are marked on these local binary features. These pairs when joint give the distance between these points. This target data set thus obtained, further goes for training and comparison.

The feature extraction module is taken from the CNN model used. This consist of convolution layers [16].

Each convolutional layer is defined with several parameters including the input size, kernel size, depth of the map stack, zero-padding and stride.

The output size can be calculated by:

$$Mx = (Ix - Kx)/Sx \quad (1)$$

$$My = (Iy - Ky)/Sy \quad (2)$$

where (Mx, My) , (Ix, Iy) , (Kx, Ky) indicate the map size, input size, kernel size, respectively, Sx, Sy indicate the stride in row & column.

A convolutional layer condenses the input by extracting the features of interest from it and produces the feature maps in response to different feature detectors as indicated above. In the first convolutional layer, simple feature as edges are filtered. The neurons of the next convolution layers learn to gather the information and gain a bigger picture of the image. Hence a high-order feature detecting is performed on each convolution layer, thus produces a feature map were all the feature maps are of the same size and thus a stack is created, were each feature map is a part of that stack.

Linearization: This is thus followed by the process of activation. It is this process that gets the output. Now, the output if directly taken will have nonlinearities due to positive and negative images. Thus to remove the nonlinearities, activation function is used which is called as rectified linear unit (ReLU), which is as follows:

$$F(x) = \max(0, x) \quad (3)$$

This gives positive value at the output if it gets positive value and 0 if it gets negative value [17].

Pooling: This is used to reduce the size of feature map if needed, i.e. the common features are generalized and represented in one window size.

Classification: The output is the final convoluted feature in terms of rows and columns, where each row represents a feature class and each column a unique vector.

Training and saving: This convoluted feature of the input image goes to the training set. If it is unique, then it is saved with its unique vector number.

Testing model: The testing model has the mapping table as shown in Table 1.

Table 1 Emotions playlist mapping

Label vector	Emotion	Playlist no
0	Angry	11: cooling song list
1	Disgust	12: happy song list
2	Fear	13: encouraging song list
3	Happy	14: normal song list
4	Sad	12: happy song list
5	Surprise	14: normal song list
6	Normal	14: normal song list

The vector map of the input image is fed in the testing model and tested against the trained labelled model. For a match the label is selected as per the detected emotions and the required song playlist is available on the client side.

2.4 Output Module

Of all the moods detected, three are shown where the mood is detected as happy, angry and surprise as shown in Figs. 5, 6 and 7.

The results of these detected moods is the playlist which will be selected and played. If the mood detected is sad then fast rhythm songs are selected to make the user feel happy as shown in Fig. 8 named as sad mood. Figure 9 named as angry

Fig. 5 Happy mood



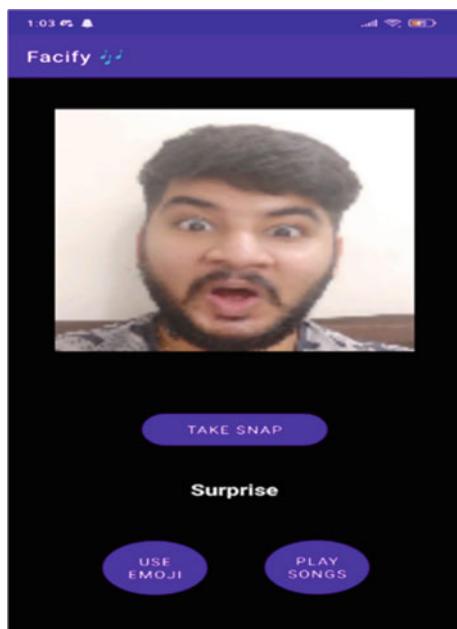
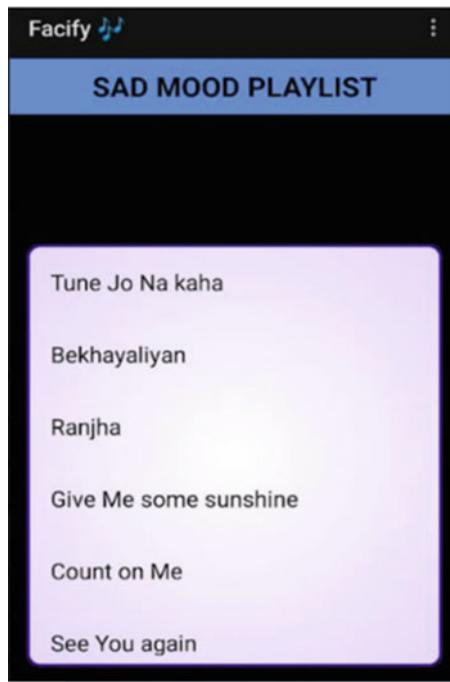
Fig. 6 Angry mood**Fig. 7** Surprise mood

Fig. 8 Sad mood

mood is provided if the mood of the user is detected as angry. The song list helps the user to release his anger slowly and make him calm. Figure 10 named as natural mood is provided when the mood detected is normal. It helps to alleviate the mood towards happy and thus increase the users efficiency.

2.5 Evaluation Module

The paper presents the proposed system evaluation under three heads, qualitative, quantitative and comparative.

Qualitative: These are the parameters that are not measured in number, but are very important for quantifying the algorithm used. These parameters are Robustness to noise (RN), Object independence (OI), Scene independence (SI), Computational load (CC) and Illumination independency (II) [18].

These parameters are quantified and represented as Absolute value measure (AVM) and Relative value measure (RVM) [19]. AVM talks about the parameters importance and RVM the relative position of the parameter with all other parameters. The numerical mapping representation is as represented as in Table 2.

The parameter mapping with respect to AVM and RVM are as follows in Table 3.

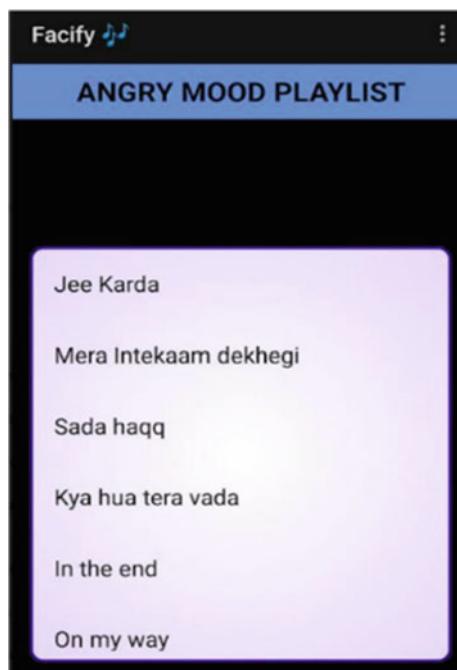
Fig. 9 Angry mood**Fig. 10** Neutral mood

Table 2 Qualitative mapping

Parameter	Numerical equivalent
Very good	4
Good	3
Average	2
Bad	1

Table 3 AVM and RVM mapping

	RN	OI	SI	CC	II
AVM	4	4	3	3	3
RVM	4	4	3	3	3

RN, Robustness to noise is high as the inner and outer curves are selected and then normalized in the latter stage. This parameter directly depends on the selected curves, thus its relative value is high. OI is the Object independency, i.e. how fast the face is detected with respect to its surroundings. As the face detection algorithm is able to select the positive images and reject the negative, thus it gives good performance as compared with others, detection is also important, thus RVM is four. SI is the scene independency, i.e. how loosely the objects within the scene are connected. For this both AVM and RVM are three. CC is the computational load which has to be low, but for this algorithm as the number of operations and number of convolutions are more and also taking the training time into account, both AVM and RVM are three. Illumination independency (II), this specifies how the illumination of the image affects its detection. This value is good, i.e. three as the faces are recognized successfully in the face of standard illumination.

Quantitative: This gives the accuracy with which the detection of features are done successfully. Of all the features, curves detection is taken for measurement.

Two accuracy types are specified for measurement, namely the Producer Accuracy, and User Accuracy [18]. These accuracies are represented as:

Producer accuracy

$$\eta_s = \text{TP}/(\text{TP} + \text{FN}) \quad (4)$$

$$\eta_n = \text{TP}/(\text{FP} + \text{TN}) \quad (5)$$

User accuracy as:

$$P_s = \text{TP}/(\text{TP} + \text{FP}) \quad (6)$$

$$P_n = \text{TN}/(\text{TN} + \text{FN}) \quad (7)$$

Table 4 Accuracy table

Images	Producer accuracy		User accuracy	
	H_s	H_n	P_s	P_n
Figure 5	0.834	0.12	0.69	0.8
Figure 6	0.675	0.23	0.65	0.7

TP: True positive, the curves detected are actual feature curves.

TN: True negative, the curves detected are not the features curves.

FP: False positive, the region not detected as curves are actual features.

FN: False positive, the region not detected as curves are not the features to be detected (Table 4).

Comparative: the accuracy of the proposed system is compared with the existing similar system of different methodologies. The proposed system accuracy was found to be almost equal to some or greater than some systems.

3 Conclusion

The accuracy of face detection can thus be increased by increasing the number of layers from outside to inside in case of deep learning. The increases in the layer number also increases the processing load. The face matching accuracy can also be improved by introducing more number of parameters along with edge detection.

It can be thus be concluded that the face is detected, matched and song list outputted with almost nearer to complete accuracy.

4 Future Scope

The future scope suggested is mainly the limitations experienced during the work due to time constraints. The entire captured image is used for training the model. This increases the overheads in calculation in terms of cost and processing time. Thus it is suggested that the feature extraction to be done first and then the model to be trained.

References

1. Singh D (2012) Human emotion recognition system. *Int J Image Graph Signal Process* 4(8). <https://doi.org/10.5815/ijigsp.2012.08.07>
2. Ruiz LZ, Alomia RPV, Dantis ADQ, San Dieg MJS (2017) Human emotion detection through facial expression for commercial analysis. In: Conference 2017, INSPEC Accession Number: 17524867, HNICEM, IEEE
3. Jonathan, Lim AP, Kusuma GP, Zahra A (2018) Facial emotion recognition using computer vision. In: IEEE 1st 2018 INAPR International Conference, Jakarta, Indonesia, 78-1-5386-9422-0/18/\$31.00 ©2018
4. Singh CB, Sarkar B, Yadav P (2021) Facial expression recognition, SSRN, Elsevier
5. Syambas NR, Purwanto UH (2012) Image processing and face detection analysis on face verification based on the age stages. In: 7th International conference on telecommunication systems, services, and applications (TSSA). <https://doi.org/10.1109/TSSA.2012.6366070IEEE>
6. Batista GEPA (2004) A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Newsl* 1. <https://doi.org/10.1145/1007730.1007735>
7. Dennis M, Fetterman D, Sechrest L (1994) Integrating qualitative and quantitative evaluation methods in substance abuse research. *Eval Program Plan* 17(4):419–427
8. Rajhi M (2017) Emotional recognition using facial expression by emoji in real time, Copyright © 2017 GRIN Verlag ISBN: 9783668728943, <https://www.grin.com/document/379502>
9. Swapna Goud N, Revanth Reddy K, Alekhya G, Sucheta GS (2019) Facial emoji recognition. *Int J Trend Sci Res Dev (IJTSRD)* 3(3), Online: www.ijtsrd.com e-ISSN: 2456-6470
10. Verni J, Image pre-processing, Chapter 2. <https://www.embedded-vision.com>
11. Meanpaa T, Pietikeinian M (2015) Texture analysis with local binary patterns. Res Gate. https://doi.org/10.1142/9789812775320_0011
12. Zou Y, Yan X, Li W (2022) Knowledge tracking model based on learning process. *J Comput Commun* 2327-5219 ISSN Online: 2327-5227 www.scirp.org/journal/jcc E-mail: jcc@scirp.org
13. Bui DT, Tsangaratos P, Nguyen VT, Van Liem N (2020) Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. Elsevier, Cantana
14. Satya R, Abraham A (2013) Comparison of supervised and unsupervised learning algorithms for pattern classification. *Int J Adv Res Artif Intell*
15. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Springer
16. O’Shea K, Nash R (2015) An introduction to convolutional neural networks. Res Gate
17. Agarap AF (2018) Deep learning using rectified linear units (ReLU). Res Gate
18. Sanin A, Sanderson C, Lovell BC (2012) Shadow detection: a survey and comparative evaluation of recent methods. *Pattern Recogn* 45(4):1684–1695
19. Kavita (2015) A comparative approach for standard shadow detection methods. Res Gate

A Method to Solve Fractional Transportation Problems with Rough Interval Parameters



Shivani and Deepika Rani

Abstract While solving real-world transportation problems, a decision-maker has to face the uncertainty and/or hesitations to define the input parameters. To deal with such situations, rough set theory is a significant tool as it includes the agreement and understanding of all the experts. In this study, a mathematical model of fractional transportation problem is developed in which all the coefficients and decision variables are rough intervals. To solve the problem, a new method is proposed in which the problem model is decomposed into two sub-models: the upper interval model and the lower interval model. These two sub-models are further solved to get the optimal rough interval solution. At last, a numerical example is solved to demonstrate the applicability of the proposed methodology in the areas of transportation and decision-making.

Keywords Fractional transportation problem · Rough interval · Upper interval model · Lower interval model · Optimal solution

1 Introduction

The transportation problem is a special type of linear programming problem which was developed by Hitchcock [1] in 1941. The main objective of the transportation problem is to transport the optimal quantity of the product from various sources to different destinations at the minimum possible transportation cost. Numerous strategies [2–5] have been developed in the literature to solve the transportation problems. In a real-life transportation system, sometimes, objectives of the problems may be fractional such as profit/cost, actual cost/standard cost, and actual time/standard time. These problems lie under the category of fractional transportation problem (FTP),

Shivani (✉) · D. Rani

Department of Mathematics, Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab 144027, India
e-mail: sainishivani2310@gmail.com

D. Rani
e-mail: ranid@nitj.ac.in

which was investigated by Swarup [6] in 1966. The efficiency of any system is represented by the interpretation of FTP as the ratio of physical and economic values. In literature, there are several studies [7–9] on the FTP.

In traditional FTP, the parameters are defined experimentally or subjectively at precise values, depending on the expert's understanding about the nature of the parameters. However, the coefficients of the real-life FTP include uncertainty due to lack of information, inaccuracy in judgment, environmental factors, etc. Therefore, solving the FTP under the aspect of uncertainty is very valuable. The fuzzy set theory introduced by Zadeh [10] in 1965 effectively cope up with the uncertainties that arise due to human imprecision. In fuzzy set theory, the degree of membership $\in [0, 1]$ is used to describe the uncertainty of the parameters. Numerous researchers [11–14] have employed fuzzy set theory to address the uncertainty in FTP. Unlike the fuzzy set theory, the degree to which an element belongs to the set of feasible solutions may not always be best represented by only the membership function. Therefore, an intuitionistic fuzzy set, a generalization and extension of the fuzzy set, was developed by Atanassov [15] in 1986. In intuitionistic fuzzy set, the uncertainty of the parameters is characterized by both the membership and the non-membership degree in such a manner that the total of the membership degree and the non-membership degree is less than or equal to one. In this regard, many researchers [16–19] have used intuitionistic fuzzy set theory to handle the uncertainty of FTP.

Although there are approaches suggested in the literature to deal with the uncertain structure of the FTP, even then there are continuing efforts to explore other uncertainties related fractional transportation problems (FTPs). In this aspect, the rough set theory developed by Pawlak [20] can be used to express the vagueness and imprecision that is caused by ambiguous information about the parameters of FTP. Rough set theory plays a crucial role for describing the uncertainty or ambiguity in optimization problems as it is free from any additional data-related information. As a result, the optimization procedure will become flexible and effective if the decision-maker uses the rough set theory to address ambiguity in optimization problems. However, the rough set has the limitation of being able to handle only discrete data and being unable to handle the continuous variables. To overcome this shortcoming of the rough set theory, Rebolledo [21] invented the idea of rough interval, a specific case of the rough set. A rough interval satisfies all of the characteristics of the rough set and is also capable of characterizing continuous variables. To deal with the uncertainty in real-world applications, many researchers [22–26] have used rough set theory in different fields. Also, various approaches have been developed in the literature to handle the uncertainty of the transportation problem using rough set tactics such as Bera et al. [27] solved the profit maximization four-dimensional transportation problem in the context of rough intervals. Midya and Roy [28] have used three distinct strategies to solve the multi-objective fixed-charge transportation problem with rough parameters. The rough and bi-rough variables are used by Bera and Mondal [29] to deal with the uncertainty of two-stage multi-objective transportation problem. To solve the rough interval multi-objective transportation problem, Garg and Rizk-Allah [30] suggested an innovative strategy. Shivani et al. [31] developed a solution mechanism for the unbalanced fully rough multi-objective fixed-charge transportation problem.

The term fully rough fractional transportation problem (FRFTP) refers to a FTP where all the parameters and decision variables are represented as rough intervals. In the present study, a new approach is proposed to solve the FRFTP under roughness tactics which depict all experts opinions in a rough interval environment. The rough interval scenario is used to specify all the parameters, including supply, demand, transportation costs, and profit. The novelty of the proposed technique is in the construction of a new formulation for the FRFTP based on rough set theory. This interpretation differs from the literary formulations, making our task original as a result.

Highlights of the Proposed Study

- (i) To represent the fuzziness or uncertainty, the model of FRFTP is formulated on the basis of the rough set theory.
- (ii) To solve the FRFTP, a new solution methodology is proposed which operates by decomposing the problem into two sub-models, the upper interval model (UIM) means respecting the information of any decision-maker, and the lower interval model (LIM) signifies the information of all decision-makers. The boundaries of these interval models are then used to decompose each interval sub-model into two separate crisp models.
- (iii) A numerical example is solved for the investigations and validations of the proposed methodology.
- (iv) To deal with real-life FTP with opposing and consenting knowledge, the proposed model can be a very useful tool for the decision-maker.

The rest of the paper is organized as follows: Sect. 2 depicts the basic definitions related to the rough set theory. The mathematical model of the proposed FRFTP is formulated in Sect. 3. Section 4 describes the solution methodology to solve the proposed FRFTP. A numerical example is solved in Sect. 5. Section 6 presents the conclusions of this study with future research directions.

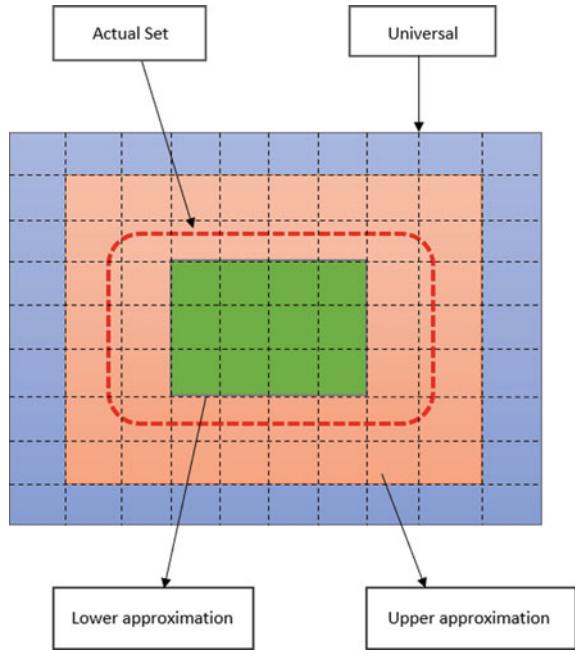
2 Preliminaries

In this section, some basic definitions related to the rough set theory are described.

Definition 1 [20] Let Ψ be a non-empty set, \mathcal{B} be a σ -algebra of subset of Ψ , Ω be an element in \mathcal{B} , and ω be positive, real-valued, additive set function. Then, $(\Psi, \Omega, \mathcal{B}, \omega)$ is called a rough space.

Definition 2 [32] Let U be a non-empty finite set and R be an equivalence relation on U and let $R(w)$ be an equivalence class of the relation that includes $w \in U$. Then, for any set $W \subseteq U$, the lower and the upper approximations of W are given as follows:

Fig. 1 Graphical depiction of rough set



(i) R -lower approximation set

$$\underline{R}(W) = \bigcup_{w \in W} \{R(w) : R(w) \subseteq W\}$$

(ii) R -upper approximation set

$$\overline{R}(W) = \bigcup_{w \in W} \{R(w) : R(w) \cap W \neq \emptyset\}$$

(iii) R -boundary region of W

$$bn_R(W) = \overline{R}(W) - \underline{R}(W).$$

The pictorial representation of the rough set is given in Fig. 1.

Definition 3 [32] Let τ^{LL} , τ^{UL} , τ^{LU} , τ^{UU} all are real numbers and $\tau^{LU} \leq \tau^{LL} \leq \tau^{UL} \leq \tau^{UU}$, then the value $\tilde{P}^{RI} = [\tau^{LL}, \tau^{UL}][\tau^{LU}, \tau^{UU}]$ is termed as rough interval. The interval $[\tau^{LL}, \tau^{UL}]$ is called lower approximation interval, and $[\tau^{LU}, \tau^{UU}]$ is called upper approximation interval such that

- (i) If $w \in [\tau^{LL}, \tau^{UL}]$, then \tilde{P}^{RI} definitely takes w .
- (ii) If $w \in [\tau^{LU}, \tau^{UU}]$, then \tilde{P}^{RI} probably takes w .
- (iii) If $w \notin [\tau^{LU}, \tau^{UU}]$, then \tilde{P}^{RI} definitely does not takes w .

Example 1 Let η represents the “waste generation rate” in a city which typically varies from 700 to 720 ton/day. During holidays, festivals, and other special occasions, the amount of waste generated per day varies from 680 to 740 ton/day.

This “waste generation rate” can be approximated by the interval as: $\eta = (\underline{\eta}, \bar{\eta}) = [700, 720][680, 740]$. It implies that the waste generation rate in a city is surely between [700, 720] ton/day, and it is possibly between [680, 740] ton/day.

Definition 4 A rough interval $\tilde{P}^{\text{RI}} = [\tau^{\text{LL}}, \tau^{\text{UL}}][\tau^{\text{LU}}, \tau^{\text{UU}}]$ is said to be positive iff $\tau^{\text{LU}} \geq 0$.

Definition 5 Let $\tilde{P}^{\text{RI}} = [\tau^{\text{LL}}, \tau^{\text{UL}}][\tau^{\text{LU}}, \tau^{\text{UU}}]$ and $\tilde{Q}^{\text{RI}} = [\phi^{\text{LL}}, \phi^{\text{UL}}][\phi^{\text{LU}}, \phi^{\text{UU}}]$ be two positive rough intervals and α be a real number. Then

- (i) $\tilde{P}^{\text{RI}} \oplus \tilde{Q}^{\text{RI}} = [\tau^{\text{LL}} + \phi^{\text{LL}}, \tau^{\text{UL}} + \phi^{\text{UL}}][\tau^{\text{LU}} + \phi^{\text{LU}}, \tau^{\text{UU}} + \phi^{\text{UU}}]$,
- (ii) $\tilde{P}^{\text{RI}} \ominus \tilde{Q}^{\text{RI}} = [\tau^{\text{LL}} - \phi^{\text{UL}}, \tau^{\text{UL}} - \phi^{\text{LL}}][\tau^{\text{LU}} - \phi^{\text{UU}}, \tau^{\text{UU}} - \phi^{\text{LU}}]$,
- (iii) $\alpha \tilde{P}^{\text{RI}} = \begin{cases} [\alpha \tau^{\text{LL}}, \alpha \tau^{\text{UL}}][\alpha \tau^{\text{LU}}, \alpha \tau^{\text{UU}}], & \text{if } \alpha \geq 0 \\ [\alpha \tau^{\text{UL}}, \alpha \tau^{\text{LL}}][\alpha \tau^{\text{UU}}, \alpha \tau^{\text{LU}}], & \text{if } \alpha < 0 \end{cases}$,
- (iv) $\tilde{P}^{\text{RI}} \otimes \tilde{Q}^{\text{RI}} = [\tau^{\text{LL}} \phi^{\text{LL}}, \tau^{\text{UL}} \phi^{\text{UL}}][\tau^{\text{LU}} \phi^{\text{LU}}, \tau^{\text{UU}} \phi^{\text{UU}}]$,
- (v) $\tilde{P}^{\text{RI}} \oslash \tilde{Q}^{\text{RI}} = [\tau^{\text{LL}}/\phi^{\text{UL}}, \tau^{\text{UL}}/\phi^{\text{LL}}][\tau^{\text{LU}}/\phi^{\text{UU}}, \tau^{\text{UU}}/\phi^{\text{LU}}]$.

Example 2 Let $\tilde{P}^{\text{RI}} = [5, 7][4, 9]$ and $\tilde{Q}^{\text{RI}} = [2, 5][1, 7]$ be two rough intervals. Then, $\tilde{P}^{\text{RI}} + \tilde{Q}^{\text{RI}} = [7, 12][5, 16]$, $\tilde{P}^{\text{RI}} - \tilde{Q}^{\text{RI}} = [0, 5][-3, 8]$, $\tilde{P}^{\text{RI}} * \tilde{Q}^{\text{RI}} = [10, 35][4, 63]$, $\tilde{P}^{\text{RI}} / \tilde{Q}^{\text{RI}} = [1, 7/2][4/7, 9]$, $4\tilde{P}^{\text{RI}} = [20, 28][16, 36]$, and $-4\tilde{P}^{\text{RI}} = [-28, -20][-36, -16]$.

Definition 6 Let $\tilde{P}^{\text{RI}} = [\tau^{\text{LL}}, \tau^{\text{UL}}][\tau^{\text{LU}}, \tau^{\text{UU}}]$ and $\tilde{Q}^{\text{RI}} = [\phi^{\text{LL}}, \phi^{\text{UL}}][\phi^{\text{LU}}, \phi^{\text{UU}}]$ are two rough intervals, then $\tilde{P}^{\text{RI}} = \tilde{Q}^{\text{RI}}$ iff $\tau^{\text{LL}} = \phi^{\text{LL}}, \tau^{\text{UL}} = \phi^{\text{UL}}, \tau^{\text{LU}} = \phi^{\text{LU}}, \tau^{\text{UU}} = \phi^{\text{UU}}$.

3 Mathematical Model of Fully Rough Fractional Transportation Problem (FRFTP)

In this section, the model of FRFTP is formulated in which the objective function represents the ratio of the total profit to the total transportation cost. In the proposed model, there are s ($g = 1, 2, \dots, s$) sources and t ($h = 1, 2, \dots, t$) destinations. The aim of the decision-maker is to determine the number of units of the item to be transported from g th source to h th destination to obtain the maximum value of the objective function, which is represented by Eq. (1). The supply, demand, and non-negativity constraints are represented by Eqs. (2), (3), and (4), respectively. The following notations and assumptions are used to formulate the proposed mathematical model:

Notations

s : number of sources

t : number of destinations

$\tilde{\rho}_{gh}^{\text{RI}}$: profit earned from shipment of the unit quantity of the item from g th source to h th destination

$\tilde{\chi}_{gh}^{\text{RI}}$: unit transportation cost of the item from g th source to h th destination

$\tilde{\alpha}_g^{\text{RI}}$: availability of the item at g th source

$\tilde{\beta}_h^{\text{RI}}$: demand of the item at h th destination

$\tilde{\xi}_{gh}^{\text{RI}}$: units of the item transported from g th source to h th destination.

Assumptions

- (i) The proposed problem is balanced problem, i.e., $\sum_{g=1}^s \tilde{\alpha}_g^{\text{RI}} = \sum_{h=1}^t \tilde{\beta}_h^{\text{RI}}$.
- (ii) All the parameters of the proposed model such as profit, cost, availability of sources, demand at destinations, and decision variables are rough interval.
- (iii) $\tilde{\rho}_{gh}^{\text{RI}} = [\rho_{gh}^{\text{LL}}, \rho_{gh}^{\text{UL}}][\rho_{gh}^{\text{LU}}, \rho_{gh}^{\text{UU}}]$, $\tilde{\chi}_{gh}^{\text{RI}} = [\chi_{gh}^{\text{LL}}, \chi_{gh}^{\text{UL}}][\chi_{gh}^{\text{LU}}, \chi_{gh}^{\text{UU}}]$, $\tilde{\xi}_{gh}^{\text{RI}} = [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}]$, $\tilde{\alpha}_g^{\text{RI}} = [\alpha_g^{\text{LL}}, \alpha_g^{\text{UL}}][\alpha_g^{\text{LU}}, \alpha_g^{\text{UU}}]$, $\tilde{\beta}_h^{\text{RI}} = [\beta_h^{\text{LL}}, \beta_h^{\text{UL}}][\beta_h^{\text{LU}}, \beta_h^{\text{UU}}]$.
- (iv) $\sum_{g=1}^s \sum_{h=1}^t [\chi_{gh}^{\text{LL}}, \chi_{gh}^{\text{UL}}][\chi_{gh}^{\text{LU}}, \chi_{gh}^{\text{UU}}] \otimes [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] > 0$.

Using these notations and assumptions, the mathematical model of the proposed FRFTP is formulated as follows:

Model 1

$$\text{Maximize } \tilde{\Delta}^{\text{RI}}(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t [\rho_{gh}^{\text{LL}}, \rho_{gh}^{\text{UL}}][\rho_{gh}^{\text{LU}}, \rho_{gh}^{\text{UU}}] \otimes [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}]}{\sum_{g=1}^s \sum_{h=1}^t [\chi_{gh}^{\text{LL}}, \chi_{gh}^{\text{UL}}][\chi_{gh}^{\text{LU}}, \chi_{gh}^{\text{UU}}] \otimes [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}]} \quad (1)$$

subject to

$$\sum_{h=1}^t [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] = [\alpha_g^{\text{LL}}, \alpha_g^{\text{UL}}][\alpha_g^{\text{LU}}, \alpha_g^{\text{UU}}], \quad g = 1, 2, \dots, s \quad (2)$$

$$\sum_{g=1}^s [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] = [\beta_h^{\text{LL}}, \beta_h^{\text{UL}}][\beta_h^{\text{LU}}, \beta_h^{\text{UU}}], \quad h = 1, 2, \dots, t \quad (3)$$

$$[\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}][\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] \geq 0, \quad g = 1, 2, \dots, s; \quad h = 1, 2, \dots, t \quad (4)$$

4 Solution Methodology

In this section, a new methodology is proposed to solve the FRFTP, i.e., Model 1. The stepwise procedure of the proposed methodology is defined below:

Step 1: Decompose the proposed problem, i.e., Model 1 into two sub-models: UIM and LIM to represent the possibly optimal solution and the surely optimal solution, respectively.

UIM is defined by using the upper interval of the Model 1 and is presented as follows:

$$\text{UIM: Maximize } \Delta^U(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t [\rho_{gh}^{\text{LU}}, \rho_{gh}^{\text{UU}}] \otimes [\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}]}{\sum_{g=1}^s \sum_{h=1}^t [\chi_{gh}^{\text{LU}}, \chi_{gh}^{\text{UU}}] \otimes [\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}]}$$

subject to

$$\sum_{h=1}^t [\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] = [\alpha_g^{\text{LU}}, \alpha_g^{\text{UU}}], \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s [\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] = [\beta_h^{\text{LU}}, \beta_h^{\text{UU}}], \quad h = 1, 2, \dots, t$$

$$[\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}}] \geq 0, \quad g = 1, 2, \dots, s; \quad h = 1, 2, \dots, t$$

And LIM is defined by using the lower interval of the Model 1 and is presented as follows:

$$\text{LIM: Maximize } \Delta^L(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t [\rho_{gh}^{\text{LL}}, \rho_{gh}^{\text{UL}}] \otimes [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}]}{\sum_{g=1}^s \sum_{h=1}^t [\chi_{gh}^{\text{LL}}, \chi_{gh}^{\text{UL}}] \otimes [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}]}$$

subject to

$$\sum_{h=1}^t [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}] = [\alpha_g^{\text{LL}}, \alpha_g^{\text{UL}}], \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s [\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}] = [\beta_h^{\text{LL}}, \beta_h^{\text{UL}}], \quad h = 1, 2, \dots, t$$

$$[\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}}] \geq 0, \quad g = 1, 2, \dots, s; \quad h = 1, 2, \dots, t$$

Step 2: Use the UIM and the LIM to construct four crisp FTPs, named as upper upper fractional transportation problem (UUFTP), upper lower fractional transportation problem (ULFTP), lower lower fractional transportation problem (LLFTP), and lower upper fractional transportation problem (LUFTP) as follows:

$$\mathbf{UUFTP:} \text{Maximize } \Delta^{\text{UU}}(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t \rho_{gh}^{\text{UU}} \otimes \xi_{gh}^{\text{UU}}}{\sum_{g=1}^s \sum_{h=1}^t \chi_{gh}^{\text{LU}} \otimes \xi_{gh}^{\text{LU}}}$$

subject to

$$\sum_{h=1}^t \xi_{gh}^{\text{UU}} = \alpha_g^{\text{UU}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{UU}} = \beta_h^{\text{UU}}, \quad h = 1, 2, \dots, t$$

$$\sum_{h=1}^t \xi_{gh}^{\text{LU}} = \alpha_g^{\text{LU}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{LU}} = \beta_h^{\text{LU}}, \quad h = 1, 2, \dots, t$$

$$\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}} \geq 0, \quad \xi_{gh}^{\text{LU}} \leq \xi_{gh}^{\text{UU}} \quad \forall g, h.$$

$$\mathbf{ULFTP:} \text{Maximize } \Delta^{\text{UL}}(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t \rho_{gh}^{\text{UL}} \otimes \xi_{gh}^{\text{UL}}}{\sum_{g=1}^s \sum_{h=1}^t \chi_{gh}^{\text{LL}} \otimes \xi_{gh}^{\text{LL}}}$$

subject to

$$\sum_{h=1}^t \xi_{gh}^{\text{UL}} = \alpha_g^{\text{UL}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{UL}} = \beta_h^{\text{UL}}, \quad h = 1, 2, \dots, t$$

$$\sum_{h=1}^t \xi_{gh}^{\text{LL}} = \alpha_g^{\text{LL}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{LL}} = \beta_h^{\text{LL}}, \quad h = 1, 2, \dots, t$$

$$\xi_{gh}^{\text{LL}} \leq \xi_{gh}^{\text{UL}}, \quad \xi_{gh}^{\text{UL}} \leq \xi_{gh}^{\text{UU}*}, \quad \xi_{gh}^{\text{LL}} \geq \xi_{gh}^{\text{LU}*} \quad \forall g, h.$$

$$\textbf{LLFTP: Maximize } \Delta^{\text{LL}}(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t \rho_{gh}^{\text{LL}} \otimes \xi_{gh}^{\text{LL}}}{\sum_{g=1}^s \sum_{h=1}^t \chi_{gh}^{\text{UL}} \otimes \xi_{gh}^{\text{UL}}}$$

subject to

$$\sum_{h=1}^t \xi_{gh}^{\text{LL}} = \alpha_g^{\text{LL}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{LL}} = \beta_h^{\text{LL}}, \quad h = 1, 2, \dots, t$$

$$\sum_{h=1}^t \xi_{gh}^{\text{UL}} = \alpha_g^{\text{UL}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{UL}} = \beta_h^{\text{UL}}, \quad h = 1, 2, \dots, t$$

$$\xi_{gh}^{\text{LL}}, \xi_{gh}^{\text{UL}} \geq 0, \quad \xi_{gh}^{\text{LL}} \leq \xi_{gh}^{\text{UL}*} \quad \forall g, h.$$

$$\textbf{LUFTP: Maximize } \Delta^{\text{LU}}(\xi) = \frac{\sum_{g=1}^s \sum_{h=1}^t \rho_{gh}^{\text{LU}} \otimes \xi_{gh}^{\text{LU}}}{\sum_{g=1}^s \sum_{h=1}^t \chi_{gh}^{\text{UU}} \otimes \xi_{gh}^{\text{UU}}}$$

subject to

$$\sum_{h=1}^t \xi_{gh}^{\text{LU}} = \alpha_g^{\text{LU}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{LU}} = \beta_h^{\text{LU}}, \quad h = 1, 2, \dots, t$$

$$\sum_{h=1}^t \xi_{gh}^{\text{UU}} = \alpha_g^{\text{UU}}, \quad g = 1, 2, \dots, s$$

$$\sum_{g=1}^s \xi_{gh}^{\text{UU}} = \beta_h^{\text{UU}}, \quad h = 1, 2, \dots, t$$

$$\xi_{gh}^{\text{LU}}, \xi_{gh}^{\text{UU}} \geq 0, \quad \xi_{gh}^{\text{LU}} \leq \xi_{gh}^{\text{LU}*} \quad \forall g, h.$$

Step 3: Solve the problem UUFTP to obtain the optimal solution $\xi_{gh}^{\text{UU}*}$ and $\xi_{gh}^{\text{LU}*}$ with objective function value $\Delta^{\text{UU}*}$.

Step 4: Solve the problem ULFTP to obtain the optimal solution $\xi_{gh}^{\text{UL}*}$ and $\xi_{gh}^{\text{LL}*}$ with objective function value $\Delta^{\text{UL}*}$.

Step 5: Solve the problem LLFTP by substituting $\xi_{gh}^{\text{LL}*}$ and $\xi_{gh}^{\text{UL}*}$ from Step 4 to obtain the objective function value $\Delta^{\text{LL}*}$.

Table 1 Profit ($\tilde{\rho}_{gh}^{\text{RI}}$) and the transportation cost ($\tilde{\lambda}_{gh}^{\text{RI}}$)

	1	2	3	4	Supply
1	[7, 7][6, 8]	[10, 11][10, 13]	[4, 6][3, 7]	[7, 10][7, 12]	[10, 12][9, 13]
	[11, 14][9, 14]	[8, 10][8, 10]	[11, 14][11, 16]	[4, 6][2, 8]	
2	[6, 6][5, 7]	[9, 11][7, 13]	[11, 14][8, 15]	[4, 7][4, 9]	[15, 19][14, 20]
	[7, 9][6, 10]	[3, 5][3, 5]	[9, 12][8, 15]	[9, 11][7, 13]	
3	[7, 9][6, 10]	[4, 5][3, 8]	[14, 15][11, 16]	[8, 9][6, 9]	[15, 17][14, 18]
	[12, 13][9, 14]	[13, 15][12, 16]	[10, 11][8, 15]	[7, 10][6, 13]	
Demand	[8, 10][7, 11]	[6, 8][5, 9]	[12, 14][11, 15]	[14, 16][14, 16]	

Step 6: Solve the problem LUFTP by substituting $\xi_{gh}^{\text{LU}*}$ and $\xi_{gh}^{\text{UU}*}$ from Step 3 to obtain the objective function value $\Delta^{\text{LU}*}$.

Step 7: The set of rough intervals $\{\tilde{\xi}_{gh}^{\text{RI}*}\} = \{[\xi_{gh}^{\text{LL}*}, \xi_{gh}^{\text{UL}*}][\xi_{gh}^{\text{LU}*}, \xi_{gh}^{\text{UU}*}]\}$ is an optimal solution to the given FRFTP with the objective value $(\tilde{\Delta}^{\text{RI}*}) = [\Delta^{\text{LL}*}, \Delta^{\text{UL}*}][\Delta^{\text{LU}*}, \Delta^{\text{UU}*}]$.

5 Numerical Example

In this section, a numerical problem has been solved to show the applicability of the proposed methodology. In this problem, there are three sources from which the items are supplied to four destinations. The main objective of this problem is to transport the items from sources to destinations by maximizing the ratio of total profit to total transportation cost. The profit, cost of transportation, supply of sources, and demand at the destinations are taken as rough intervals due to the incomplete information about the market conditions and are shown in Table 1.

For finding the rough optimal solution, the problem is solved using the solution methodology defined in Sect. 4. According to Step 1 of the proposed methodology, decompose the problem into two sub-problems, namely the UIM and the LIM. From these two sub-problems, four crisp FTPs have been constructed as follows:

UUFTP

$$\text{Max } \Delta^{\text{UU}} = \frac{8\xi_{11}^{\text{UU}} + 13\xi_{12}^{\text{UU}} + 7\xi_{13}^{\text{UU}} + 12\xi_{14}^{\text{UU}} + 7\xi_{21}^{\text{UU}} + 13\xi_{22}^{\text{UU}} + 15\xi_{23}^{\text{UU}} + 9\xi_{24}^{\text{UU}} + 10\xi_{31}^{\text{UU}} + 8\xi_{32}^{\text{UU}} + 16\xi_{33}^{\text{UU}} + 9\xi_{34}^{\text{UU}}}{9\xi_{11}^{\text{LU}} + 8\xi_{12}^{\text{LU}} + 11\xi_{13}^{\text{LU}} + 2\xi_{14}^{\text{LU}} + 6\xi_{21}^{\text{LU}} + 3\xi_{22}^{\text{LU}} + 8\xi_{23}^{\text{LU}} + 7\xi_{24}^{\text{LU}} + 9\xi_{31}^{\text{LU}} + 12\xi_{32}^{\text{LU}} + 8\xi_{33}^{\text{LU}} + 6\xi_{34}^{\text{LU}}}$$

subject to

$$\begin{aligned}
 \xi_{11}^{\text{UU}} + \xi_{12}^{\text{UU}} + \xi_{13}^{\text{UU}} + \xi_{14}^{\text{UU}} &= 13, \quad \xi_{21}^{\text{UU}} + \xi_{22}^{\text{UU}} + \xi_{23}^{\text{UU}} + \xi_{24}^{\text{UU}} = 20, \\
 \xi_{31}^{\text{UU}} + \xi_{32}^{\text{UU}} + \xi_{33}^{\text{UU}} + \xi_{34}^{\text{UU}} &= 18, \\
 \xi_{11}^{\text{UU}} + \xi_{12}^{\text{UU}} + \xi_{31}^{\text{UU}} &= 11, \quad \xi_{12}^{\text{UU}} + \xi_{22}^{\text{UU}} + \xi_{32}^{\text{UU}} = 9, \quad \xi_{13}^{\text{UU}} + \xi_{23}^{\text{UU}} + \xi_{33}^{\text{UU}} = 15, \\
 \xi_{14}^{\text{UU}} + \xi_{24}^{\text{UU}} + \xi_{34}^{\text{UU}} &= 16, \\
 \xi_{11}^{\text{LU}} + \xi_{12}^{\text{LU}} + \xi_{13}^{\text{LU}} + \xi_{14}^{\text{LU}} &= 9, \quad \xi_{21}^{\text{LU}} + \xi_{22}^{\text{LU}} + \xi_{23}^{\text{LU}} + \xi_{24}^{\text{LU}} = 14, \\
 \xi_{31}^{\text{LU}} + \xi_{32}^{\text{LU}} + \xi_{33}^{\text{LU}} + \xi_{34}^{\text{LU}} &= 14, \\
 \xi_{11}^{\text{LU}} + \xi_{21}^{\text{LU}} + \xi_{31}^{\text{LU}} &= 7, \quad \xi_{12}^{\text{LU}} + \xi_{22}^{\text{LU}} + \xi_{32}^{\text{LU}} = 5, \quad \xi_{13}^{\text{LU}} + \xi_{23}^{\text{LU}} + \xi_{33}^{\text{LU}} = 11, \\
 \xi_{14}^{\text{LU}} + \xi_{24}^{\text{LU}} + \xi_{34}^{\text{LU}} &= 14, \\
 \xi_{gh}^{\text{UU}}, \quad \xi_{gh}^{\text{LU}} \geq 0 \quad \xi_{gh}^{\text{LU}} \leq \xi_{gh}^{\text{UU}} \quad \forall g = 1, 2, 3; \quad h = 1, 2, 3, 4
 \end{aligned}$$

Solving the problem UUFTP by LINGO 18.0 optimization solver, we have

$$\begin{aligned}
 \xi_{11}^{\text{UU}} &= 0, \quad \xi_{12}^{\text{UU}} = 2, \quad \xi_{13}^{\text{UU}} = 0, \quad \xi_{14}^{\text{UU}} = 11, \quad \xi_{21}^{\text{UU}} = 7, \quad \xi_{22}^{\text{UU}} = 7, \quad \xi_{23}^{\text{UU}} = 6, \quad \xi_{24}^{\text{UU}} = 0, \\
 \xi_{31}^{\text{UU}} &= 4, \quad \xi_{32}^{\text{UU}} = 0, \quad \xi_{33}^{\text{UU}} = 9, \quad \xi_{34}^{\text{UU}} = 5, \\
 \xi_{11}^{\text{LU}} &= 0, \quad \xi_{12}^{\text{LU}} = 0, \quad \xi_{13}^{\text{LU}} = 0, \quad \xi_{14}^{\text{LU}} = 9, \quad \xi_{21}^{\text{LU}} = 7, \quad \xi_{22}^{\text{LU}} = 5, \quad \xi_{23}^{\text{LU}} = 2, \quad \xi_{24}^{\text{LU}} = 0, \\
 \xi_{31}^{\text{LU}} &= 0, \quad \xi_{32}^{\text{LU}} = 0, \quad \xi_{33}^{\text{LU}} = 9, \quad \xi_{34}^{\text{LU}} = 5.
 \end{aligned}$$

At this solution, the objective function value is $\Delta^{\text{UU}*} = 3.1969$.

ULFTP

$$\text{Max } \Delta^{\text{UL}} = \frac{7\xi_{11}^{\text{UL}} + 11\xi_{12}^{\text{UL}} + 6\xi_{13}^{\text{UL}} + 10\xi_{14}^{\text{UL}} + 6\xi_{21}^{\text{UL}} + 11\xi_{22}^{\text{UL}} + 14\xi_{23}^{\text{UL}} + 7\xi_{24}^{\text{UL}} + 9\xi_{31}^{\text{UL}} + 5\xi_{32}^{\text{UL}} + 15\xi_{33}^{\text{UL}} + 9\xi_{34}^{\text{UL}}}{11\xi_{11}^{\text{LL}} + 8\xi_{12}^{\text{LL}} + 11\xi_{13}^{\text{LL}} + 4\xi_{14}^{\text{LL}} + 7\xi_{21}^{\text{LL}} + 3\xi_{22}^{\text{LL}} + 9\xi_{23}^{\text{LL}} + 9\xi_{24}^{\text{LL}} + 12\xi_{31}^{\text{LL}} + 13\xi_{32}^{\text{LL}} + 10\xi_{33}^{\text{LL}} + 7\xi_{34}^{\text{LL}}}$$

subject to

$$\begin{aligned}
 \xi_{11}^{\text{UL}} + \xi_{12}^{\text{UL}} + \xi_{13}^{\text{UL}} + \xi_{14}^{\text{UL}} &= 12, \quad \xi_{21}^{\text{UL}} + \xi_{22}^{\text{UL}} + \xi_{23}^{\text{UL}} + \xi_{24}^{\text{UL}} = 19, \\
 \xi_{31}^{\text{UL}} + \xi_{32}^{\text{UL}} + \xi_{33}^{\text{UL}} + \xi_{34}^{\text{UL}} &= 17, \\
 \xi_{11}^{\text{UL}} + \xi_{21}^{\text{UL}} + \xi_{31}^{\text{UL}} &= 10, \quad \xi_{12}^{\text{UL}} + \xi_{22}^{\text{UL}} + \xi_{32}^{\text{UL}} = 8, \quad \xi_{13}^{\text{UL}} + \xi_{23}^{\text{UL}} + \xi_{33}^{\text{UL}} = 14, \\
 \xi_{14}^{\text{UL}} + \xi_{24}^{\text{UL}} + \xi_{34}^{\text{UL}} &= 16, \\
 \xi_{11}^{\text{LL}} + \xi_{12}^{\text{LL}} + \xi_{13}^{\text{LL}} + \xi_{14}^{\text{LL}} &= 10, \quad \xi_{21}^{\text{LL}} + \xi_{22}^{\text{LL}} + \xi_{23}^{\text{LL}} + \xi_{24}^{\text{LL}} = 15, \\
 \xi_{31}^{\text{LL}} + \xi_{32}^{\text{LL}} + \xi_{33}^{\text{LL}} + \xi_{34}^{\text{LL}} &= 15, \\
 \xi_{11}^{\text{LL}} + \xi_{21}^{\text{LL}} + \xi_{31}^{\text{LL}} &= 8, \quad \xi_{12}^{\text{LL}} + \xi_{22}^{\text{LL}} + \xi_{32}^{\text{LL}} = 6, \quad \xi_{13}^{\text{LL}} + \xi_{23}^{\text{LL}} + \xi_{33}^{\text{LL}} = 12, \\
 \xi_{14}^{\text{LL}} + \xi_{24}^{\text{LL}} + \xi_{34}^{\text{LL}} &= 14, \\
 \xi_{gh}^{\text{UL}}, \quad \xi_{gh}^{\text{LL}} \geq 0 \quad \xi_{gh}^{\text{LL}} \leq \xi_{gh}^{\text{UL}}, \quad \xi_{gh}^{\text{UL}} \leq \xi_{gh}^{\text{UU}*}, \quad \xi_{gh}^{\text{LL}} \geq \xi_{gh}^{\text{LU}*}, \quad \forall g = 1, 2, 3; \quad h = 1, 2, 3, 4
 \end{aligned}$$

Solving the problem ULFTP using LINGO 18.0 optimization solver, we have

$$\begin{aligned}\xi_{11}^{\text{UL}} &= 0, \quad \xi_{12}^{\text{UL}} = 1, \quad \xi_{13}^{\text{UL}} = 0, \quad \xi_{14}^{\text{UL}} = 11, \quad \xi_{21}^{\text{UL}} = 7, \quad \xi_{22}^{\text{UL}} = 7, \quad \xi_{23}^{\text{UL}} = 5, \\ \xi_{24}^{\text{UL}} &= 0, \quad \xi_{31}^{\text{UL}} = 3, \quad \xi_{32}^{\text{UL}} = 0, \quad \xi_{33}^{\text{UL}} = 9, \quad \xi_{34}^{\text{UL}} = 5, \\ \xi_{11}^{\text{LL}} &= 0, \quad \xi_{12}^{\text{LL}} = 1, \quad \xi_{13}^{\text{LL}} = 0, \quad \xi_{14}^{\text{LL}} = 9, \quad \xi_{21}^{\text{LL}} = 7, \quad \xi_{22}^{\text{LL}} = 5, \quad \xi_{23}^{\text{LL}} = 3, \quad \xi_{24}^{\text{LL}} = 0, \\ \xi_{31}^{\text{LL}} &= 1, \quad \xi_{32}^{\text{LL}} = 0, \quad \xi_{33}^{\text{LL}} = 9, \quad \xi_{34}^{\text{LL}} = 5.\end{aligned}$$

At this solution, the objective function value is $\Delta^{\text{UL}*} = 1.9007$.

LLFTP

$$\text{Max } \Delta^{\text{LL}} = \frac{7\xi_{11}^{\text{LL}} + 10\xi_{12}^{\text{LL}} + 4\xi_{13}^{\text{LL}} + 7\xi_{14}^{\text{LL}} + 6\xi_{21}^{\text{LL}} + 9\xi_{22}^{\text{LL}} + 11\xi_{23}^{\text{LL}} + 4\xi_{24}^{\text{LL}} + 7\xi_{31}^{\text{LL}} + 4\xi_{32}^{\text{LL}} + 14\xi_{33}^{\text{LL}} + 8\xi_{34}^{\text{LL}}}{14\xi_{11}^{\text{UL}} + 10\xi_{12}^{\text{UL}} + 14\xi_{13}^{\text{UL}} + 6\xi_{14}^{\text{UL}} + 9\xi_{21}^{\text{UL}} + 5\xi_{22}^{\text{UL}} + 12\xi_{23}^{\text{UL}} + 11\xi_{24}^{\text{UL}} + 13\xi_{31}^{\text{UL}} + 15\xi_{32}^{\text{UL}} + 11\xi_{33}^{\text{UL}} + 10\xi_{34}^{\text{UL}}}$$

subject to

$$\begin{aligned}\xi_{11}^{\text{LL}} + \xi_{12}^{\text{LL}} + \xi_{13}^{\text{LL}} + \xi_{14}^{\text{LL}} &= 10, \quad \xi_{21}^{\text{LL}} + \xi_{22}^{\text{LL}} + \xi_{23}^{\text{LL}} + \xi_{24}^{\text{LL}} = 15, \\ \xi_{31}^{\text{LL}} + \xi_{32}^{\text{LL}} + \xi_{33}^{\text{LL}} + \xi_{34}^{\text{LL}} &= 15, \\ \xi_{11}^{\text{LL}} + \xi_{12}^{\text{LL}} + \xi_{31}^{\text{LL}} &= 8, \quad \xi_{12}^{\text{LL}} + \xi_{22}^{\text{LL}} + \xi_{32}^{\text{LL}} = 6, \quad \xi_{13}^{\text{LL}} + \xi_{23}^{\text{LL}} + \xi_{33}^{\text{LL}} = 12, \\ \xi_{14}^{\text{LL}} + \xi_{24}^{\text{LL}} + \xi_{34}^{\text{LL}} &= 14, \\ \xi_{11}^{\text{UL}} + \xi_{12}^{\text{UL}} + \xi_{13}^{\text{UL}} + \xi_{14}^{\text{UL}} &= 12, \quad \xi_{21}^{\text{UL}} + \xi_{22}^{\text{UL}} + \xi_{23}^{\text{UL}} + \xi_{24}^{\text{UL}} = 19, \\ \xi_{31}^{\text{UL}} + \xi_{32}^{\text{UL}} + \xi_{33}^{\text{UL}} + \xi_{34}^{\text{UL}} &= 17, \\ \xi_{11}^{\text{UL}} + \xi_{21}^{\text{UL}} + \xi_{31}^{\text{UL}} &= 10, \quad \xi_{12}^{\text{UL}} + \xi_{22}^{\text{UL}} + \xi_{32}^{\text{UL}} = 8, \quad \xi_{13}^{\text{UL}} + \xi_{23}^{\text{UL}} + \xi_{33}^{\text{UL}} = 14, \\ \xi_{14}^{\text{UL}} + \xi_{24}^{\text{UL}} + \xi_{34}^{\text{UL}} &= 16, \\ \xi_{gh}^{\text{LL}}, \quad \xi_{gh}^{\text{UL}} &\geq 0 \quad \xi_{gh}^{\text{LL}} \leq \xi_{gh}^{\text{UL}*} \quad \forall g = 1, 2, 3; \quad h = 1, 2, 3, 4\end{aligned}$$

Solving the problem LLFTP by substituting $\xi_{gh}^{\text{LL}*}, \xi_{gh}^{\text{UL}*}$ from (ULFTP) and using LINGO 18.0 optimization solver, we have

$$\begin{aligned}\xi_{11}^{\text{LL}} &= 0, \quad \xi_{12}^{\text{LL}} = 1, \quad \xi_{13}^{\text{LL}} = 0, \quad \xi_{14}^{\text{LL}} = 9, \quad \xi_{21}^{\text{LL}} = 7, \quad \xi_{22}^{\text{LL}} = 5, \quad \xi_{23}^{\text{LL}} = 3, \quad \xi_{24}^{\text{LL}} = 0, \\ \xi_{31}^{\text{LL}} &= 1, \quad \xi_{32}^{\text{LL}} = 0, \quad \xi_{33}^{\text{LL}} = 9, \quad \xi_{34}^{\text{LL}} = 5, \\ \xi_{11}^{\text{UL}} &= 0, \quad \xi_{12}^{\text{UL}} = 1, \quad \xi_{13}^{\text{UL}} = 0, \quad \xi_{14}^{\text{UL}} = 11, \quad \xi_{21}^{\text{UL}} = 7, \quad \xi_{22}^{\text{UL}} = 7, \quad \xi_{23}^{\text{UL}} = 5, \quad \xi_{24}^{\text{UL}} = 0, \\ \xi_{31}^{\text{UL}} &= 3, \quad \xi_{32}^{\text{UL}} = 0, \quad \xi_{33}^{\text{UL}} = 9, \quad \xi_{34}^{\text{UL}} = 5.\end{aligned}$$

At this solution, the objective function value is $\Delta^{\text{LL}*} = 0.8673$.

LUFTP

$$\text{Max } \Delta^{\text{LU}} = \frac{6\xi_{11}^{\text{LU}} + 10\xi_{12}^{\text{LU}} + 3\xi_{13}^{\text{LU}} + 7\xi_{14}^{\text{LU}} + 5\xi_{21}^{\text{LU}} + 7\xi_{22}^{\text{LU}} + 8\xi_{23}^{\text{LU}} + 4\xi_{24}^{\text{LU}} + 6\xi_{31}^{\text{LU}} + 3\xi_{32}^{\text{LU}} + 11\xi_{33}^{\text{LU}} + 6\xi_{34}^{\text{LU}}}{14\xi_{11}^{\text{UU}} + 10\xi_{12}^{\text{UU}} + 16\xi_{13}^{\text{UU}} + 8\xi_{14}^{\text{UU}} + 10\xi_{21}^{\text{UU}} + 5\xi_{22}^{\text{UU}} + 15\xi_{23}^{\text{UU}} + 13\xi_{24}^{\text{UU}} + 14\xi_{31}^{\text{UU}} + 16\xi_{32}^{\text{UU}} + 15\xi_{33}^{\text{UU}} + 13\xi_{34}^{\text{UU}}}$$

subject to

$$\xi_{11}^{\text{LU}} + \xi_{12}^{\text{LU}} + \xi_{13}^{\text{LU}} + \xi_{14}^{\text{LU}} = 9, \quad \xi_{21}^{\text{LU}} + \xi_{22}^{\text{LU}} + \xi_{23}^{\text{LU}} + \xi_{24}^{\text{LU}} = 14,$$

$$\xi_{31}^{\text{LU}} + \xi_{32}^{\text{LU}} + \xi_{33}^{\text{LU}} + \xi_{34}^{\text{LU}} = 14,$$

$$\xi_{11}^{\text{LU}} + \xi_{21}^{\text{LU}} + \xi_{31}^{\text{LU}} = 7, \quad \xi_{12}^{\text{LU}} + \xi_{22}^{\text{LU}} + \xi_{32}^{\text{LU}} = 5, \quad \xi_{13}^{\text{LU}} + \xi_{23}^{\text{LU}} + \xi_{33}^{\text{LU}} = 11,$$

$$\xi_{14}^{\text{LU}} + \xi_{24}^{\text{LU}} + \xi_{34}^{\text{LU}} = 14,$$

$$\xi_{11}^{\text{UU}} + \xi_{12}^{\text{UU}} + \xi_{13}^{\text{UU}} + \xi_{14}^{\text{UU}} = 13, \quad \xi_{21}^{\text{UU}} + \xi_{22}^{\text{UU}} + \xi_{23}^{\text{UU}} + \xi_{24}^{\text{UU}} = 20,$$

$$\xi_{31}^{\text{UU}} + \xi_{32}^{\text{UU}} + \xi_{33}^{\text{UU}} + \xi_{34}^{\text{UU}} = 18,$$

$$\xi_{11}^{\text{UU}} + \xi_{21}^{\text{UU}} + \xi_{31}^{\text{UU}} = 11, \quad \xi_{12}^{\text{UU}} + \xi_{22}^{\text{UU}} + \xi_{32}^{\text{UU}} = 9, \quad \xi_{13}^{\text{UU}} + \xi_{23}^{\text{UU}} + \xi_{33}^{\text{UU}} = 15,$$

$$\xi_{14}^{\text{UU}} + \xi_{24}^{\text{UU}} + \xi_{34}^{\text{UU}} = 16,$$

$$\xi_{gh}^{\text{LU}}, \quad \xi_{gh}^{\text{UU}} \geq 0 \quad \xi_{gh}^{\text{LU}} \leq \xi_{gh}^{\text{LL*}} \quad \forall i = 1, 2, 3; \quad j = 1, 2, 3, 4$$

Solving the problem LUFTP by substituting $\xi_{gh}^{\text{LU*}}, \xi_{gh}^{\text{UU*}}$ from (UUFTP) and using LINGO 18.0 optimization solver, we have

$$\xi_{11}^{\text{LU}} = 0, \quad \xi_{12}^{\text{LU}} = 0, \quad \xi_{13}^{\text{LU}} = 0, \quad \xi_{14}^{\text{LU}} = 9, \quad \xi_{21}^{\text{LU}} = 7, \quad \xi_{22}^{\text{LU}} = 5, \quad \xi_{23}^{\text{LU}} = 2, \quad \xi_{24}^{\text{LU}} = 0,$$

$$\xi_{31}^{\text{LU}} = 0, \quad \xi_{32}^{\text{LU}} = 0, \quad \xi_{33}^{\text{LU}} = 9, \quad \xi_{34}^{\text{LU}} = 5,$$

$$\xi_{11}^{\text{UU}} = 0, \quad \xi_{12}^{\text{UU}} = 2, \quad \xi_{13}^{\text{UU}} = 0, \quad \xi_{14}^{\text{UU}} = 11, \quad \xi_{21}^{\text{UU}} = 7, \quad \xi_{22}^{\text{UU}} = 7, \quad \xi_{23}^{\text{UU}} = 6,$$

$$\xi_{24}^{\text{UU}} = 0, \quad \xi_{31}^{\text{UU}} = 4, \quad \xi_{32}^{\text{UU}} = 0, \quad \xi_{33}^{\text{UU}} = 9, \quad \xi_{34}^{\text{UU}} = 5.$$

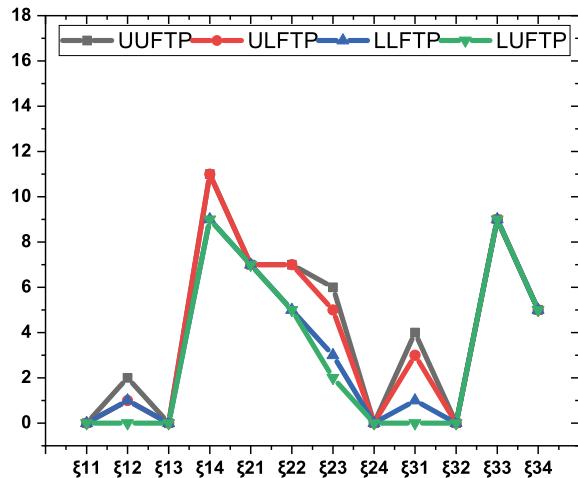
At this solution, the objective function value is $\Delta^{\text{LU*}} = 0.4973$.

After solving the above four crisp FTPs, the optimal transported amount is obtained in terms of rough intervals which are shown in Table 2 and Fig. 2. Also, at this solution, the rough interval optimum ratio of total profit to total transportation cost is obtained as [0.8673, 1.9007][0.4973, 3.1969].

Table 2 Rough optimal transported amount

Destinations → Sources ↓	1	2	3	4
1	[0, 0][0, 0]	[1, 1][0, 2]	[0, 0][0, 0]	[9, 11][9, 11]
2	[7, 7][7, 7]	[5, 7][5, 7]	[3, 5][2, 6]	[0, 0][0, 0]
3	[1, 3][0, 4]	[0, 0][0, 0]	[9, 9][9, 9]	[5, 5][5, 5]

Fig. 2 Graphical representation of the rough optimal transported amount



6 Conclusions

Rough set theory is practically more applicable to express uncertainty in optimization problems, since it takes into consideration the opinion of all the involved experts (intersection) and respects their knowledge (union) by lower and upper approximation intervals, respectively. To analyze this advantage of the rough set theory, a new solution methodology is proposed to solve the fully rough fractional transportation problem. The systematic process of the proposed approach operates by decomposing the problem into two sub-models as the upper interval model and the lower interval model. From these two sub-models, four crisp fractional transportation problems such as UUFTP, ULFTP, LLFTP, and LUFTP are constructed. Finally, solving these four crisp fractional transportation problems using LINGO 18.0 optimization solver gives the rough interval optimal solution of the proposed model.

The proposed work is innovative in the sense that the fractional transportation problem has been formulated in terms of the rough set theory, and the solution approach has been developed for the proposed fully rough fractional transportation problem. Unlike existing approaches that provide crisp solutions to a rough problem, the suggested algorithms provide a rough interval solution to a problem in the rough environment, which is a significant advantage. This technique thus broadens the range of obtained solutions and handles uncertainty and inaccuracy more effectively. The extension of the proposed methodology for the unbalanced fully rough fractional transportation problem will be an interesting future research direction. Also, to examine the benefits of the rough set theory, it can be applied in a wide variety of fields, for example, scheduling, sequencing, supply chain management, etc.

Acknowledgements The first author is thankful to the Ministry of Human Resource Development, India, for providing financial support, to carry out this work.

Compliance with Ethical Standards

Conflict of Interest All the authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Hitchcock FL (1941) The distribution of a product from several sources to numerous localities. *J Math Phys* 20(1):224–230. <https://doi.org/10.1002/sapm1941201224>
2. Ahmed MM, Khan AR, Uddin MS, Ahmed F (2016) A new approach to solve transportation problems. *Open J Optim* 5(1):22–30. <https://doi.org/10.4236/ojop.2016.51003>
3. Amaliah B, Fatichah C, Suryani E (2022) A new heuristic method of finding the initial basic feasible solution to solve the transportation problem. *J King Saud Univ Comput Inf Sci* 34(5):2298–2307. <https://doi.org/10.1016/j.jksuci.2020.07.007>
4. Amaliah B, Fatichah C, Suryani E (2022) A supply selection method for better feasible solution of balanced transportation problem. *Expert Syst Appl* 203:117399. <https://doi.org/10.1016/j.eswa.2022.117399>
5. Karagul K, Sahin Y (2020) A novel approximation method to obtain initial basic feasible solution of transportation problem. *J King Saud Univ Eng Sci* 32(3):211–218. <https://doi.org/10.1016/j.jksues.2019.03.003>
6. Swarup K (1966) Transportation technique in linear fractional functional programming. *J R Naval Sci Serv* 21(5):256–260
7. Gupta A, Khanna S, Puri M (1993) A paradox in linear fractional transportation problems with mixed constraints. *Optimization* 27(4):375–387. <https://doi.org/10.1080/02331939308843896>
8. Khurana A, Arora S (2006) The sum of a linear and a linear fractional transportation problem with restricted and enhanced flow. *J Interdiscip Math* 9(2):373–383. <https://doi.org/10.1080/09720502.2006.10700450>
9. Joshi VD, Gupta N (2011) Linear fractional transportation problem with varying demand and supply. *Le Matematiche* 66(2):3–12. <https://doi.org/10.4418/2011.66.2.1>
10. Zadeh L (1965) Fuzzy sets. *Inf Control* 8(3):338–353
11. Anukokila P, Radhakrishnan B (2019) Goal programming approach to fully fuzzy fractional transportation problem. *J Taibah Univ Sci* 13(1):864–874. <https://doi.org/10.1080/16583655.2019.1651520>
12. Bhatia TK, Kumar A, Sharma MK (2022) Mehar approach to solve fuzzy linear fractional transportation problems. *Soft Comput* 26:11525–11551. <https://doi.org/10.1007/s00500-022-07408-x>
13. Khalifa HAEW, Kumar P, Alharbi MG (2021) On characterizing solution for multi-objective fractional two-stage solid transportation problem under fuzzy environment. *J Intell Syst* 30(1):620–635. <https://doi.org/10.1515/jisys-2020-0095>
14. Liu ST (2016) Fractional transportation problem with fuzzy parameters. *Soft Comput* 20(9):3629–3636. <https://doi.org/10.1007/s00500-015-1722-5>
15. Atanassov K (1986) Intuitionistic fuzzy sets. *Fuzzy Sets Syst* 20(1):87–96
16. Anukokila P, Anju A, Radhakrishnan B (2019) Optimality of intuitionistic fuzzy fractional transportation problem of type-2. *Arab J Basic Appl Sci* 26(1):519–530. <https://doi.org/10.1080/25765299.2019.1691895>
17. Bharati SK (2019) Trapezoidal intuitionistic fuzzy fractional transportation problem. In: Soft computing for problem solving, pp 833–842. <https://doi.org/10.1007/978-981-13-1595-4-66>
18. El Sayed M, Abo-Sinna MA (2021) A novel approach for fully intuitionistic fuzzy multi-objective fractional transportation problem. *Alex Eng J* 60(1):1447–1463. <https://doi.org/10.1016/j.aej.2020.10.063>

19. El Sayed MA, El-Shorbagy MA, Farahat FA, Fareed AF, Elsisy MA (2021) Stability of parametric intuitionistic fuzzy multi-objective fractional transportation problem. *Fract Fract* 5(4):233–250. <https://doi.org/10.3390/fractfract5040233>
20. Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11(5):341–356
21. Rebolledo M (2006) Rough intervals enhancing intervals for qualitative modeling of technical systems. *Artif Intell* 170(8):667–685. <https://doi.org/10.1016/j.artint.2006.02.004>
22. Bouzayane S, Saad I (2020) A multi-criteria approach based on rough set theory for the incremental periodic prediction. *Eur J Oper Res* 286(1):282–298. <https://doi.org/10.1016/j.ejor.2020.03.024>
23. Ruidas S, Seikh MR, Nayak PK (2021) A production-repairing inventory model considering demand and the proportion of defective items as rough intervals. *Oper Res* 22:2803–2829. <https://doi.org/10.1007/s12351-021-00634-5>
24. Seikh MR, Dutta S, Li DF (2021) Solution of matrix games with rough interval pay-offs and its application in the telecom market share problem. *Int J Intell Syst* 36(10):6066–6100. <https://doi.org/10.1002/int.22542>
25. Sharma HK, Kumari K, Kar S (2020) A rough set theory application in forecasting models. *Decis Mak Appl Manag Eng* 3(2):1–21. <https://doi.org/10.31181/dmame2003001s>
26. Stankovic M, Gladovic P, Popovic V (2019) Determining the importance of the criteria of traffic accessibility using fuzzy AHP and rough AHP method. *Decis Mak Appl Manag Eng* 2(1):86–104. <https://doi.org/10.31181/dmame1901086s>
27. Bera S, Giri PK, Jana DK, Basu K, Maiti M (2018) Multi-item 4D-TPs under budget constraint using rough interval. *Appl Soft Comput* 71:364–385. <https://doi.org/10.1016/j.asoc.2018.06.037>
28. Midya S, Roy SK (2020) Multi-objective fixed-charge transportation problem using rough programming. *Int J Oper Res* 37(3):377–395
29. Bera RK, Mondal SK (2020) Credit linked two-stage multi-objective transportation problem in rough and bi-rough environments. *Soft Comput* 24(23):18129–18154. <https://doi.org/10.1007/s00500-020-05066-5>
30. Garg H, Rizk-Allah RM (2021) A novel approach for solving rough multi-objective transportation problem: development and prospects. *Comput Appl Math* 40(4):1–24. <https://doi.org/10.1007/s40314-021-01507-5>
31. Shivani, Rani D, Ebrahimnejad A (2022) An approach to solve an unbalanced fully rough multi-objective fixed-charge transportation problem. *Comput Appl Math* 41(4):1–27. <https://doi.org/10.1007/s40314-022-01830-5>
32. Xu J, Tao Z. Rough multiple objective decision making. CRC Press Taylor and Francis Group. <http://www.copyright.com/>

Applications of IoT and Various Attacks on IoT



Sumeet Dhillon, Nishchol Mishra, and Devendra Kumar Shakya

Abstract Internet of Things (IoT) is the unification of digital/technical world with the real world and its applications, enabling the ease of communication between people, devices, processes, and objects present in the surroundings. IoT is the newest technological modernization with the ability to boost efficiency, enhance environmental sustainability, and upgrade security covering numerous evolving domains as well as assist in smoothing the everyday task and activities. The IoT researches accomplished till date certify the enormous progress in the field of Internet of Things. IoT applications are progressively fabricating their existence in nearly every sector surrounding humans. Few of the omnipresent applications of IoT are smart homes and city, smart health care, hi-tech wearable's, and smart agriculture which are further elaborated in the survey. This survey paper presents an exhaustive description of applications of IoT, with the possible vulnerabilities and attacks on numerous IoT applications. Thus, the study also includes the mitigation methods to prevent these attacks and attain trustworthy and reliable IoT applications.

Keywords IoT · Attacks · Security

1 Introduction

IoT refers to a platform consisting of embedded sensors/devices linking and communicating via Internet to accumulate and interchange information within the network. IoT technology and applications are flourishing and enhancing human lives and activities

S. Dhillon (✉)

Computer Science and Engineering Department, SATI, Vidisha, India
e-mail: sumeetdhillon.cse@satiengg.in

N. Mishra

School of Information Technology, RGPV Bhopal, Bhopal, India
e-mail: Nishchol@rgtu.net

D. K. Shakya

Department of Electronics Engineering, SATI, Vidisha, India
e-mail: dkshakya.ec@satiengg.in

in the field of education, health care, transportation, and living accommodations. With the frequent increment in IoT connected devices, numerous challenging obstacles are also arising with IoT technology including scalability, security, heterogeneity, service quality, etc. IoT is a structure with embedded objects and sensor technology to connect with other devices in the network using wireless medium for transmission and communication of data to initiate, interchange, and transmit information without any human interactivity. Internet of Things allows inter-linkage between the cyber world and physical world referred to as cyber physical system. Moreover, IoT can be accomplished as smart homes, smart management, smart city, and smart transportation and in many other developing fields.

Internet of Things can also be stated as a flourishing network of software as well as hardware modules associated with diurnal living activities to maintain comfort, convenience, and reliability. This trending technology is expanding and introducing numerous IoT gadgets in the market, consisting of heterogeneous and myriad devices with divergent protocols. Due to their potentialities like data storage, processing, and networking, IoT devices assist users in leading a comfortable lifestyle. IoT can convert homes smarter and improve security by boosting convenience, energy efficiency, enhancing security, and relaxation in day-to-day human activities. Thus, it is helpful in enriching the data analytics, automation, and healthcare sector.

Smart wearables, smart home and city, advance health care, smart-grids, and smart transportation are some of the voguish IoT applications. In accordance with the research analysis, it has been corroborated that any system manifesting the properties of dynamic topology, remote management, resource restraints, and wireless means of communication are invariably vulnerable to security concern. Thus, the Internet of Things (IoT) applications require escalating the data accumulation from IoT devices and sensors to transmit these data collection to IoT server and applications where the data is analyzed, clustered, and cumulated for the purpose of end users. Hence, it is eminently crucial to secure, protect, and ensure the data from not reaching the malicious threats and adversary trap (Fig. 1).

The paper is structured into five sections where Sect. 1 encapsulates the introduction to Internet of Things and its applications. Section 2 is literature survey of recent research papers on IoT and its applications. Sections 3 and 3.1 provides a description of IoT applications with attack and vulnerabilities to IoT application (Table 1) and types of attacks with its definition and countermeasures associated to it (Table 2), respectively. Sections 4 and 5 covers the conclusion and future scope concerning IoT application and its security.

2 Literature Review

The IoT industries are attempting to bring about more projects allied to IoT in the developing sectors of health care, transportation, aquaculture, farming, and many more. The statistical data indicates the profitable growth of approx. 24% for the IoT industries. The records perceived from recent years information not only includes

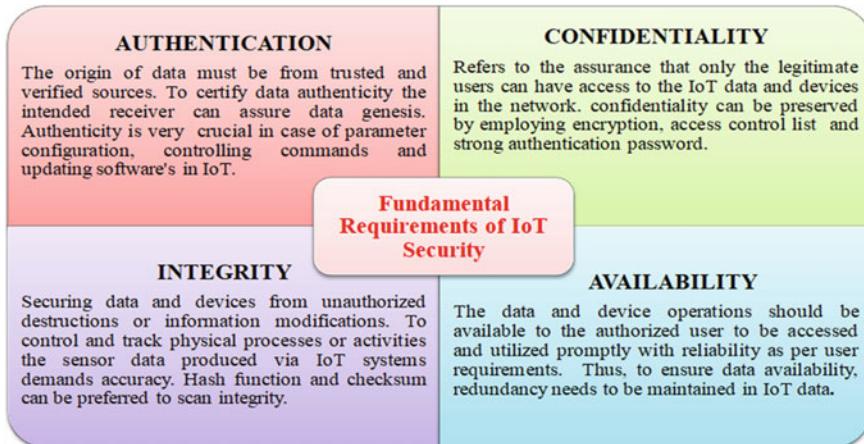


Fig. 1 Fundamental requirements of IoT security

the widening of IoT industries but also focuses on the increasing number of IoT publications and research articles.

Encapsulating paper [14] mentioning the modern arising IoT security attacks and threats including their growing vulnerabilities and possible countermeasures to secure network, software, and hardware infrastructure. To prevent the attack and attacker from blemishing the IoT network and resources, it is crucial to secure IoT by employing cyber security and other security mechanism. The initial goal of IoT systems and devices is to guarantee data integrity, confidentiality, security, and availability. The paper [14] covers IoT challenges like storage issue, physical tampering, insecure web interface, not updating software and firmware, and hardcoded credentials. It also discusses major IoT attacks such as brute force attack, Sybil attack, DDoS attack, replay attack, MITM attack, malicious application, and weak authentication. It should be noted that the security measures taken to protect the devices must not impact the device performance or system usability instead applying lightweight protocols to escalate the system performance. Availing of intrusion detection/prevention system to secure from network attack like malware infection, brute force, and DDoS. Further, research can be continued to implement complex security algorithm and enhance performance.

The survey paper [12] explores numerous security attacks like eavesdropping, booting attack, sleep deprivation, malicious code injection, distributed denial-of-service, routing attack, MITM, sniffing attack, etc., on various layers of IoT applications including the emerging open issues targeting sensing layer, middleware layer, application and network layer, and gateways. Thus, covers the existing and forthcoming solutions such as block-chain, machine learning, edge and fog computing to safeguard from IoT threats.

Numerous IoT devices make use of cloud for data storage and synchronization. To prevent data breaches and other threats in the route to cloud, it is crucial to

Table 1 Applications of IoT, its services, attacks, and vulnerability

References	Application	Description	Attacks	Vulnerabilities
1 [1–3]	Smart home	A huge number of IoT applications have empowered the automation of different household activities. Adopting numerous embedded devices for saving energy, smart metering, and to manage home security	Eavesdropping, DDoS, man-in-the-middle (MITM) attack, impersonation, and malware attack	(1) Shortage of effectual cryptographic support (2) Web-based interface reliability (3) Limited authentication, authorization, and accounting (AAA) services
2 [4–6]	Smart health care	IoT has introduced various opportunities and benefits in the area of health care. It brought development and improvement in the field of health services. This includes hospital and medical management, drug control, and intelligent machines	Botnet attack, DoS denial-of-service, identity theft, MITM, black hole attack, and routing diversion	(1) Insecure device and mobile connectivity (2) Uncertain cloud connectivity (3) Bounded privacy and limited availability (4) Lack of web-based interfaces security
3 [7, 8]	Smart city	A middleware can be provided through IoT for future smart cities. Gathering data from IoT devices, sensing infrastructure, and placing it in a consistent manner. Thus, giving rise to smart services that can communicate with surrounding scenario like smart streetlight, smart-grids, smart-transport, etc.	Eavesdropping, malware, black hole attack, hardware or software failure, data theft, and DoS attack	(1) Lack of privacy (2) Restricted cloud connectivity (3) Lack of effectual cryptographic support (4) Insecure device linking

(continued)

Table 1 (continued)

References	Application	Description	Attacks	Vulnerabilities
4 [9]	Smart wearable	Wearables are the electronic devices that are worn by people in order to analyze, track, and transmit personal information. These smart devices can even track biometric data from human heart rate to sleeping patterns. This technology is gaining popularity in the fashion and gaming industries	Brute force attack, phishing attack, draining of battery, DDoS attack, and sniffing	(1) Data security and privacy (2) Insecure data communication (3) Lack of encryption and authentication (4) Physical damage to device (5) Tracking of location and data leakage
5 [10, 11]	Smart farming	A smart system designed to monitor crop field using sensors and observing crop health, temperature, humidity, light, soil moisture, etc. Moreover, making irrigation system automatic. Thus, enabling farmers to enhance their productivity. The farmers can supervise their land from anywhere	Sybil attack, distributed DDoS denial-of-service attack, and spoofing attack	(1) Data privacy and holding are vital security concern within the agriculture sector (2) Integrity and availability of data (3) Harsh environmental conditions can affect the devices (4) Sensors are vulnerable to malfunction

Table 2 Attacks on IoT applications and their countermeasures

References	Attacks	Illustration	Countermeasures to attacks
1 [12, 13, 6]	Eavesdropping attack	These attack take place when an intruder interrupt, alter, steal, or eradicate data which is being transmitted and communicated between two or more connected devices, also named as snooping or sniffing attack. By exploiting insecure network and device communications, the hacker gains access to user confidential information when transmitted between accredited sender and receiver	Data encryption and authentication, segmenting and monitoring network by employing endpoint and intrusion detection system, exploiting security technologies like VPN's, packet filtering, antimalware's, firewalls
2 [14, 6, 7, 11]	Sybil attack	A solo malicious node can execute the attack by creating numerous dupe identities and broadcasting data to infect the comprehensive network performance. The intruders make use of fake devices to form congestion in the IoT network and decrement the device performance, thus are hazardous as these strikes on availability and functioning of network as well as devices	Verify the exact location, ensure message proceedings and authentication, certifying trust and privilege attenuation, maintain received signal strength indication (RSSI), testing and validating resources
3 [14, 12, 15, 10, 16]	DDoS attack	This sort of attack probably takes place by compromising a huge range of computing and communicating devices with Internet access, thus a botnet. Distributed denial-of-service arises when several systems overwhelm the bandwidth or resources of the intended user. Thus, making the services unavailable to the user by overloading the devices and network with enormous traffic. Moreover, exhausting the device performance and network resources via flooding data	Bandwidth increment to control heavy traffic, fortify effective network security, maintain server redundancy, restrict network broadcasting among the devices, continuous monitoring to inspect the traffic in real time

(continued)

Table 2 (continued)

References	Attacks	Illustration	Countermeasures to attacks
4 [17]	Key-logger	It is a kind of surveillance technique or a spyware tool preferred by cyber-attackers to monitor, track, and record every keystroke on targeted computer/system in order to purloin or collect sensitive identities, login passwords, etc., by installing a spying software or hardware device to capture the key-logging patterns	Avail two-step-verification, install key encryption software to avert hardware key-logger attack, host intrusion prevention system to monitor keyboard/disk/memory, auto containment, and website filtering to prevent system from malicious software and websites
5 [14, 51]	Brute force attack	It is a hacking strategy where the attacker uses hit and try technique to gain authentication/login details, guess, and crack password or rift the encryption keys of the smart devices. The hacker attempts several combinations via computer to make out username and password of the victim. This is an old, easy still a reliable trick for attaining illegitimate access to the targeted network or individual credential as well organization system	Never use personal credentials as password. Do prefer unique combinations, restricting number of login attempts on sites and devices, validate user identity using multi-factor authentication, using CAPTCHAS, and unique URL's for online login
6 [18]	Root-kit attack	Referred to as hidden malicious software's consisting of huge number of tools used for gaining root access of the appropriated operating system permitting to execute the commands with admin prerogatives and providing entrance to root-kit initiator. Thus, provide attacker the ability to disable or corrupt security software, track user and device functioning to hijack the system	Never download attachments send via unknown contacts, stay alert from phishing/scam e-mails, stop ignoring system/software's updates, be cautious from drive-by downloads to prevent automatic root-kit download

(continued)

Table 2 (continued)

References	Attacks	Illustration	Countermeasures to attacks
7 [6, 19, 20]	Botnet attack	A group of hijacked devices is termed as botnet. IoT botnets are a network of IoT devices linking to the malware controlled infected routers. These are misused in instigating DDoS attack on the target unit to distort its performance and functioning services. Botnets are administrated via single command and control server (C- & C) linking every infected bots in the network	Installing antimalware/antivirus programs, maintain up-to-date operating system, do not open/download files from suspicious source, create robust password, and employing (2FA) two factor authentication
8 [13, 10, 17]	Mirai botnet	Introduced in 2016 infecting upmost services and websites on Internet. Mirai is a sort of malicious software targeting end user devices including smart home, router, cameras and other artifacts, converting them in remote restrained bots of zombie network. Mirai botnets are exploited by the cyber-attackers to perform massive DDoS attack onto victim's computer system	Changing the default system administrator credentials constantly, preserve antivirus tools/anti-software's and their updates, incessantly updating the firmware and system software
9 [17]	Reaper botnet	Referred to as "IoT troop" which has infected approx. one million corporation in year 2017. Reaper escalates like computer worms and spread impulsively between infected devices. It employs a fusion of nine attacks to target familiar IoT susceptibility including server network, storage device, and IP camera. Thus, it has the capabilities to perform complex attacks	Track and reinforce vulnerable tools and software's, strengthen the password harder to crack, robust communication, enable encryption and ensure authentication, secure via SSH (secure shell) and HTTPS protocol, incapacitate redundant services on system
10 [7, 21, 22]	Black hole attack	When an infected malicious node claims to provide shortest path to packets in the network till its destination, where as it drops/discards all the received packets from being redirected to the adjacent node is known as black hole attack. The attack turns more challenging and severe when merged with sinkhole as it cease complete data traffic encompassing black hole	Exploiting multiple base stations to enhance the possibilities of successful packet transmission in the network, enabling structured security and effective cryptographic algorithm to ensure safe and reliable communication

secure IoT applications via end to end security and trust. The author in paper [13] presents a 3D model for threat identification consisting of three dimensions as: The targeted security services responsible for preserving data integrity, confidentiality and availability, second is attack venue incorporating three route as communication network, IoT devices, and cloud for data storage plus processing, third is planes of attack comprising four layers named as firmware, application software, hardware, and operating system. It also defines IoT application architecture covering four domains namely: IoT devices, communication, the cloud, and presentation. The paper [13] also examined nine security challenges including inadequate integrity and trust, software vulnerabilities, lack of standardization, insecure web interface, open architecture, system limitations, weak security link, privacy issue, and malware targeting IoT device. Furthermore, mitigation methods to security risk are discussed as fog computing, location tracking, device identification, and security policies.

IoT is being used to assist humans and their activities via numerous applications such as smart cities, health care, and smart accounting, some of the major issues faced by IoT applications are reliability, scalability, security, consistency, and privacy. Researchers have suggested numerous solutions to secure IoT, where block-chain is one of the latterly solution to overpower IoT obstacles. The aim of paper [15] is to lay out a survey on utilization of block-chain in IoT framework. It constitutes the attributes of block-chain technology such as decentralization, immutability, transparency, and security for exploiting IoT applications. Thus, the author explores the broadly preferred and evaluative block-chain platform for IoT applications like Ethereum platforms and Hyperledger fabric. Hence, the survey is helpful in identification of complications related to block-chain-based integration and design for applications of Internet of Things.

Cloud-based technology enables analysis, accessing, and controlling of precise information to serve smart policies to people, businesses, and organizations in order to upgrade human's lifestyle. IoT assist smart cities in escalating resource and transportation management, reducing pollutants, decrease accident rate, garbage collection management, etc. People connect in smart city environment via android devices connecting smart homes and smart vehicles. Connecting devices and data to cities physical system improves efficiency and reduces expenditures. The paper [23] analyzes and explores the role of cloud-IoT applications in smart cities. Moreover, covering cloud-IoT convergence and associated services are infrastructure as a service (IaaS), software as a service (SaaS), platform as a service (PaaS), cloud-based IoT services/solutions such as Thing Speak (holding precious IoT data and securing in public/private cloud), Xively (supports interactive and secure records), Plotly (open source API for data analysis and identification tools), Yaler (provides cloud connectivity services), and applications like ABB Robotics, Amazon Warehouse, Airbus, Real Time Innovation (RTI), and many more for smart city.

The author in paper [24] set forth a comprehensive review of arising IoT technologies for smart farming. Firstly, the author summarizes relevant studies and explores emerging IoT technologies for smart agriculture like open source IoT platforms, middleware platforms, wireless technology, network function virtualization (NFV), software defined networking (SDN), and fog/cloud computing. It contributes IoT

applications categorization for smart farming into seven classes namely smart water and smart harvesting management, supply chain management, farm disease controlling, device monitoring, agro-chemical utilization, and smart farming practices. The paper [24] also presents taxonomies and comparability of state-of-the-art methodology regarding block-chain technology-based supply chain management for IoT smart agriculture. Moreover, real-time projects using existing technologies are added to indicate high performance in the area of smart farming.

The aim of paper [1] is to furnish a baseline survey regarding Internet of Things, machine learning, smart cities, smart health care, and their relationship. Further, it elaborates the usage of IoT, block-chain, artificial intelligence, and machine learning technologies for the advancement of smart health care and smart city system to optimize and enrich the accuracy/correctness of expected outcomes. IoT, sensor network, and machine learning provide potentiality to better recognize diseases and relieve doctors. The major challenges faced by healthcare/medical system are due to increasing digital and analog data, financial/economical hindrance, and aging society. Thus, giving rise to robots utilization in day-to-day healthcare scenario, assigning IoT enabled objects and sensors to cure health issue and advanced city environment can ease and enhance the quality/standard of human lifestyle. Believing the continuous growth in sensor technology combining with IoT, machine learning, artificial intelligence, the upcoming future of healthcare facilities with modern opportunities is highly auspicious for doctors, hospital staff, patients, and other medical tool manufacturers.

In paper [2], the author firstly examines the smart home design methodologies and then suggested a comprehensive architecture for further progress of smart homes. The recommended smart home structure contains device group, central controller, monitor group, and user interfaces. The representative in device and monitor group constantly reports the progress to central controller. To attain privacy and security, the author recommended a lightweight security preserving communication protocol supporting message authentication code (MAC) to maintain data authenticity, integrity, and chaos-based encryption. Accounting to the employed agents restricted computing abilities in smart home to encode the data transmission, the author espouses symmetric cryptographic scheme. For the purpose of MAC computation and encryption, one time secret keys are applied initiated on the basis of divergent chaotic approach. As per resultant, the proposed strategy attains great efficiency and higher security status by simplifying key management and minimizing memory capacity via proposed chaos-based cryptographic method.

IoT wearables put forward infinite advance possibilities in several real-life applications, with huge potential when merged with accessible IoT technology. The author in paper [3] comprehensively explored numerous latest and valuable studies in the field of IoT wearable's and categorized IoT wearables in four notable sections according to their significance on the basis of: health, localization and tracking, sports and diurnal activities, security, and safety. Cellular Internet of Things (CIoT) can reform wearable IoT industry where CIoT still needs to gain recognition in research society. IoT smart wearables accumulate and explore data, also take acute decisions and deliver outcome to end users, thus discovering applications progressively to

upgrade human's day-to-day life. The paper [3] also encloses essential divergence of algorithms related to wearable's cluster and inspected the open challenges/future direction such as privacy, power consumption, sensor resolution, safety, and wearing comfort-ability in above mentioned four groups. The study discloses several CIoT benefits and elevation to IoT wearable gadgets. Moreover, directing the possibilities and obstacles while administrating CIoT enable wearables.

The research paper [4] mentions a consummated suggestion for illustrating IoT security solutions to secure different aspects of IoT including detailed description of IoT security attacks, threats, and susceptibilities to ensure the service level performance of complete information on the basis of IoT technology. Moreover, it spotlights the major challenges and solutions to prevent data loss in IoT systems and secure data from various attacks that are required to be observed at organizational and system level. The author recommended the IoT devices security management system to enable the IT association assembling, break-down, account security data and instance to differentiate, analyze, observe, and review IT concerning threats which can lead to the association operative jeopardize, thus operated IoT security approach with the high distinction model.

The aim of researcher in paper [5] is to examine and determine the variety of IoT applications, scan the possible security threats, and their consequences on IoT applications. A detailed survey of previous studies to better understand IoT applications and allied cyber susceptibilities is mentioned. The paper [5] encloses IoT applications namely smart home, smart health, smart city, smart transportation and traffic management, smart wearable devices along with their vulnerabilities such as limited device memory, insecure web interface, network services, compromised device to device and cloud connectivity, non-updated software, thus giving rise to IoT attack like brute force, denial-of-service, injection attack, man-in-the-middle attack, data loss and corruption, and weak authentication. This research did not include and suggested the possible solutions as well as preventive measures to secure with attacks and vulnerabilities. Hence, in the future, the author will concentrate on implementing the countermeasures to the above-mentioned threats.

The rapid evolution in IoT applications in human lifestyle by availing multiple communication medium and devices, protecting data, and devices has turned to be more complex job. The paper [6] has reviewed the complications and security weakness such as take control via active or passive attack and information theft are faced in the development of IoT environment and its applications. The researcher encloses the requirements of IoT applications as reliability, scalability, availability, data management, efficient energy, openness, analytical tools, security, device sensing, and intelligence for secure communication. In addition, the IoT challenges concerning hardware, software, and network such as mobility, device multiplicity, scalability, and multiplicity of communication channel are discussed, the categorization of attacks namely eavesdropping attacks, malware, DoS, impersonation, MITM, botnet attacks, identity fraud, replay attack, routing diversion, spoofing, and Sybil attacks in IoT applications like smart farming, smart medical facilities, smart car/transportation, and smart campus are encapsulated.

Concatenation of cloud computing with IoT has escalated the broad diversity of IoT applications in numerous fields adding commercial sector, engineering department, manufacturing field, as well as supply chain sector by furnishing higher processing potentialities, adaptability, high availability, and affordability services. Still, there are few security challenges that need to be addressed are data access control and management, hardware protection, denial-of-service, and power constraints. The paper [7] introduces the framework regarding cloud assisting IoT and its applications in promoting smart cities, transportation, and telemedicine's. Auditing the risk of security attack and threats because of illegitimate access and misemployed information accumulated via IoT nodes.

The author targets to survey modern arising vulnerabilities and security obstacles in IoT applications, including the proposed solutions to mitigate the possible attacks in IoT applications. In smart city sector, the feasible threats are eavesdropping, hardware or software failure, DoS attack, lack of testing, manufacturer defects, and natural hazards. The methods to alleviate security risk can be advance proactive antivirus defense and intelligent threat detection. Moreover, protect telemedicine's and health care by maintaining the integrity, availability, confidentiality, and authenticity of data and devices. Inspecting intelligent transportation framework where the intruder can physically attain access via exposed port and can gain login details using brute force attack. Thus, the major IoT security issues that need to be considered are admission control and system confirmation, lack of authorization/authentication/privacy/security frame, and requisite storage administration. Hence a dynamic defence-based method for securing IoT and real-life application from harmful attack should be used.

The paper [25] analyzes how the IoT has revolutionized almost every area of our life. The author also discussed about the application of IoT in hospitals like with the installation of sensors, the beds in hospital are operated and the use of IoT apps for monitoring the health of patient. Thus, any vulnerability in IoT-based health apps can have disastrous impact on the patient life. Health apps are prone to DDoS attacks and sensor attacks. There is a need to develop system to monitor and detect attacks so that these attacks can be prevented to ensure security.

The paper [26] introduces the role of agriculture in GDP of India and also mentioned the problems faced by Farmers to meet the demand chain resulting in inflation. These problem are due to lack of any developed system in place which can predict the weather conditions and soil condition. The author proposed a IoT-based system which can predict the weather and soil condition in advance, which will alert the farmer on time to take preventive measures and meet the demand-supply chain. There is a possibility of side channel attacks on this system. Hence, developing a privacy preserving method to prevent attack is required.

3 IoT Applications

IoT is playing a versatile role in various aspects of public and private sectors in today's scenario. Because of IoT, people can maintain home and appliance security, track devices or things, monitor health, make reservation, etc., easily. Yet, there are countless probabilities for Internet of Things to interconnect autonomously with each and every object/device accessible in cyberspace. Thus, adopting IoT for betterment and upgrading of organizations, businesses, and government division is a suitable choice. Moreover, Table 1 lists out the various application of Internet of Things with its utility, vulnerability, and types of possible attacks on each application.

Need of IoT applications: To explore the veracious potentiality of IoT technology, IoT applications exploits businesses, enhance and ease human life, automate device, handle processes, performance, serving appropriate information, monitoring, and controlling environmental set up remotely. IoT applications assist in building business models and other stream by providing real-time data required to expand services and smart products.

3.1 Attacks on IoT Applications

These days, the technology is becoming to be an essential part of human life. Internet of Things can be interpreted as a miscellaneous technology applicable in numerous application domain and environments at system level. IoT systems are general purpose, network connected devices and nodes that can be attacked/hacked by intruders using malware's and other harmful attacks giving rise to complications in IoT security. Thus, the security of IoT is one of the major concern in the trending and future applications of IoT.

By exploiting IoT, the physical gadgets or objects can be accredited to generate, collect, and interchange data and information in a seamless way. The connected IoT network and devices turn out to be a challenging and complex system when it becomes vulnerable to hazardous attacks and thus makes it difficult to control and handle its impact on IoT objects. According to recent observations, there are varieties of intricate attacks targeting IoT as a tool, to block the services and make them useless for the authorized users.

In this section, the prevalent attacks on the IoT applications including description and proposed solutions are mentioned. The recommended solutions may have restrictions, however, the most relevant and satisfactory approach, preventive measure can be preferred to deal with these complex threats and thus, securing IoT environment.

4 Conclusion

Several applications of IoT prioritize on automating numeral task and activities, therefore, are making effort to authorize the inanimate physical devices to function and react successfully without any human interference. The major trouble arises during the identification of affected devices within that time period, the accessed services as well as resources and most importantly the impacted areas of the system, in order to take strict actions to rectify the impact and prevent further risk. This survey presents numerous attacks and related vulnerabilities on smart IoT applications, and it also suggests the finest countermeasures to overcome the attacks on IoT applications. Thus, the study is anticipated to serve as a beneficial source for escalating the security of forthcoming IoT applications.

5 Future Scope

Security and privacy of data and devices are the top priorities of every existing and upcoming IoT applications. Still the researchers are constantly working for enhancing the security and reliability of IoT technology and its applications from imminent threats. For securing IoT devices and systems from upcoming obstructions adopt a multi-layer security approach, block-chain technology, because it is crucial to ensure the ease of communication as well as data transmission between the devices and maintain the security of evaluative data. Thus, it is helpful in managing the services, IoT apps, and devices effortlessly.

References

1. Ghazal TM, Hasan MK, Alshurideh MT, Alzoubi HM, Ahmad M, Akbar SS, Al Kurdi B, Akour IA (2021) IoT for smart cities: machine learning approaches in smart healthcare—a review. *Fut Internet* 13(8):218
2. Song T, Li R, Mei B, Yu J, Xing X, Cheng X (2017) A privacy preserving communication protocol for IoT applications in smart homes. *IEEE Internet Things J* 4(6):1844–1852
3. Dian FJ, Vahidnia R, Rahmati A (2020) Wearables and the Internet of Things (IoT), applications, opportunities, and challenges: a survey. *IEEE Access* 8:69200–69211
4. Ghazal TM, Hasan MK, Hassan R, Islam S, Abdullah HS, Afifi MA, Kalra D (2020) Security vulnerabilities, attacks, threats and the proposed countermeasures for the Internet of Things applications. *Solid State Technol* 63(1s):2513–2521
5. Ahamed J, Rajan AV (2016) Internet of Things (IoT): application systems and security vulnerabilities. In: 2016 5th International conference on electronic devices, systems and applications (ICEDSA). IEEE, pp 1–5
6. Mohammad Z, Abu Qattam T, Saleh K (2019) Security weaknesses and attacks on the Internet of Things applications. In: 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, pp 431–436

7. Alsaidi A, Kausar F (2018) Security attacks and countermeasures on cloud assisted IoT applications. In: 2018 IEEE international conference on smart cloud (SmartCloud). IEEE, pp 213–217
8. Du J, Jiang C, Gelenbe E, Xu L, Li J, Ren Y (2018) Distributed data privacy preservation in IoT applications. *IEEE Wirel Commun* 25(6):68–76
9. Jabbar WA, Alsibai MH, Amran NSS, Mahayadin SK (2018) Design and implementation of IoT-based automation system for smart home. In: 2018 International symposium on networks, computers and communications (ISNCC). IEEE, pp 1–6
10. Kolias C, Kambourakis G, Stavrou A, Voas J (2017) DDoS in the IoT: Mirai and other botnets. *Computer* 50(7):80–84
11. Chehida S, Baouya A, Bozga M, Bensalem S (2020) Exploration of impactful countermeasures on IoT attacks. In: 2020 9th Mediterranean conference on embedded computing (MECO). IEEE, pp 1–4
12. Hassija V, Chamola V, Saxena V, Jain D, Goyal P, Sikdar B (2019) A survey on IoT security: application areas, security threats, and solution architectures. *IEEE Access* 7:82721–82743
13. Bhattacharai S, Wang Y (2018) End-to-end trust and security for Internet of Things applications. *Computer* 51(4):20–27
14. Rajendran G, Ragul Nivash RS, Parthy PP, Balamurugan S (2019) Modern security threats in the Internet of Things (IoT): attacks and countermeasures. In: 2019 International Carnahan conference on security technology (ICCST). IEEE, pp 1–6
15. Abdelmaboud A, Ahmed AIA, Abaker M, Eisa TAE, Albasheer H, Ghorashi SA, Karim FK (2022) Blockchain for IoT applications: taxonomy, platforms, recent advances, challenges and future research directions. *Electronics* 11(4):630
16. Mishra N, Pandya S (2021) Internet of things applications, security challenges, attacks, intrusion detection, and future visions: a systematic review. *IEEE Access*
17. Podder P, Mondal M, Bharati S, Paul PK (2021) Review on the security threats of internet of things. *arXiv preprint arXiv:2101.05614*
18. Nagy R, Németh K, Papp D, Buttyán L (2021) Rootkit detection on embedded IoT devices. *Acta Cybernet* 25(2):369–400
19. Ali I, Ahmed AIA, Almogren A, Raza MA, Shah SA, Khan A, Gani A (2020) Systematic literature review on IoT-based botnet attack. *IEEE Access* 8:212220–212232
20. Alhammadi NAM, Zaboon KH (2022) A review of IoT applications, attacks and its recent defense methods. *J Glob Sci Res* 7(3):2128–2134
21. Kaur V, Rani S (2017) Prevention/detection methods of black hole attack: a review. *Advances in wireless and mobile communications. ISSN 0973-6972* 10:747–756
22. Kassim MRM (2020) IoT applications in smart agriculture: Issues and challenges. In: 2020 IEEE conference on open systems (ICOS). IEEE, pp 19–24
23. Alam T (2021) Cloud-based IoT applications and their roles in smart cities. *Smart Cities* 4(3):1196–1219
24. Friha O, Ferrag MA, Shu L, Maglaras LA, Wang X (2021) Internet of Things for the future of smart agriculture: a comprehensive survey of emerging technologies. *IEEE CAA J Autom Sinica* 8(4):718–752
25. Khan RL, Priyanshu D (2022) Internet of Things-based human healthcare monitoring system. In: *Soft computing: theories and applications*. Springer, Singapore, pp 869–879
26. Choubey DK, Gupta A, Suvvari S, Pathak N (2022) IoT driven precision cultivation for diverse Indian climate conditions. In: *Soft computing: theories and applications 2022*. Springer, Singapore, pp 275–282

Real-Time Implementation of Laguerre Neural Network-Based Adaptive Control of DC-DC Converter



Sasank Das Gangula, Tousif Khan Nizami, U. Ramanjaneya Reddy, and Priyanka Singh

Abstract Applications of power electronic converters have increased invariably in fields of engineering such as robotics, e-mobility and smart grids. DC-DC converters are employed as a switching devices to obtain a required amount of DC voltage in various industrial applications. Under the class of non-isolated DC-DC power converters, the buck converters are of specific interest, as they provide lower DC output voltage than the source DC voltage. In order to obtain a faithful output voltage tracking despite disturbances affecting the system, the converter is connected in the closed feedback loop. In this respect, this paper presents the design, development and experimental findings of Laguerre neural network driven adaptive control of DC-DC buck power converter. The stability of the proposed controller is established through Lyapunov stability criterion. Further, the results are compared with adaptive backstepping control method, by subjecting the converter to start-up test, step changes in the load resistance, input voltage and reference voltage tests. Thereafter, the performance is evaluated on DSP-based dSPACE 1104 processor in the laboratory. Finally, the results are compared in terms of settling time of output voltage state. The results indicate an enhanced dynamic performance of both output voltage and inductor current with the action of proposed controller, thus making it suitable for fast practical applications.

Keywords DC-DC converter · Load disturbances · Neural network · Output voltage · Transient performance

1 Introduction

DC-DC converters are essential components in power systems for providing a fixed output DC voltage across its load ends. They are well known for their high efficiency and compact size. They are used in smart grids, telecommunication, renewable energy and automobile industry. Amongst them buck, boost and buck-boost are

S. D. Gangula (✉) · T. K. Nizami · U. Ramanjaneya Reddy · P. Singh

School of Engineering and Sciences, SRM University AP, Guntur, Andhra Pradesh 522502, India
e-mail: sasankdas_g@srmmap.edu.in

the non-isolated DC-DC converter typologies. In specific, DC-DC buck converter is non-smooth, nonlinear and time varying systems. Therefore, a closed-loop feedback controller is provided to faithfully track the output DC voltage, besides guaranteeing the stability of the DC-DC converter. It must be noted that the challenges faced during the design of robust controller for output voltage regulation are as follows: (i) unaccounted voltage drop in power switching devices, (ii) unaccounted feedback sensor dynamics, (iii) stray inductances in wires, (iv) time delays, (v) input voltage disturbances and (vi) load perturbations. In this connection, authors in the past have proposed proportional integral derivative (PID)/proportional integral (PI) controllers, wherein small signal transfer function model is adopted for the design. Though PI/PID controllers are cost effective and provide a satisfactory response during start-up, however, the transient response is unsatisfactory during large load and input voltage perturbations. Besides, multiple objective evolutionary algorithms have also been proposed for tuning these linear controllers such as artificial ecosystem-based optimization, hybrid whale optimization and multi-loop constant on-time controller. However, these offline tuned controllers are computationally complex and time consuming.

Of late, advanced controller has been proposed such as fuzzy logic control (FLC), which is a rule-based method and does not require complete mathematical model of the plant. In [1], genetic algorithm (GA)-based FLC is proposed for battery charging application. Adaptive fuzzy synergistic controller has also been reported in [2]. Nonetheless, the transient response of output voltage is near-satisfactory, besides increasing the computational burden.

Amongst nonlinear controllers, sliding mode control (SMC) based on variable structure systems control theory [3] has received considerable attention in past few years. It claims to increase the robustness and provides fair dynamic response under matched perturbations. However, because of their inherent chattering problem, the converter operates at high and variable switching frequency. This leads to an excessive power loss and electromagnetic interference through the power semiconductor switches [4]. H_∞ controller is also another type of nonlinear controller proposed in the literature. It is based on Riccati equations and linear matrix inequalities (LMI) methods to derive a robust optimal control law, while minimizing the H_∞ norm of the output state with respect to the uncertainty. In the recent past, neural network-based controllers have also been proposed for DC-DC converters [5]. A backpropagation artificial neural network (ANN)-based controller is proposed in [6] and evaluated against PID for different DC-DC converters.

Backstepping control (BSC), developed by Kokotovic in 1990, is another efficient nonlinear control methodology based on recursive design method. Some of the earlier works for DC-DC converters can be seen in [7–10]. In this connection, this article proposes design, development and experimental validation of a novel online adaptive Laguerre neural network-based backstepping control for the output voltage regulation in DC-DC buck converter [11]. The proposed design simplifies the control structure and ensures sufficient performance under wide operating domain. Stability analysis and convergence to the stable tracking solution have been proved under Lyapunov sense. Experimental validation of the proposed control on buck converter

using dSPACE DS1104-based MPC8240 processor is conducted, to showcase the output tracking performance of the proposed controller and conventional adaptive backstepping control (ABSC) technique intended for same application.

Remaining part of the article is summarized as follows: Sect. 2 presents the mathematical modelling of the DC-DC converter followed by the problem formulation. Section 3 describes the proposed control algorithm with detailed stability analysis. Section 4 presents the real-time experimental findings followed by its discussion. Finally, conclusions are drawn in Sect. 5.

2 Converter Modelling and Problem Formulation

Figure 1 presents the topology of a DC-DC buck converter with resistive load R at its output terminals. The input DC voltage source, inductor, capacitor and the freewheeling diode are denoted as E , L , C and D , respectively. The switching action is performed by the switch S_w . During ON mode of the converter, when S_w is closed the freewheeling diode is reverse biased and hence, the energy is directly supplied from source to the load. However, during OFF mode of the converter, the input voltage E is disconnected and the diode allows the inductor to supply the stored energy across the load resistor. Assuming the converter operation in a continuous conduction mode, the mathematical description is presented as follows; let the output voltage across the load resistance R be v_o and the current through inductor as i_L . Selecting state variables as $x_1 = v_o$ and $x_2 = i_L$, the state space averaged model [12] can be written as can be described as

$$\dot{x}_1 = -\frac{x_1}{RC} + \frac{x_2}{C}$$

$$\dot{x}_2 = -\frac{x_1}{L} + \frac{uE}{L}$$

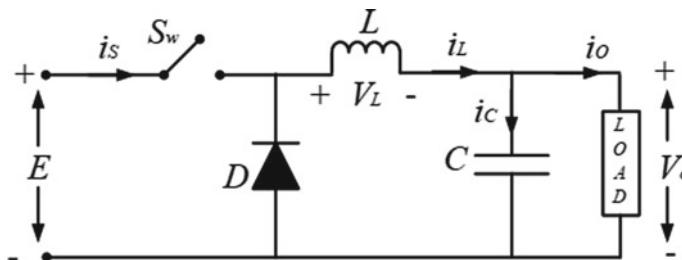


Fig. 1 DC-DC buck converter

where the control input $u \in [0, 1]$ represents the opening and closing of the switch S_w . In addition, the control law is fed to a PWM generator for producing the control signal u_d .

Here, the control objective is to attain a faithful output voltage tracking of v_o with respect to reference voltage v_r , under both nominal condition and at the event of any variation in load and input affecting the converter system.

3 Proposed Control Scheme

The proposed integrated control structure includes the neural network for the approximation of load uncertainties within the framework of adaptive backstepping control [11] for DC-DC buck converter. The is explained below.

3.1 Laguerre Neural Network

Motivated from the universal approximation characteristics of Laguerre neural networks, the proposed control structure exploits Laguerre neural network to online estimates the unknown time varying load term R . Therefore, the unknown term to be approximated here is

$$f(\cdot) = x_1/RC \quad (1)$$

from (1). The unknown term can be estimated as

$$f(\cdot) = W^* \Phi(x_1) + \varepsilon \quad (2)$$

with $W^* := [w_1^* \ w_2^* \ w_3^* \ \dots \ w_\rho^*]^T$ as the weights of the neural network and the estimation error is mentioned as ε . The regressor $\Phi(x_1) := [L_0(x_1) \ L_1(x_1) \ \dots \ L_{\rho-1}]^T$ is used as a bases vector. In this paper, Laguerre polynomial in the single functional layer is used. Here, $L_i(x_1)$, $i = 0, \dots, (\rho - 1)$ indicates the Laguerre polynomials represented by $L_{i+1}(x) = (2i + 1 - x)L_i(x) - iL_{i-1}(x)i + 1$ and ρ is the size of the polynomial function. Moreover $L_0(x) = 1$ and $L_1(x) = 1 - x$.

3.2 Controller Design

Adaptive controller is designed based on the backstepping control [13] and then integrated with the Laguerre Neural Network. The design procedure is as follows, the first error variables be defined as

$$z_1 = x_1 - v_r, \quad z_2 = \frac{x_2}{C} - \alpha(\cdot) \quad (3)$$

where v_r is the reference output voltage. Next finding

$$\dot{z}_1 = -W^{*T} \Phi(x_1) + z_2 + \alpha(\cdot) \quad (4)$$

and selecting virtual control input $\alpha(\cdot)$ for z_1 . Thereby considering the Lyapunov function

$$V_{z_1} : \mathbb{R} \times \mathbb{R}^\rho \times \mathbb{R}_+ \rightarrow \mathbb{R}_+ \quad (5)$$

as

$$V_{z_1} = \frac{1}{2} z_1^2(t) + \frac{1}{2\gamma} \tilde{W}^T(t) \tilde{W}(t) \quad (6)$$

Here, $z_1(t)$ is the error for voltage tracking and

$$\tilde{W}(t) = W^*(t) - \hat{W}(t) \quad (7)$$

is the estimation error. Further differentiating V_{z_1} , we obtained as follows:

$$\dot{V}_{z_1} = z_1 \left(-W^{*T} \Phi(x_1) + z_2 + \alpha(\cdot) \right) - \tilde{W}(t) \dot{\hat{W}}(t) \quad (8)$$

Now taking

$$\alpha = -c_1 z_1 + \hat{W}^T \Phi(x_1) \quad (9)$$

and

$$\dot{\hat{W}}(t) = -\gamma \Phi(x_1) z_1(t) \quad (10)$$

will yield

$$\dot{V}_{z_1} = -c_1 z_1^2 + z_1 z_2 \quad (11)$$

Further, to $u(t)$, so that it leads to asymptotically stabilizing the error variable z_2 . The Lyapunov function is taken as $V_{z_2} : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ given by

$$V_{z_2} = \frac{z_2^2}{2} \quad (12)$$

Taking derivative

$$\dot{V}_{z_2} = z_2 \left(-\frac{x_1}{LC} + \frac{uE}{LC} - \dot{\alpha} \right) \quad (13)$$

Expressing

$$u(t) = \frac{LC}{E} \left(-z_1 - c_2 z_2 + \frac{x_1}{LC} + \dot{\alpha} \right) \quad (14)$$

results in

$$\dot{V}_{z_2} = -c_2 z_2^2 - z_1 z_2 \quad (15)$$

Next, taking the derivative of V_z

$$\dot{V}_z = - \sum_{i=1}^2 c_i z_i^2 < 0 \quad (16)$$

Therefore, the negative definiteness of \dot{V}_z is proved. This means

$$-\int_{t_0}^{\infty} \dot{V}_z dt \leq V_z(t_0) - V_z(\infty) \leq \int_{t_0}^{\infty} z^T \mathcal{Q} z dt \leq \infty \quad (17)$$

where $z := [z_1 \ z_2]^T$ and $\mathcal{Q} := \text{diag}\{c_1, c_2\} > 0$.

The above expression on closed-loop signals results in $z(t) \in \mathcal{L}_2 \cap \mathcal{L}_{\infty}$ and $\dot{z}(t) \in \mathcal{L}_{\infty}$ besides boundedness of weights, i.e. $W(t), \dot{W}(t) \in \mathcal{L}_{\infty}$. Hence from Barbalat's signal convergence lemma, the stability of $z(t)$, i.e. $\lim_{t \rightarrow \infty} z(t) = 0$ means that $z_1(t)$ finally reaches to zero asymptotically. The closed-loop block diagram of the proposed control is shown in Fig. 2.

4 Experimental Results and Discussion

A real-time experimental test bench is made in the laboratory as shown in Fig. 3. It has a DC-DC buck converter and a controller dSPACE DS1104 with MPC8240 processor. The input voltage is supplied from a DC power supply. Power MOSFET IRFP460 with a rating of 500 V, 20 A is used as a switching device. The voltage and the current states are measured using LV-25P and LA-55P sensors. The dSPACE Control Desk DS1104 is used with a desktop computer with Intel® Core™ i7 with 3.2 GHz processor frequency.

To evaluate the proposed controller, the tests undertaken are (i) start-up, (ii) reference tracking, (iii) load disturbance and (iv) input voltage supply perturbations. Moreover, the dynamic performance has been compared with adaptive backstepping control (ABSC) [8]. The specifications of the converter under consideration are given in Table 1. The controller design parameters are selected as $c_1 = 10^4$, $c_2 = 1000$ and $P = 0.001 \times 10^{-1} I_2$. The tests conducted are discussed in detail hereafter.

- (i) **Start-up:** The output voltage response during the DC-DC converter start-up from 0 to 10 V is plotted in Fig. 4a, b under the operation of conventional ABSC and the proposed control. The ABSC method takes a long settling time of 0.650 s to reach the reference voltage, whereas the proposed control takes

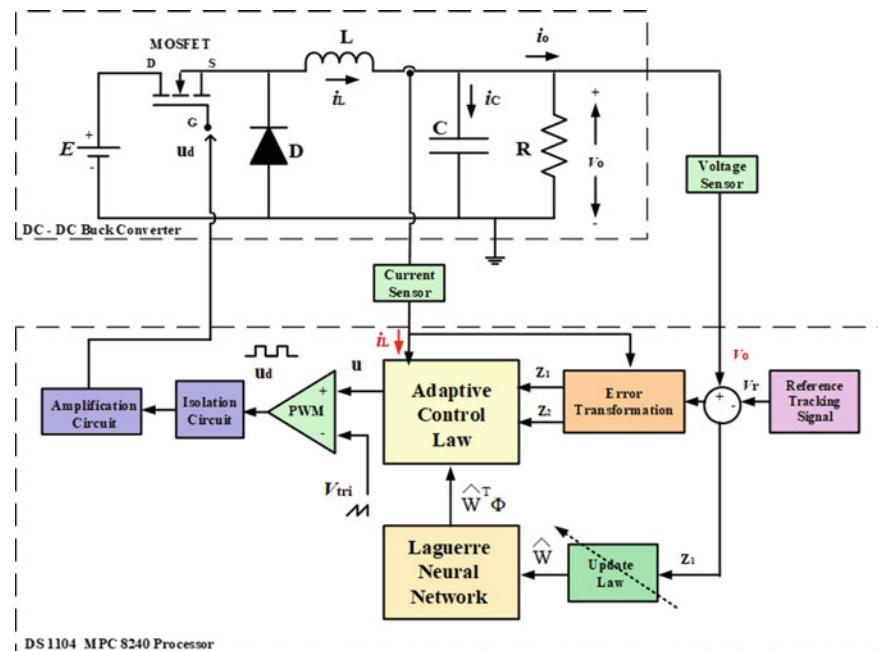


Fig. 2 Block diagram of the proposed controller

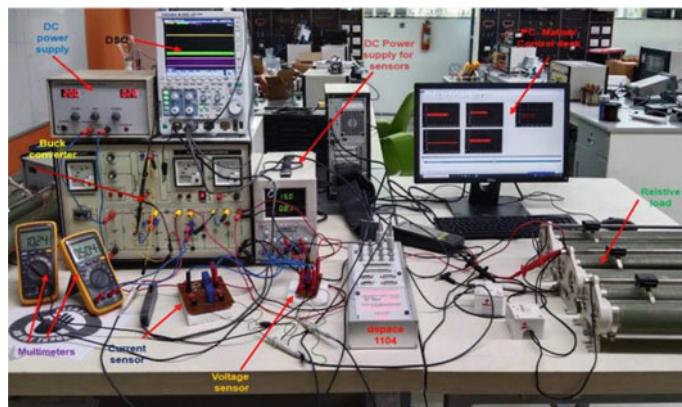


Fig. 3 Experimental laboratory set-up

Table 1 DC-DC buck converter specifications

Parameter	Rating
Input voltage, E	20 V
Nominal output voltage, v_o	10 V
Nominal output current, i_o	0.5 A
Nominal load resistance, R	20 Ω
Inductor, L	10 mH
Power rating, P	5 W
Capacitor, C	200 μF
PWM frequency, f	10 kHz

0.525 s. Besides, it can be seen that the proposed control offers better inductor current response as depicted in Fig. 4b compared to ABSC in Fig. 4a.

- (ii) **Reference trajectory tracking:** Figure 4c, d presents the output trajectory tracking performance with respect to a step change of reference voltage from 10 to 15 V. As it can be noticed, the ABSC methods take 0.15 s and the proposed LGNN-ABSC takes 0.1 s to track the desired output voltage. Similarly, the dynamic response of inductor current during reference trajectory tracking is also plotted in Fig. 4c, d, which indicates the fast response of inductor current under the proposed control method.
- (iii) **Load disturbance:** Further, the real-time buck converter system is subjected to a change in load resistance from 20 to 40 Ω and vice-versa as shown in Fig. 4e, f under ABSC and the proposed control. As Fig. 4e shows that the output voltage takes longer time of 0.8 s to reach to nominal value of 10 V with ABSC and with the proposed control it takes 0.45 s. Similarly, during the load change from 20 to 10 Ω and vice-versa, as shown in Fig. 5a, b under ABSC and the proposed control. The results confirm superior performance by the proposed method proposed by yielding fast settling and lesser peak overshoot/undershoot.
- (iv) **Input voltage disturbance:** At the end to ensure the performance of the proposed control method, the buck converter system is subjected to a sudden change in input DC voltage from 20 to 25 V and vice-versa. During this test, the effectiveness of proposed control and ABSC are evaluated and thus, the results are shown in Fig. 5c, d.

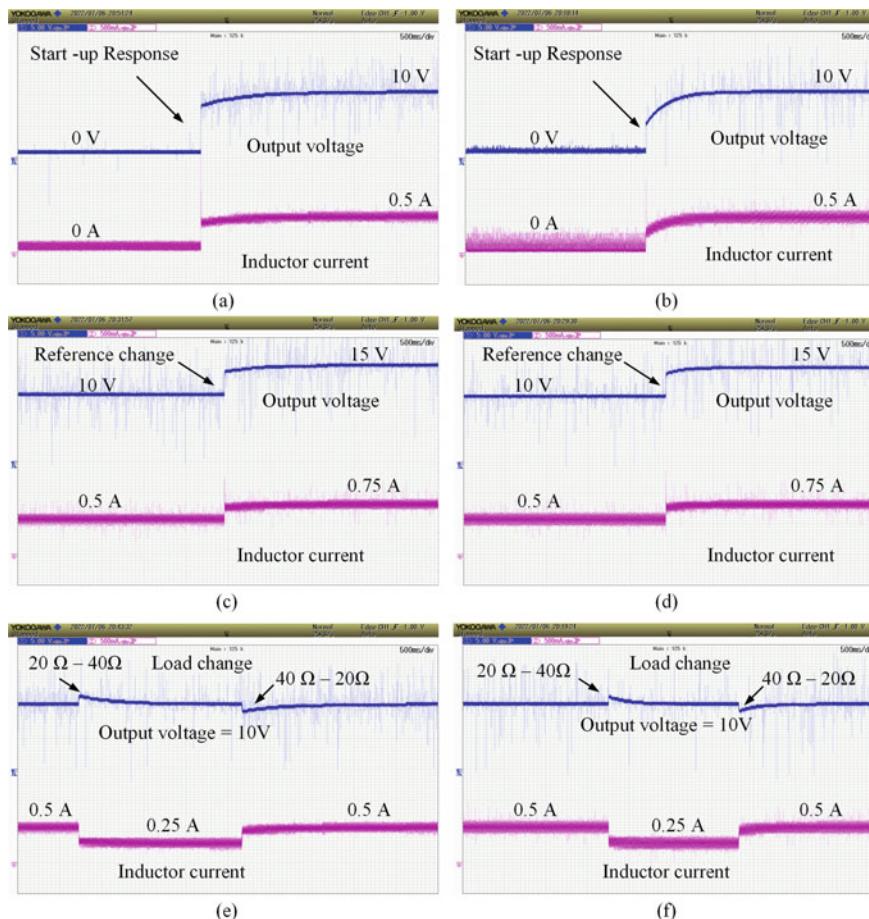


Fig. 4 Experimental results: output voltage and inductor current response during start-up **a** ABSC and **b** proposed control; output voltage and inductor current response during reference tracking from 10 to 15 V **c** ABSC and **d** proposed control; output voltage and inductor current response during load change from 20 to 40 Ω and vice-versa **e** ABSC and **f** proposed control

5 Conclusion

This article presents experimental validation of Laguerre neural network (LGNN)-based adaptive backstepping control for DC-DC buck converter. The proposed controller approximates the mismatched load uncertainty in an online manner and compensates during the control action through the backstepping controller. The closed-loop stability of the controller proposed is established through Lyapunov stability criterion. Extensive real-time experimentation has been conducted to find out the superiority of proposed controller over a wide operating regime under output voltage

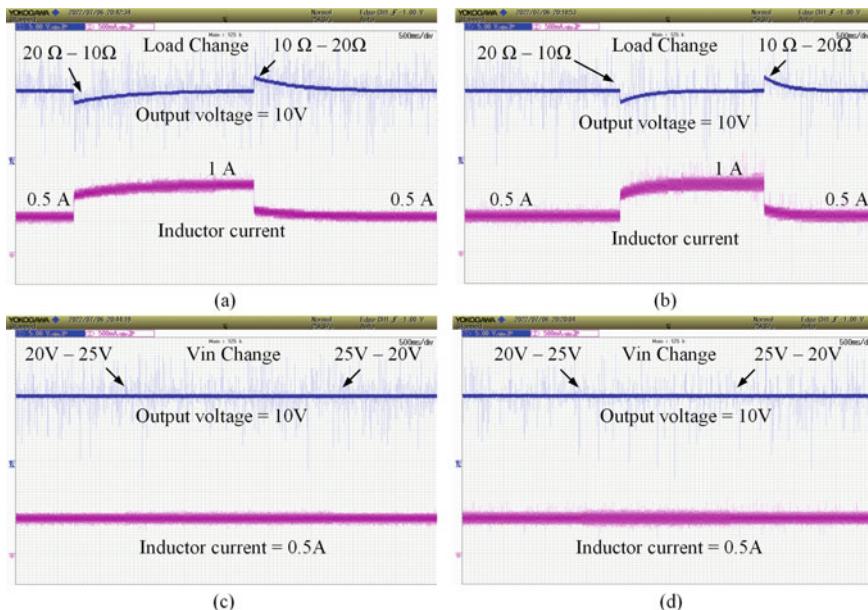


Fig. 5 Experimental results: output voltage and inductor current response during load change from 20 to 10 Ω and vice-versa **a** ABSC and **b** proposed control; output voltage and inductor current response during input voltage change test 20–25 V and vice-versa **c** ABSC and **d** proposed control

tracking, input and load disturbances. It is concluded that the proposed controller preforms faster in estimating the unknown load function and thereby yielding an enhanced dynamic performance of output voltage in DC-DC converter.

References

1. Shieh C-S (2014) Fuzzy PWM based on genetic algorithm for battery charging. *Appl Soft Comput* 21:607–616
2. Bouchama Z, Khatir A, Benaggoune S, Harmas MN (2020) Design and experimental validation of an intelligent controller for DC-DC buck converters. *J Franklin Inst* 357:10353–10366
3. Utkin V (2013) Sliding mode control of DC/DC converters. *J Franklin Inst* 350:2146–2165
4. Tan S-C, Lai Y-M, Chi KT (2008) General design issues of sliding-mode controllers in DC-DC converters. *IEEE Trans Ind Electron* 55:1160–1174
5. Renjini G, Devi V (2022) Artificial neural network controller based cleaner battery-less fuel cell vehicle with EF2 resonant DC-DC converter. *Sustain Comput Inform Syst* 35:100667
6. Ramirez-Hernandez J, Juarez-Sandoval O-U, Hernandez-Gonzalez L, Hernandez-Ramirez A, Olivares-Dominguez R-S (2020) Voltage control based on a back-propagation artificial neural network algorithm. In: 2020 IEEE international autumn meeting on power, electronics and computing (ROPEC). IEEE, pp 1–6

7. Sira-Ramirez H, Rios-Bolivar M, Zinober AS (1995) Adaptive input-output linearization for PWM regulation of DC-to-DC power converters. In: Proceedings of 1995 American control conference-ACC'95. IEEE, pp 81–85
8. Sureshkumar R, Ganeshkumar S (2011) Comparative study of proportional integral and backstepping controller for buck converter. In: 2011 international conference on emerging trends in electrical and computer technology. IEEE, pp 375–379
9. Nizami TK, Chakravarty A, Mahanta C (2017) Design and implementation of a neuro-adaptive backstepping controller for buck converter fed PMDC-motor. *Control Eng Pract* 58:78–87
10. Nizami TK, Chakravarty A, Mahanta C, Iqbal A, Hosseinpour A (2022) Enhanced dynamic performance in DC-DC converter-PMDC motor combination through an intelligent non-linear adaptive control scheme. *IET Power Electron*
11. Nizami TK, Chakravarty A (2020) Laguerre neural network driven adaptive control of DC-DC step down converter. *IFAC-PapersOnLine* 53:13396–13401
12. Komurcugil H (2012) Adaptive terminal sliding-mode control strategy for DC-DC buck converters. *ISA Trans* 51:673–681
13. Kokotovic PV (1992) The joy of feedback: nonlinear and adaptive. *IEEE Control Syst Mag* 12:7–17

5G New Radio Physical Downlink Shared Channel Throughput Analysis with Different Numerology and Modulation Schemes



Rajesh Kumar^{ID}, Deepak Sinwar^{ID}, and Vijander Singh^{ID}

Abstract 5G new radio provides support for flexibility and scalability in the form of multiple numerologies and the air interface to meet the different use case scenarios that were proposed by International Mobile Telecommunications-2020. Broadly physical channels are categorized into the physical downlink shared channel (PDSCH), the physical downlink control channel (PDCCH), and the physical uplink shared channel (PUSCH). Among these PDSCH is proposed for downlink shared channel that supports multiple numerologies and flexibility in modulation scheme. In this paper, an assessment of PDSCH throughput is presented on several sub carrier spacing (SCS) and modulation schemes. A comparative study of throughput and signal-to-noise ratio (SNR) is also carried out that indicates the impact on throughput due to a complex modulation scheme. Based on experimental evaluations, it can be stated that 16-QAM outperformed, 64 and 256-QAMs on 120 kHz SCS which indicates a significant improvement in the PDSCH throughput.

Keywords New radio · Physical downlink shared channel · Subcarrier spacing · Throughput · Signal to noise ratio

1 Introduction

5G new radio (NR) design supports flexible numerology and scalability in using air interface to meet the enhanced mobile broadband services (eMBB), ultra-reliable low latency communication (URLLC), and massive machine type communication (mMTC) services [1]. These services support a wide range of applications and usage scenarios in the wireless environment. eMBB provides higher broadband services

R. Kumar · D. Sinwar (✉)

Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur 303007, India

e-mail: deepak.sinwar@gmail.com

V. Singh

Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur 303007, India

that demand high data rate requirements. URLLC provides low latency communication, especially for a mission-critical application, whereas mMTC supports the connection of billions of IoT and machine-to-machine (M2M) devices, where low data rate transfer is required [2]. Designing the physical layer architecture of 5G NR is a critical part as it required effective communication with the hardware devices. Massive multiple input multiple output (mMIMO) and beamforming technology are enhancements adopted by the 5G NR as compared to LTE architecture [3]. According to 3GPP, broadly 5G NR provides several physical channels, i.e., physical downlink shared channel (PDSCH), physical downlink control channel (PDCCH), physical uplink shared channel (PUSCH), etc., as mentioned below:

- PDSCH is designed for downlink data transmission that supports up to 400 MHz bandwidth band with orthogonal frequency division multiplexing (OFDM) and low-density parity check (LDPC). The modulation scheme supported by PDSCH is QPSK, 16-QAM, 64-QAM, and 256-QAM.
- PDCCH is designed to transmit control information in the downlink direction. It also supports OFDM (with polar coding) and QPSK modulation [4].
- PUSCH is designed to transfer uplink data transmission, which also supports up to 400 MHz bandwidth with LDPC encoding technique with QPSK, 16-QAM, 64-QAM, and 256-QAM modulation schemes. However, the 400 MHz channel bandwidth depends on the type of sub-carrier spacing utilized [5].

As per user equipment is concerned, it typically monitors the PDCCH signal as one slot in the resource grid. In case, a mini slot (URLLC data) is being utilized for data transmission, then it can utilize more than one slot. Once the PDCCH is detected, UEs receive one unit of data named as transport block on the PDSCH and follow the scheduling decision of gNB. UEs afterward respond with a hybrid automatic repeat request (HARQ) to denote that the data was successfully decoded or not using ACK/NACK. In case, it receives NACK, data retransmission will be further initiated [6]. A set of physical signals are also utilized for effective data transmission. These signals are separated in uplink and downlink data transmission. For the downlink data transmission, 5G NR maintains several signals viz. demodulation reference signals, phase tracking reference signals, channel state reference signals, primary synchronization signals, and secondary synchronization signals [4].

5G NR supports multiple SCS that provides users with the following advantages, i.e., scalability in the network; wider SCS that makes the system more robust for phase noise; low latency communication; deployment of smaller cell sizes; supports a wider range of deployment scenario starting from 1 GHz to millimeter-wave; and flexible channel bandwidth utilization in PDSCH for efficient spectrum utilization [7].

In comparison with 5G, LTE was having a single subcarrier spacing of 15 kHz only. On the other hand, there are differences between the 15 kHz SCS of LTE and NR [8]; in 15 kHz SCS of LTE, there are only seven OFDM symbols, whereas 15 kHz SCS of NR contains 14 OFDM symbols [9]. 3GPP also introduced the concept of a mini slot in 5G NR that was not available in LTE. In the mini slot, the transmission slot is shorter than the regular slot duration. It is also possible to fit an integer number

Table 1 Multiple numerology with range of frequency [12]

Numerology (μ)	SCS (kHz)	OFDM symbol	Slot duration (time domain) in ms	PRB bandwidth (kHz)	Max channel bandwidth (MHz)	Range of frequency
0	15	14	1	180	49.5	FR1
1	30	14	0.5	360	99	FR1
2	60	14/12	0.25	720	198	FR1/FR2
3	120	14	0.125	1440	396	FR2
4	240	14	0.0625	2880	397.44	FR2

of slot in the narrow band subcarrier spacing into the wide band subcarrier spacing slot, but time alignment is important for time division duplex (TDD) network [10]. One slot of 30 kHz SCS is having a duration of 0.5 ms in the NR frame structure, 60 kHz SCS is having 0.25 ms, and 120 kHz has 0.125 ms [11]. Table 1 shows multiple numerologies of 5G NR. Here, frequency range-1 (FR1) indicates low-band frequency (less than seven GHz) and frequency range-2 (FR2) indicates high-band frequency (millimeter waves having a frequency of more than 24 GHz).

Few researchers also contributed to the throughput analysis of 5G NR using different channel modeling schemes.

In this regard, Ömer et al. [13] analyzed physical downlink shared throughput using 15, 30, and 60 SCS and different modulation schemes. The authors concluded that the 60 kHz SCS provides better throughput, and a higher modulation scheme can contribute more to the PDSCH throughput. However, the analysis depends on the size of the antenna and PDSCH mapping layer. The authors in [14] analyzed the PUSCH throughput in the uplink direction and observed that in low SNR value, the QPSK modulation scheme provides finer output, whereas, with a high SNR value, 256-QAM contributed to higher throughput. Similarly, PDSCH throughput increases if higher SCS is chosen. They also analyzed the behavior of the propagation model and observed that a clustered delay line (CDL) model contributed to higher throughput in the uplink direction.

Beamforming vectors also plays important role in PDSCH throughput, therefore an experimental studied carried out by Panda and Ramanath [15] for a multiuser framework analysis the codebook-based beamforming in 5G NR. Various parameters were considered for analyzing beamforming viz. numbers of layers, antenna configuration, etc. The codebook generated, and accurate beamforming contributes to the subsequent improvement in PDSCH throughput.

Verdecia-Peña and Alonso [16] proposed a decode and forward (D&F) cooperative hardware network for 5G NR, it is a layer 2 MIMO relay technology. The key performance indicators (KPI) such as error vector magnitude (EVM), bit error rate (BER), and throughput were measured with 64 and 256-QAM modulation schemes. D&F network made subsequent improvements in KPIs while communicating gNB to UEs with the outdoor-to-indoor scenario.

NR air interface is developed by 3GPP for the 5th generation mobile communication system with a wide range of use case scenarios. Takeda et al. [17] discussed the physical downlink control channel release for 5G mobile communication with the functionality of physical design framework, monitoring scheme, beamforming operation, and other carried information. On the other hand, Launay [18] presented 5G NR system characteristics, master information block messages, and system information block messages for the broadcast control channel (BCCH).

Flexibility and scalability issues in 5G NR can be resolved using different numerologies support and different use case scenarios. In this regard, Indoondon and Fowdur [19] compared various coding techniques to meet the end-user requirements and discussed several challenges of implementation in actual environments. The authors also assessed the performance of LDPC coding and the polar coding scheme that is used in 5G NR. They observed the best bit error rate (BER) and block error rate (BLER) by implementing the lower code rate with lower QAM. In addition, they observed higher throughput gain using a higher code rate and higher QAM.

In short, the major contributions of this work are highlighted as follows:

- To analyze the PDSCH throughput by varying different SCS (i.e., 15, 30, 60, and 120 kHz) and different modulation schemes (i.e., 16, 64, 256-QAM).
- To compare the performance of PDSCH throughput with different SNR ranges (i.e., -5, 0, 5).
- To find out the best modulation scheme for PDSCH that can contribute to higher throughput.

The rest of the paper is organized as follows. Section 2 describes the model of PDSCH, Sect. 3 presents an illustration of PDSCH using an experimental setup followed by results and discussions in Sect. 4. Finally, Sect. 5 concludes the work.

2 System Model of PDSCH

Physical channels define the flow of information that is transmitted from UEs to gNB and through various physical channels as discussed in the previous section. The main work of PDSCH is to perform slot-based mapping. This mapping information is further received by UEs in the form of PDSCH mapping type and PDCCH signals [20]. Figure 1 shows the downlink shared channel and “PDSCH transmit and receive processing chain”.

One or two transport block is mapped with the DL-SCH encoder processing chain. The DL-SCH encoding process consists of cyclic redundancy check (CRC), code block segmentation and CRC, low-density parity-check (LDPC) encoding, rate matching, and code block concatenation as per the phase defined in the 3GPP standard [21]. The scramble code is received by PDSCH from DL-SCH, and it is further transmitted to the precoding. PDSCH will decide which encoding technique and modulation technique will be applied to the scramble data. After the decision execution of a complex modulation scheme, these scramble data will be converted into

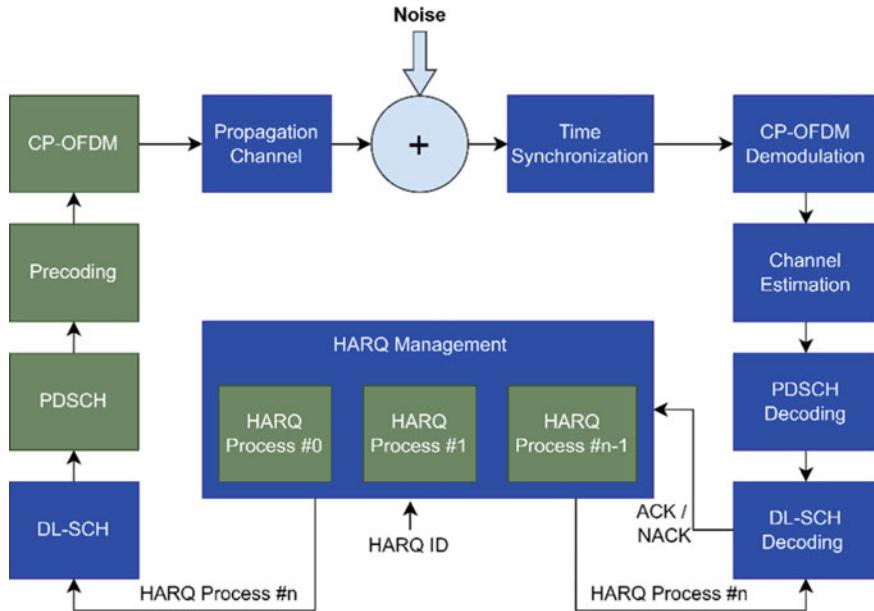


Fig. 1 Downlink shared channel (DL-SCH) and PDSCH transmit and receive processing chain [22]

complex value symbols. PDSCH adopts modulation schemes such as QPSK, 16-QAM, 64-QAM, and 256-QAM. In the QPSK, single pair of scramble bits is transferred for the modulation, whereas in 16, 64, and 256-QAM, a pair of 2, 3, and 4 scramble bits, respectively, are being transferred to the complex value symbols. The role of the antenna port comes in the action afterward. These individual complex value modulated symbols are then mapped with the given antenna ports as per the layered mapping technique. The precoding block will perform the mapping in all the transmission layers of the complex value modulated signal. Finally, to activate the transmission, PDSCH mapped all antenna ports with layer mappings that results in complex symbols. These symbols are then mapped to the resource blocks (RBs) for the generation of OFDM signals in all antenna ports [13].

3 A PDSCH Illustration Using Experimental Setup

A PDSCH throughput model is implemented using MATLAB 2022a with the simulation parameters that are presented in Table 2. The authors used a CDL propagation model for an urban microcell scenario with multiple SCS, normal cyclic prefix, and the channel bandwidth of 20 MHz (except 120 kHz SCS that was implemented with 50 MHz). The simulation also used 8×2 PDSCH transmitting and receiving

Table 2 Parameter setting for the different numerology in the simulation

Numerology (μ)	SCS	N_{rb}	SNR range	Number of 10 ms frame	Number of PDSCH layers	QAM
0	15	106	[- 5 0 5]	2	2	16, 64, 256
1	30	51	[- 5 0 5]	2	2	16, 64, 256
2	60	24	[- 5 0 5]	2	2	16, 64, 256
3	120	32	[- 5 0 5]	2	2	16, 64, 256

antennas and a 490/1024 code rate to calculate the transport block size. Here, the SNR represents the average SNR per resource element (RE) per receiving antenna and REs in the resource grid (that is in the frequency domain). To achieve the desired SNR, the illustration introduces an equivalent noise level in the time domain.

The number of resource blocks can be calculated using (1) as follows:

$$N_{\text{rb}} = \frac{\text{CB} - 2 \times \text{GB}}{\text{One resource block bandwidth}} \quad (1)$$

Here, N_{rb} represents a number of resource blocks, CB the channel bandwidth, and GB the guard-band bandwidth.

The SNR is calculated using (2) as follows:

$$\text{SNR} = \frac{S_{\text{RE}}}{N_{\text{RE}}} \quad (2)$$

Here, S_{RE} represents the average signal power per RE per receive antenna, and N_{RE} represents the average noise power per RE per receive antenna. N_{RE} models the AWGN that is added to the signal.

4 Results and Discussions

This section presents the performance evaluation of PDSCH throughput that was carried out by varying different SCS and modulation schemes. Figures 2 and 3 depicts the PDSCH throughput versus SNR using 16, 64, and 256-QAM on 15 kHz SCS. It is clear from Fig. 3 that the higher throughput is obtained using 16-QAM as compared to 64-QAM and 256-QAM. Similarly, it is observed from Figs. 4 and 5 that the 16-QAM modulation scheme (with 30, 60, and 120 SCS, respectively) provides higher throughput as compared to 64 and 256-QAM.

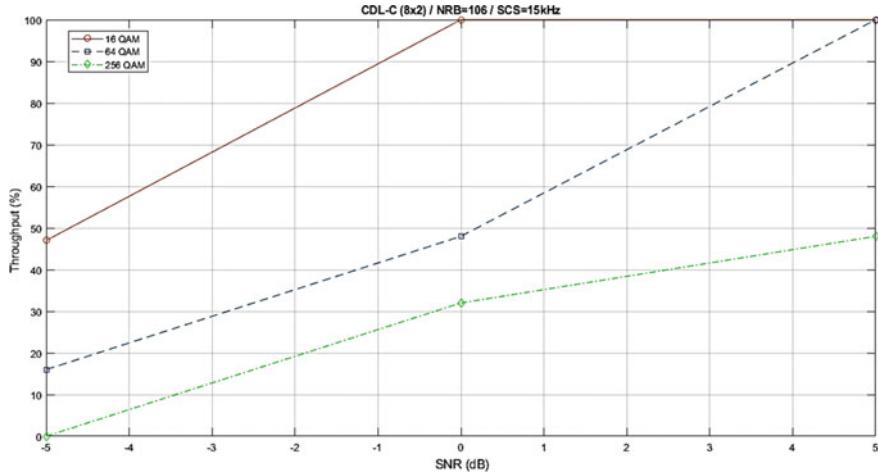


Fig. 2 PDSCH throughput evaluation of 15 kHz SCS with 16, 64, and 256-QAM modulation

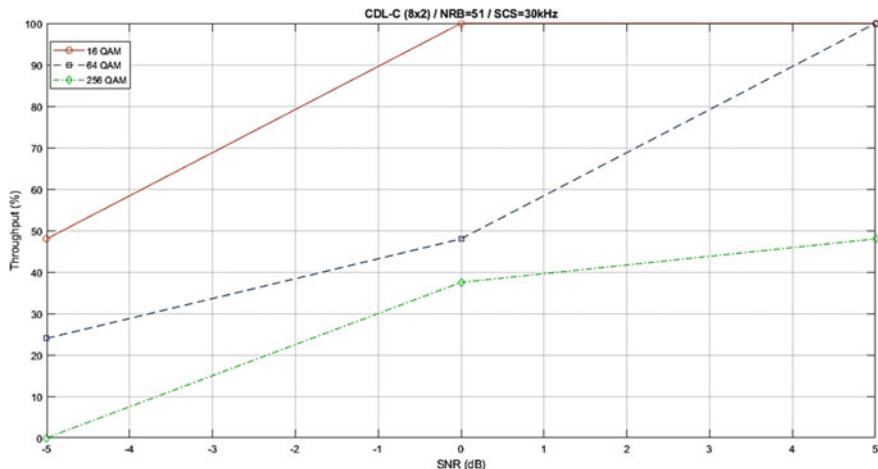


Fig. 3 PDSCH throughput evaluation of 30 kHz SCS with 16, 64, and 256-QAM modulation

4.1 Discussion

In the case of 256-QAM modulation, when the noise is present in the signal, the HARQ scheme is not able to decode the signal in the given time frame and went to the complete signal loss. The fact was also supported by Verdencia-Peña and Alonso [16] indicating that the 64-QAM is less prone to error as compared to 256-QAM. Based on experimental evaluations, it can be stated that the impact of QAM is more as compared to varied SCS for measuring signal quality when performing the HARQ

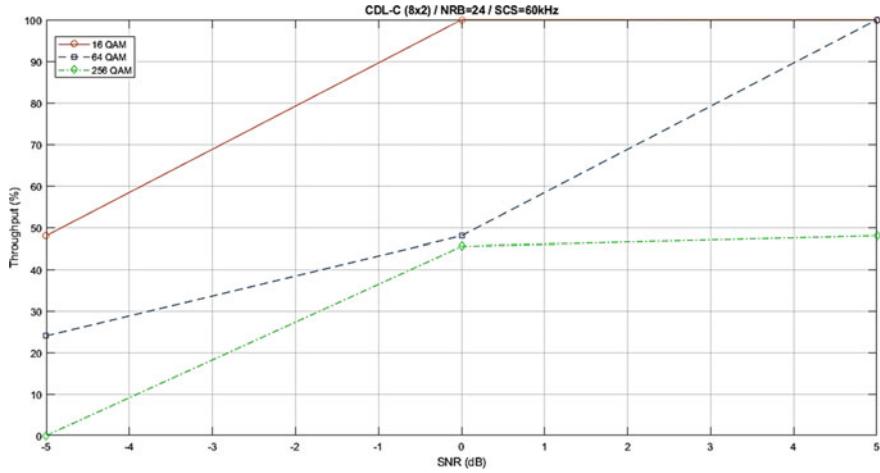


Fig. 4 PDSCH throughput evaluation of 60 kHz SCS with 16, 64, and 256-QAM modulation

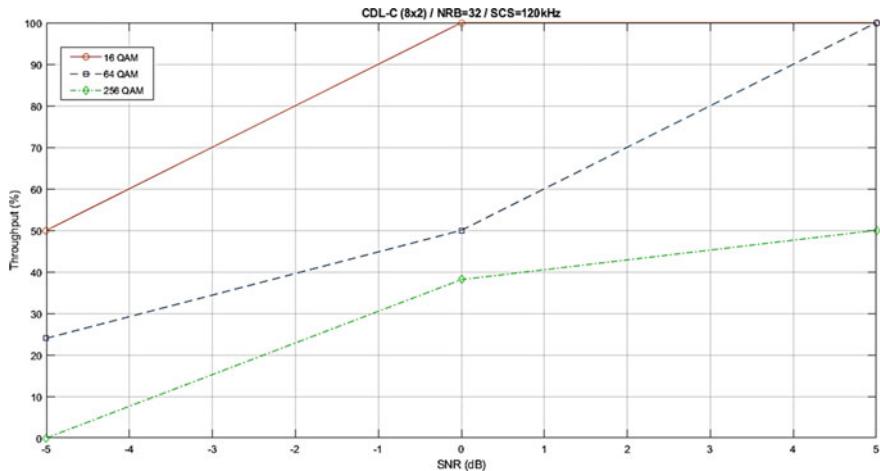


Fig. 5 PDSCH throughput evaluation of 120 kHz SCS with 16, 64, and 256-QAM modulation

process. The choice of modulation scheme greatly affects the throughput capacity. So, it is recommended to use 16-QAM to obtain higher PDSCH throughput. In addition, an SCS of 120 kHz with a 16-QAM modulation scheme provided a throughput of 50% in the case of -5 dB SNR. The PDSCH throughput is directly proportional to the increasing size of the antenna grid; the higher the antenna grid, the higher the chances of decoding signals accurately.

It is shown in Figs. 3, 4 and 5 that the performance of 256-QAM was not found satisfactory due to several reasons such as presence of noise in different scenarios and complex encoding scheme of 256-QAM. However, the result may vary if different code rate and antennas size are utilized.

5 Conclusion

The new radio is the new radio air interface developed by the 3GPP to achieve greater scalability, flexibility, and efficiency among different use case scenarios. In this paper, the throughput assessment of PDSCH is carried out to observe the impact of varying sub carrier spacing and modulation schemes. As per the 256-QAM is concerned, it should provide more throughput as compared to 16- and 64-QAM, but due to its complex encoding scheme (8-bits per symbol), it results in incorrect decoding of the signal using HARQ. The same kind of behavior was observed during experimental evaluations that indicates very less throughput (near to zero) of 256-QAM with 120 kHz SCS in noisy environments (-5db SNR). The result may vary if different propagation models and antenna sizes (grid) are utilized. On the other hand, it was also observed that the PDSCH throughput is influenced majorly by QAM rather than SCS. The impact of changing SCS does not affect more to the throughput as QAM does. In the future, the work may be extended to perform a comparative analysis of throughput with block error rate (BER) using a low-density parity check (LDPC) and polar coding scheme.

References

1. Vook FW, Ghosh A, Diarte E, Murphy M (2019) 5G new radio: overview and performance. Conf Rec—Asilomar Conf Signals, Syst Comput 2018:1247–1251. <https://doi.org/10.1109/ACSSC.2018.8645228>
2. Tian W, Lin K (2022) Chapter 2—Requirements and scenarios of 5G system. In: Shen J, Du Z, Zhang Z, Yang N, Tang H (eds) 5G NR and enhancements. Elsevier, pp 41–52. <https://doi.org/10.1016/B978-0-323-91060-6.00002-7>
3. Specification T (2021) TS 123 501—V16.6.0—5G; system architecture for the 5G system (5GS) (3GPP TS 23.501 version 16.6.0 Release 16). 0.251 [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.09.00_60/ts_123501v150900p.pdf
4. 3GPP (2020) Physical channels and modulation (3GPP TS 38.211 version 15.3.0 Release 15) 0 [Online]. Available: <https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx>
5. Morais DH (2022) 5G NR overview and physical layer. In: Key 5G physical layer technologies: enabling mobile and fixed wireless access. Springer International Publishing, Cham, pp 233–297. https://doi.org/10.1007/978-3-030-89209-8_10
6. Zaidi A, Athley F, Medbo J, Gustavsson U, Durisi G, Chen X (2018) NR physical layer: overview. 5G Phys Layer pp 21–34. <https://doi.org/10.1016/b978-0-12-814578-4.00007-2>
7. Correia N, Al-Tam F, Rodriguez J (2021) Optimization of mixed numerology profiles for 5g wireless communication scenarios[†]. Sensors 21(4):1–22. <https://doi.org/10.3390/s21041494>

8. Differences Between 4G-LTE and 5G-NR Physical Layer—Long Term Evolution 4G. [Online]. Available: <https://ltebasics.wordpress.com/2022/01/22/differences-between-4g-lte-and-5g-nr-physical-layer/>
9. Yazar A, Peköz B, Arslan H (2018) Fundamentals of multi-numerology 5G new radio [Online]. Available: <http://arxiv.org/abs/1805.02842>
10. Zaidi AA, Baldemir R, Moles-Cases V, He N, Werner K, Cedergren A (2018) OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN. *IEEE Commun Stand Mag* 2(2):78–83. <https://doi.org/10.1109/mcomstd.2018.1700021>
11. de Valgas JF, Monserrat JF, Arslan H (2021) Flexible numerology in 5G NR: interference quantification and proper selection depending on the scenario. *Mob Inf Syst* 2021. <https://doi.org/10.1155/2021/6651326>
12. Dilli R (2020) Analysis of 5G wireless systems in FR1 and FR2 frequency bands. In: 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA), 2020, pp 767–772. <https://doi.org/10.1109/ICIMIA48430.2020.9074973>
13. Ömer N, Üniversitesi H, Fakültesi M, Mühendisliği EE (2020) Performance analysis of physical downlink shared channels for 5G new radio international turkic world congress on science and engineering performance analysis of physical downlink shared channels for 5G new radio Niğde Ömer Halisdemir Üniversitesi , Fen B. Int Turkic World Congr Sci Eng no. Aug 2019
14. Kabalci Y, Ali M (2020) Throughput analysis over 5G NR physical uplink shared channels. Proc—2020 IEEE 2nd Glob Power, Energy Commun Conf GPECOM 2020 (November):345–349. <https://doi.org/10.1109/GPECOM49333.2020.9247906>
15. Panda I, Ramanath S (2021) Analysis of beamforming in dense urban deployments. *Int Conf Commun Syst Netw (COMSNETS)* 2021:29–33. <https://doi.org/10.1109/COMSNETS51098.2021.9352933>
16. Peña, Alonso JI (2022) Design and synchronization procedures of a D&F co-operative 5G network based on SDR hardware interface: performance analysis. *Sensors* 22(3). <https://doi.org/10.3390/s22030913>
17. Takeda K, Xu H, Kim T, Schober K, Lin X (2020) Understanding the heart of the 5G air interface: an overview of physical downlink control channel for 5G new radio. *IEEE Commun Stand Mag* 4(3):22–29. <https://doi.org/10.1109/MCOMSTD.001.1900048>
18. Launay F (2021) 5G-NR radio interface—Radio access procedure. NG-RAN and 5G-NR, pp 237–261. <https://doi.org/10.1002/9781119851288.ch9>
19. Indoondon M, Pawan Fowdur T (2021) Overview of the challenges and solutions for 5G channel coding schemes. *J Inf Telecommun* 5(4):460–483. <https://doi.org/10.1080/24751839.2021.1954752>
20. Mediatek (2020) 5G NR and 4G LTE coexistence a comprehensive deployment guide to dynamic spectrum sharing, p 3
21. Specification T (2018) TS 138 212 - V15.2.0—5G; NR; multiplexing and channel coding (3GPP TS 38.212 version 15.2.0 Release 15). 0:100 [Online]. Available: https://www.etsi.org/dl-liver/etsi_ts/138200_138299/138212/15.02.00_60/ts_138212v150200p.pdf
22. DL-SCH and PDSCH Transmit and Receive Processing Chain—MATLAB & Simulink—MathWorks India [Online]. Available: <https://in.mathworks.com/help/5g/gs/dl-sch-and-pdsch-transmit-and-receive-processing-chain.html>

Automated Text Summarization Using Transformers



Yogesh Kumar, Ashish Jangir, Bhavya Meena, and Isha Pathak Tripathi

Abstract Text summarization is the process of creating a condensed form of text document which maintains significant information and general meaning of source text. Automatic text summarization becomes an important way of finding relevant information precisely in large text in a short time with little efforts. There are two main strategies involved in text summarization such as abstractive and extractive. In extractive method, the algorithm generates the summary by just picking up the words and line from the corpus. On the other hand in abstractive method, the algorithm generates the summary by rewriting the sentences. The main idea of this paper is to summarize text and know how transformers work in case of text summarization.

Keywords Abstractive · Extractive · Text summarization · Transformers · BERT · RNN

1 Introduction

In the modern Internet age, textual data is ever increasing. We need some ways to condense this data while preserving the information and meaning of the original data intact. From data on the Internet to news articles in newspapers to books in online libraries, a huge amount of useful data is available. This information is left unusable, unless we find a solution to make it available for users. Main idea is then to decrease the volume of the available data so it can be easily shared, read, and interpreted by users.

Y. Kumar (✉) · A. Jangir · B. Meena · I. P. Tripathi
Indian Institute of Information Technology, Kota, Rajasthan, India
e-mail: 2019kucp1108@iiitkota.ac.in

A. Jangir
e-mail: 2019kucp1109@iiitkota.ac.in

B. Meena
e-mail: 2019kucp1110@iiitkota.ac.in

I. P. Tripathi
e-mail: isha@iiitkota.ac.in

Text summarization is the process of decreasing the size of the corpus of the text to a few sentences which contains the most important information of the source content. The decreased size of the corpus of text makes it easy for the user to read the text and understand the context of text in less time. Time efficiency is the biggest advantage of text summarization with other advantages like clear, transparent, and precise summary as we remove all unnecessary and redundant data from text. Let us look at the two approaches which we are going to work with in this paper:

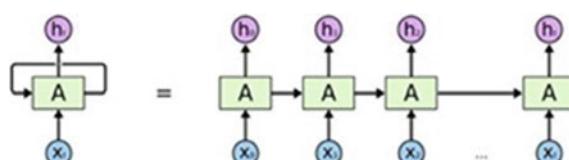
1.1 *Transformers*

Transformer networks can be pre-trained and are able to provide top performance in standard NLP benchmarks. Compared to the sequential models that had gone before, they deliver better results while making more efficient use of the available processing power. The transformer model introduced an “attention” mechanism that takes into account the relationship between all the words in the sentence. It creates differential weightings indicating which other elements in the sentence are most critical to the interpretation of a problem word. In this way, ambiguous elements can be resolved quickly and efficiently.

1.2 *RNN*

Recurrent neural network (RNN) is a type of neural network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required, and hence, there is a need to remember the previous words. RNN has a memory which remembers all information about what has been calculated. RNNs make use of their memory for processing a sequential form of input. RNN stores everything and, thus, is good at recollecting the previous data. But in RNN, all the inputs are dependent on one another. The RNN will not forget all the relations it previously encountered while training itself. For accomplishing this, the RNN forms many layers of interconnected system with loops in them, which allows it to remember the patterns (Fig. 1).

Fig. 1 RNN flowchart



Transformer is a sequence-to-sequence (S2S) architecture originally proposed for neural machine translation that rapidly replaces recurrent neural networks (RNNs) in natural language processing tasks. Previously, RNN-based methods were used for NLP models. This method is not very efficient as it has long-term dependencies. The model forgets the distant position content and mixes with the adjacent positions and makes it difficult to make decisions. This sequential process does not allow to take full advantage of fast computation.

Nowadays, we use a methodology known as attention mechanism. The attention mechanism works differently from RNN methods by considering the relation between the lexemes, irrespective of their placement position. A transformer performs small and constant steps (with the help of self-attention mechanism) by maintaining relation between all the words.

2 Related Work

Allahari et al. [1] emphasized various extractive approaches for single and multi-document summarization. They described some of the most extensively used methods such as topic representation approaches, frequency-driven methods, graph-based, and machine learning techniques. Although it was not feasible to explain all diverse algorithms and approaches comprehensively in that paper, they believe it provides a good insight into recent trends and progresses in automatic summarization methods and described the state-of-the-art in this research area.

Zolotareva et al. [2] dealt with the demanding task of abstractive document summarization. They have used the transformer framework, to create a multi-sentence summary. Experiments were carried out to verify the effectiveness of the proposed method. Experimental results performed well in the abstractive document summarization. And they stated, the future direction is to study the transformer method for the task of summarizing multiple documents.

Vaswani et al. [3] stated that the dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder and decoder configuration. They propose such models also connect the encoder and decoder through attention mechanisms and a simple network architecture based on an attention mechanism.

Nallapati et al. [4] model abstractive text summarization using attentional encoder-decoder recurrent neural networks. They proposed several models that address critical problems in summarization, such as modeling keywords, capturing the hierarchy, and emitting rare and stopwords.

Masum et al. [5] suggested a text summarization approach based on sequence-to-sequence RNN, which is basically abstractive summarization method that leverages the Amazon Fine Food Reviews dataset. Their approach includes a bidirectional RNN which have LSTMs in the encoding layer with an attention model in the decoding layer. Talukder et al. [6] discussed about abstractive text summarization. They compared various abstractive approaches in their paper.

3 Problem Statement

Text summarization is a process in which a large text is converted into a summarized form with lesser text, keeping the meaning similar to the original text, saving time and effort of reading the whole text. For example: If we have a sentence like “Nescafe is a good coffee brand and it is rich in taste”, then with the help of text summarization, it can simply be reduced to “Nescafe coffee is good in taste”. Here, we can see that the meaning of the obtained text is somewhat similar to the original text with lesser number of words. In a similar way, a larger paragraph can also be reduced by this technique.

With the exponential increase in the availability of data, the need to summarize text is imperative. With the ever increasing pace of life, the public no longer has the time to read long articles, so a tool is needed to develop that can effectively summarize such data. The application of summarization is valuable in domains like finance and retail, search engines, business analysis, and research as it helps save time and improve productivity.

Since a longtime RNNs have been used for text summarization purpose, but they are not too much efficient on large chunks of data due to the inability to handle long short-term dependencies. So we have tried to conduct a comparative study on abstractive text summarization by transformer method and RNN method on Amazon Fine Food Reviews dataset which is a collection of more than 500000 reviews and compare the results generated by both the models. The prime focus is to find which method is better over the other one and how closely the generated results resemble the referenced or human generated summary in order to get better insights into which method can be used to get better text summarization results.

4 Methodology

4.1 *Transformers*

The transformer model uses attention mechanism to boost the speed to beef-up the model training. There are various types of transformers available but we have used the BERT transformer model.

BERT refers to bidirectional encoder representations from transformers. It is a transformer used to overcome the limitations of RNN and other neural networks as long-term dependencies. It is a pre-trained model that is naturally bidirectional. This pre-trained model can be tuned to easily perform the NLP tasks as specified, as summarization in our case. It uses a powerful flat architecture with inter sentence transform layers so as to get the best results in summarization. The summary sentences are assumed to be representing the most important points of a document. BERT is able to express the semantics of a document and obtain representations for its sentences.

- Pre-processing: In this phase, we removed all unwanted symbols and cleaned the data. Then, tokenization is performed to convert word vector to numeric vector to help model to gain information about sequence of data.
- Dataset analysis and modulations: Created a custom dataset for reading the data frame and loading it into the data loader to pass it at a later stage for fine-tuning the model and to prepare it for predictions.
- Encoder: The encoder in transformer model is comprised of several self-attention layers. In the encoding component, there is a number of encoders that are stacked on top of each other. All the encoders are similar in structure. Each one is bifurcated into two sub-layers as follows: The self-attention layer and the feedforward neural network layer. Initially, the inputs of the encoder pass via the self-attention layer, which helps it to analyze other words in the input. The outputs from the self-attention layer are then transferred to the feedforward neural network layer of the encoder. The identical feedforward neural network is also autonomously applied to each and every position.
- Decoder: The decoding component is also comprised of a number of decoders. The number of decoders is the same as the number of encoders in the encoding component. The decoder has both the layers that are present in encoders, viz., the self-attention layer and the feedforward neural network layer, but in between them is an encoder-decoder attention layer that helps the decoder to focus on relevant parts of the input sentence (just like what attention does in seq2seq models). The encoder starts by processing the input sequence. The top encoder's output is then converted into a set of attention vectors. These are employed by each decoder in its “encoder-decoder attention” layer, which helps the decoder in focusing on the appropriate places in the input sequence. The self-attention layer in the decoder is only permitted to pay attention to earlier points in the output sequence. The “encoder-decoder attention” layer functions in the same way that multiheaded self-attention does.
- Fine-tuning BERT for Summarization It is conceivable that there is a mismatch between the encoder and the decoder, since the former is pre-trained so we fine-tuned the model on our dataset.

For fine-tuning, most model hyperparameters are the same as in pre-training, with the exception of the batch size, learning rate, and number of training epochs. The

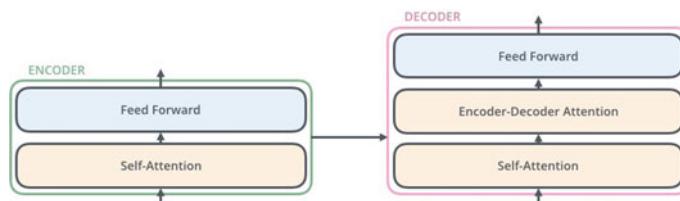


Fig. 2 Encoder-decoder model

optimal hyperparameter values are task-specific, but we perform with the following range of possible values to work across all tasks:

- Batch size = 64
- Epochs = 2
- Encoder length = 16
- Decoder length = 6
- Summary Generation: We have generated the summary using the encoder–decoder functions using pre-trained BERT model with fine-tuning on our dataset.
- ROUGE Score: Here, we have produced ROUGE scores for the summary generated and compared it with original summary in the dataset (Fig. 2).

Self-attention is one of the most important concepts to perform summarization task. It is as important to understand what self-attention is as it is to understand the problem statement. Let us look in detail what self-attention is and how it works.

4.2 *Self-attention in Detail*

First of all, we create three vectors from each of the encoder’s input vectors. For every word in the sentence, we create a query vector, a Key vector, and a value vector. All these vectors created are smaller in dimension than the embedding vector.

Calculating a score is the second stage in calculating self-attention. Let us assume, we are working out the self-attention for the first word. Each word in the input sentence must be compared to the first word. For that, we need to score every word of the sentence against the first word. As we encode a word at a specific position, the score directs how much emphasis to place on other portions of the input sentence.

This score is obtained by the dot product of the query vector and the key vector(word being scored). The dot product of q_1 and k_1 would be the first score if we were processing self-attention for the word in position 1. The dot product of q_1 and k_2 would be the second score.

In the third step, the scores are divided by the square root of the key vectors’ dimension. This is done to produce more stable gradients.

In the fourth step, the result is passed via a softmax operation. The softmax does the normalization of the scores, so they all lie between 0 and 1 and add up to 1. The softmax score decides how much each word will be expressed at the particular position.

Then apply a softmax operation to the result. All the scores are then normalized and added up to 1 by softmax. This position’s softmax score decides how much each word will be expressed. The softmax score is multiplied by each value vector in the fifth phase.

The final step is to add the weighted value vectors together. The output of the self-attention layer is created as a result of this. We may now transmit the generated vector to the feedforward neural network.

4.3 Calculation of Self-attention

To calculate self-attention, query, key, and value matrices are required, so first, we have to calculate these, and this can be done by assuming embeddings into matrix X and multiplying it by the weight matrices.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

Mathematically, attention can be represented as shown above in Eq. (1). After getting all the required attributes, the attention is obtained by its mathematical calculations using the above equation.

4.4 RNN

To implement the RNN model, we have used the LSTM-based neural network to enhance the efficiency of generated summaries. Long short-term memory networks are a special kind of RNN, capable of learning long-term dependencies. LSTM is designed so that it has an activation state that can act like weights and preserves the information over long distances, hence the name long short-term memory. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn. All recurrent neural networks have the form of a chain of repeating modules of neural network.

LSTM uses the encoder–decoder architecture that are defined as follows:

4.4.1 Encoder

An encoder LSTM model reads the entire input sequence, wherein, at each time-step, one word is fed into the encoder. It then processes the information at every time-step and captures the contextual information present in the input sequence.

4.4.2 Decoder

The decoder is also an LSTM network which reads the entire target sequence word-by-word and predicts the same sequence offset by one time-step. The decoder is trained to predict the next word in the sequence given the previous word.

- Pre-processing: This is the process of deleting extraneous words, symbols, and tags from provided data that may be any like a document or a article from Wikipedia.

There may be several steps in this procedure, tokenization, and stopword detection which are part of pre-processing.

- Tokenisation: The entire text is transformed into a continuous stream of words.
- Stopword detection: Identifying and removing words with no useful information, for example: a, the, is, are, etc.
- Analyzing Dataset: Here, we analyze the length of the original text and the summary to get an overall idea about the distribution of length of the text. This helps to fix the maximum length of the sequence.
- Preparing the Tokenizer: A tokenizer builds the vocabulary and converts the word sequence to an integer sequence. We have used tokenizer from preprocessing module of the keras library.
We also take a look at the proportion of rare words and its total coverage in the original text and summary to decide length for summary generation.
- Model Building: We are finally at the model building part; here, we have performed a 3 stacked LSTM for the encoder. This leads to a better representation of the sequence. We have used sparse categorical cross-entropy as a loss function since it converts the integer sequence to a one-hot vector on the fly to overcome any memory issues.
- Model Training: The following parameters were selected by taking into account the computation power and resources at hand. Therefore, we selected the hyper-parameters using the manual configuration method. The dataset is split into 90% training data and 10% testing data. The concept of early stopping is very important. It is used to stop training the neural network at the right time by monitoring a user-specified metric. Our model stops training once the validation loss increases. This is done using the train _test _split with

- Epochs = 20
- Batch size = 128
- Latent dimension = 300
- Embedding dimension = 100
- Optimizer = “RMSprop”
- Loss function = “sparse_categorical_crossentropy”
- Attention Mechanism: The key intuition behind the attention mechanism concept is how much attention do we need to pay to every word in the input sequence for generating a word at a time-step t . So, instead of looking at all the words in the source sequence, we can increase the importance of specific parts of the source sequence that results in the target sequence.
- Softmax Layer: The final linear layer followed by a softmax Layer performs its job through the decoder, as decoder stack outputs a vector of floats. Next, the softmax of the scaled score is taken to get the attention weights, which gives the probability values between 0 and 1. By doing a softmax, the higher scores get heightened, and lower scores are depressed. This allows the model to be more confident about which words to attend too.

- **Summary Generation:** Using LSTM, we defined the functions to convert an integer sequence to a word sequence for summary and generated the summary for dataset.
- **ROUGE Scores:** Here, we have produced ROUGE scores for the summary generated and compared it with original summary in the dataset.

5 Experiments and Results

5.1 Dataset

The experiment was carried out on Amazon Fine Food Reviews Dataset provided by Kaggle. The dataset consists of around 500,000 reviews from Amazon. This includes nine class labels which are ProductId, UserId, ProfileName, HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary, Text. We have taken into consideration only two parameters which are summary and text.

5.1.1 Fine-Tuning

We have fine-tuned the pre-trained transformer model on the same dataset to generate summaries.

5.2 Evaluation Metrics

We used ROUGE metrics for our evaluation process. ROUGE refers to recall-oriented understudy for gisting evaluation. This is an automatic summary evaluation benchmarking metric used to determine the quality of summary produced by comparing the machine generated summary with the reference summary. ROUGE scores are computed from the number of overlapping words between the reference summary and machine generated summary, and its mathematical representation is shown in Eq. (2).

$$P = \frac{\text{Number of overlapping words}}{\text{Total words in reference summary}} \quad (2)$$

5.2.1 ROUGE-N

This measures the number of matching n-grams between our model-generated text and a reference. The N represents the n-gram that we are using. For ROUGE-1, we would be measuring the match-rate of unigrams between our model output and reference.

5.2.2 ROUGE-L

This measures the longest common subsequence (LCS) between our model output and reference. All this means is that we count the longest sequence of tokens that is shared between both.

5.2.3 ROUGE-S

This allows us to add a degree of leniency to our n-gram matching. The skip-gram metric allows us to search for consecutive words from the reference text that appear in the model output but are separated by one-or-more other words.

5.3 Results

The experimental results of text summarization done using transformer and LSTM-based RNN model are presented in Figs. 3, 4, 5 and 6.

The summaries generated by the transformer method are quite closer to the human generated summaries unlike the ones generated by RNN method. Both the methods are applied to nearly 500,000 reviews, and the results obtained by the transformer method are far good in terms of both recall and precision resulting in good F1-scores as well.

```
[ 'Classic and classy hot sauce', ['very tasty!',  
  'Five Stars', 'five stars',  
  'Five Stars', 'five stars',  
  'One Star', 'three stars',  
  'Peppered with Praise', 'great flavor and flavor',  
  'Great little hard candies!'] ] 'five stars'] ]
```

Fig. 3 Summaries generated by transformer method

```
Review: auto ship terrific way get hard find dog  
tive wrestle co footprint shipped  
Original summary: awesome product  
Predicted summary: great product
```

```
Review: far taste buds concerned perfect cup coi  
wimpy macho thanks wolfgang love  
Original summary: perfect cup of coffee  
Predicted summary: best coffee ever
```

Fig. 4 Summaries generated by RNN method

Fig. 5 Results of transformer method

	Rouge	f	p	r
0	Rouge 1	0.461607	0.484375	0.461805
1	Rouge 2	0.4375	0.4375	0.4375
2	Rouge I	0.461607	0.484375	0.461805

Fig. 6 Results of RNN method

	Rouge	f	p	r
0	Rouge 1	0.216487	0.234169	0.220418
1	Rouge 2	0.101648	0.095162	0.110700
2	Rouge I	0.246894	0.253975	0.240963

The result is represented with the help of ROUGE scores which includes three evaluating parameters that are r, p and f. Here, “r” stands for “recall”, “p” stands for “precision”, and “f” stands for “F1-score”.

The recall basically counts the number of intersecting n-grams from the model-generated summary and the human summary and is divided by the no. of n-grams in the human summary. While in precision, we divide the same by the count of n-grams in the model-generated summary. The F1-score gives a better understanding of how well the model actually works by taking into consideration both recall and precision as recall alone cannot produce desired results most of the times.

The results shown in Fig. 5 are of transformer method, and the results in Fig. 6 are of RNN method.

According to the experimental results presented, the transformer-based abstractive text summarization outperformed the RNN method.

6 Conclusion

In this study, we come to the conclusion that the transformer model performs better than the RNN model in summarizing text efficiently. The transformer model predicts meaningful and clear context. The ROUGE scores show that the summary generated by transformer model is quite closer to the main idea of the input data as compared to the RNN model. Time consumption is a big issue in RNN method as compared to the transformer method. When compared to a set of human summaries, the transformer-based method had a greater percentage level of match. As a result, we can infer that a transformer-based system is superior to an RNN-based system in terms of offering a better comprehension and summary in a shorter amount of time.

References

1. Allahari M, Safae S, Pouriyeh S, Assefi M (2017) Text summarization techniques : a brief survey. *Int J Adv Comput Sci Appl* <https://doi.org/10.48550/arXiv.1707.02268>
2. Zolotareva E, Zolotareva E, Horvath T (2020) Abstractive text summarization using transfer learning. In: 20th conference information technologies applications and theory. *ceur-ws.org/Vol-2718/paper28*
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need, NIPS. <https://doi.org/10.48550/arXiv.1706.03762>
4. Nallapati R, Zhou B, dos santos CN, Gulcehre C, Xiang B (2016) Abstractive text summarization using sequence-to-sequence RNNs. In: Proceedings of The 20th SIGNLL conference on computational natural language learning. <https://doi.org/10.48550/arXiv.1602.06023>
5. Masum AKM, Abujar S, Talukder MA (2019) Abstractive method of text summarization with sequence to sequence RNNs. <https://doi.org/10.1109/ICCCNT45670.2019.8944620>
6. Talukder MAI, Abujar S, Masum AKM, Akter S, Hossain SA (2020) Comparative study on abstractive text summarization. <https://doi.org/10.1109/ICCCNT49239.2020.9225657>

Minimizing Building Energy Waste by Detecting and Addressing HVAC Issues



Anshul Agarwal 

Abstract Every building is equipped with heating, ventilation and air conditioning (HVAC) systems to ensure user comfort. These are acknowledged to be the primary factors to the structures' excessive energy use. Inefficient operation of these HVAC systems frequently results in a significant amount of energy loss. Therefore, it is essential that these always perform effectively. To address this problem, the research focuses on a novel framework for detecting HVAC problems in buildings, such that energy consumption waste due to HVAC defects may be reduced. Existing techniques suffer from challenges such as the deployment of expensive and specialized gear, manual examination of HVAC systems and the use of vast volumes of training data for fault detection. However, the proposed framework overcomes these obstacles and is applied to a building's forty-eight HVACs (situated in Mumbai, India). A total of thirty-one HVAC systems were determined to be defective. It was noticed that the malfunctioning of these HVACs led to a 48% increase in energy usage.

Keywords Energy savings · Fault detection · Fault identification · HVACs · Air conditioners · Smart buildings

1 Introduction

Buildings contribute significantly to the increased energy consumption. Consequently, there is a greater demand than ever for intelligent buildings, given that one of their key objectives is energy efficiency. It is a challenge to satisfy these rising demands in developing countries like India [1]. The building management systems construct a sensor infrastructure that monitors the status and energy usage of various equipment [2, 3]. Inefficient and defective appliance operation typically uses large quantities of energy, resulting in substantial energy waste. Manually detecting and identifying defective equipment may be arduous and often impossible. HVAC

A. Agarwal (✉)

Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, Maharashtra 440010, India

e-mail: anshulagarwal@cse.vnit.ac.in

systems generate around 50% of building energy usage [4–6]. A series of investigations on around 13,000 air conditioners revealed that over 65 per cent were defective [7]. This needs the development of a pragmatic and efficient method for HVAC problem detection and diagnosis, so that faulty components may be fixed or replaced to prevent energy waste.

There exist several ways for HVAC fault detection. However, many of these strategies include the use of cumbersome or specialized sensor equipment [8–10], which raises the price and complicates matters. Many methods rely exclusively on model-based methods [11–14]. They demand particular understanding of the inner workings of normal HVAC operation. However, in real-world settings, it is extremely complicated and not always possible [8]. The remaining methods employ data-driven models [15–17], which need no HVAC-specific domain expertise. However, the fundamental limitation of these methods is that they require an abundance of high-quality sensor data regarding HVAC operation, which is not always accessible. These challenges were the impetus for developing a unique framework to detect malfunctioning HVACs for minimizing energy waste in smart buildings without the need of extra hardware.

This paper's main contribution is the creation and use of a unique framework with the following benefits: (a) applicable to currently existing HVACs; (b) utilizing just temperature and power consumption data; (c) requiring no additional sensing instrumentation and (d) capable of updating with fresh information and problem diagnosis. In order to evaluate the efficacy of the created framework, it is applied to a real-world situation with 48 HVAC systems placed in a building in Mumbai, India.

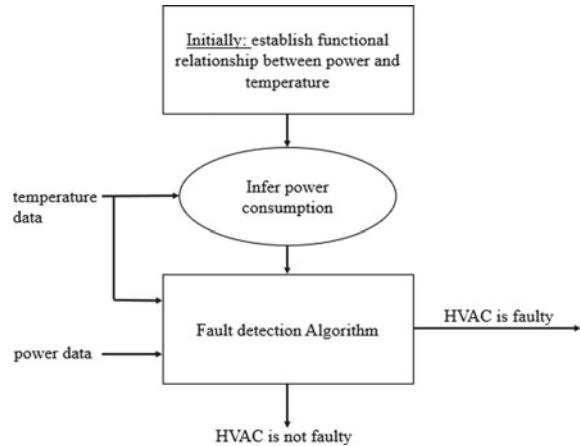
2 Methodology

The primary idea behind the developed framework is the intelligent use of functional relationship between the facets that are affected by the functioning of HVACs. Temperature and power consumption are the primary facets that immediately reflect the change in working of HVAC operation. For example, if an air conditioner is switched ON, a corresponding change (rise) in power consumption and change (drop) in temperature of the area around the air conditioner will be observed. Using these kinds of functional relationship in an intelligent way has a significant advantage since these types of data is generally present in a smart building, thereby, eliminating the need for any additional sensing and instrumentation. The summary of the research methodology adopted in this paper is presented in Fig. 1.

2.1 Initialization

The following are the functional links between temperature and power consumption numbers for a well-functioning HVAC system.

Fig. 1 Flow diagram summarizing the research methodology



- When a strong negative gradient of temperature values is detected (i.e., when the temperature is falling rapidly), a commensurate increase in energy consumption will be noticed. The rise has a certain fixed value for the HVAC i and is recorded as Δ_i^{rise} . This relationship is represented by Eq. 1.

$$((T_t - T_{t-1}) < 0) \Rightarrow (P_t = P_{t-1} + \Delta_i^{\text{rise}}) \quad (1)$$

where T_t and P_t denote temperature and power, respectively, at time t . This indicates that the HVAC system has begun its work, such as cooling the area. During the period of time $(t_1 \dots t_n)$ when the temperature is falling fast, the increase in energy consumption is noted just once.

- When a steep positive gradient of the temperature values of HVAC i is detected (i.e., a sudden increase in temperature), a decrease in power usage for HVAC i is noted. This decrease in HVAC i power consumption has a set value and is reported as Δ_i^{drop} . This relationship is specified in Eq. 2.

$$((T_t - T_{t-1}) > 0) \Rightarrow (P_t = P_{t-1} - \Delta_i^{\text{drop}}) \quad (2)$$

This indicates that the HVAC system has stopped operating, such as by turning OFF or ceasing to cool the room. During the time $(t_1 \dots t_k)$ while the temperature is steadily increasing, the power consumption decrease is noticed just once.

- When the temperature of HVAC i falls below Llimit for the first time, HVAC i is turned on and the power consumption is calculated using Eq. 3.

$$(T_t < Llimit) \Rightarrow (P_t = P_{t-1} + \Delta_i^{\text{rise}}) \quad (3)$$

The power consumption will remain undisturbed if the HVAC was already switched ON; else it will rise by Δ_i^{rise} amount. This scenario indicates that the

HVAC system will operate as long as the temperature remains below L_{limit} . In a similar fashion, when the temperature of HVAC i exceeds U_{limit} for the first time, HVAC i is turned off and the power consumption is inferred.

- When the temperature is seen to be stable within the range $[L_{\text{limit}}, U_{\text{limit}}]$ for HVAC i the related power consumption figures are likewise stable. Temperature readings that are constant imply that the difference between successive temperature values is negligible, or less than. Therefore, with stable temperature values recorded for HVAC i , the power consumption may be deduced using Eq. 5.

$$((T_t - T_{t-1}) \leq \delta) \text{ AND } (L_{\text{limit}} \leq T_t \leq U_{\text{limit}}) \Rightarrow (P_t = P_{t-1}) \quad (5)$$

2.2 Algorithm for Fault Detection

The temperature and power consumption statistics of HVAC i (whose fault state is to be identified) are inputs for the fault detection algorithm FD. Using the calculated functional relationships, the temperature measurements are utilized to infer the HVAC's energy usage. If the difference between the inferred and actual power consumption is more than δ^P , the HVAC system is labelled as defective; otherwise, it is working normally.

Algorithm FD: Fault detection

Input:

- (a) $\{T\}$: Temperature time series data of HVAC i
 - (b) $\{P\}$: Actual power consumption time series data of HVAC i
-

Output: fault status of HVAC i

1. If the temperature of HVAC i is consistently below L_{limit} , then
 - HVAC i is defective
 - Remove this HVAC unit's baseline power consumption P_t from the overall power consumption under consideration
 2. Collect temperature and power usage data from HVAC units that are not defective
 3. Determine the functional connections described in the preceding section
 4. Using the given functional relationships, calculate the power consumption (P'_t) at time t for HVAC system i from the temperature data $\{T\}$
 5. If $|P'_t - P_t| < \delta^P$ then
 - Print "HVAC is not faulty"
 6. Else
 - Print "HVAC is faulty"
-

The steps of the algorithm are described in full below.

- Step 1 specifies that the relevant temperature values for HVACs that operate continuously and are never turned off are always less than L_{limit} . This is the behaviour of a defective HVAC system. In order to determine the state of other

HVACs, the power consumption associated with the malfunctioning HVAC is subtracted from the total power consumption.

- The functional connections between temperature and power consumption data that characterize the operation of a non-defective HVAC system of type i are derived in Steps 2 and 3.
- Step 4: Using these functional connections, infer the power consumption of HVAC system i based on the temperature data T supplied as input.
- Step 5: If the inferred power consumption P'_t and the actual power consumption P_t at time t (provided as input) are near to one another (i.e. the difference is smaller than p), then the HVAC i is labelled as non-faulty.
- Step 6: If the difference is greater, it indicates that the HVAC's power consumption and temperature do not display the characteristics of a non-faulty HVAC; hence, HVAC i is labelled as defective.

If algorithm FD identifies the HVAC as defective, the next step is to inform the maintenance company responsible for AC repair.

3 Results and Discussion

For testing the developed framework, initially, a set of four HVACs (a, b, c and d) are considered. Their aggregate power consumption and temperature of individual HVACs (denotes as T_a, T_b, T_c and T_d , respectively) over a duration of more than 25 min are shown in Fig. 2.

3.1 Fault Detection

For initialization, the temperature and power consumption behaviour, of a non-faulty HVAC of the given type i considered in the application area, are shown in Fig. 3.

It can be observed from this figure that the value of different variables is initialized as follows: $\Delta^{\text{rise}} = 4.6 \text{ kW}$, $\Delta^{\text{drop}} = 4.6 \text{ kW}$, $L\text{limit} = 11^\circ\text{C}$ and $U\text{limit} = 21^\circ\text{C}$.

Application of Algorithm FD

- Fig. 3 demonstrates that the temperature T of HVAC b is always below the $L\text{limit}$ value of 11°C . Consequently, the algorithm marks HVAC b as defective. The approximately 6200 W power usage of this HVAC is subtracted from the total power consumption of all HVACs. This is illustrated by Fig. 4.
- The power consumption of HVACs a, c and d is deduced using the functional relationships found for the HVAC without a malfunction. This is seen in Fig. 4. This graph reveals that for HVAC c , the temperature (denoted by T_c) lowers drastically fourfold. Consequently, a power consumption increase of $\text{rise} = 4.6 \text{ kW}$

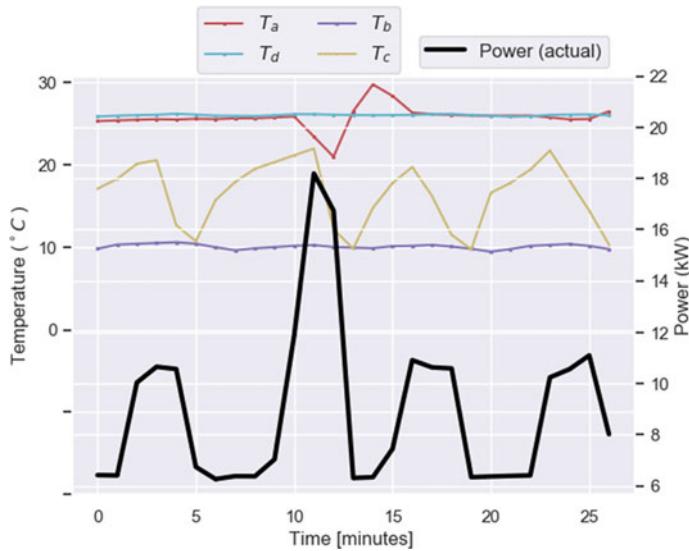


Fig. 2 Graph of aggregate power consumption and temperature of individual HVACs

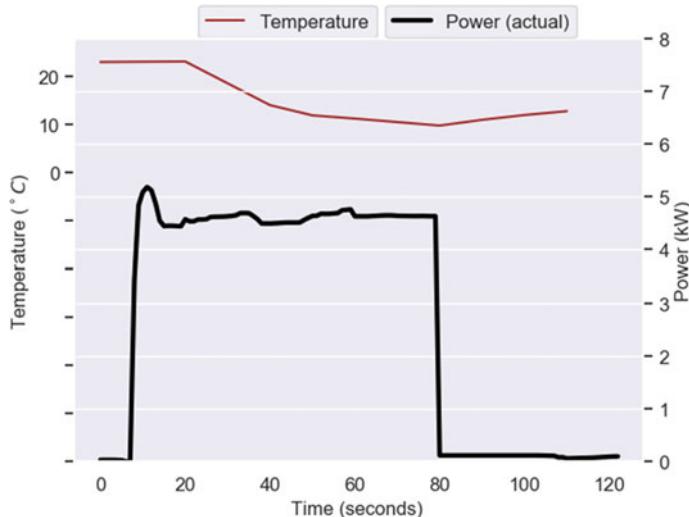


Fig. 3 Temperature and power consumption behaviour of the non-faulty HVAC i

is estimated when the temperature lowers significantly. For HVAC a , T_a temperature drops dramatically just once (after about ten minutes). Thus, the programme infers a current power usage of 4.6 kW. Since at this moment, the temperature corresponding to both HVAC systems drops significantly, a significant increase in power usage of around 8.8 kW may be noted. For HVAC d , the temperature

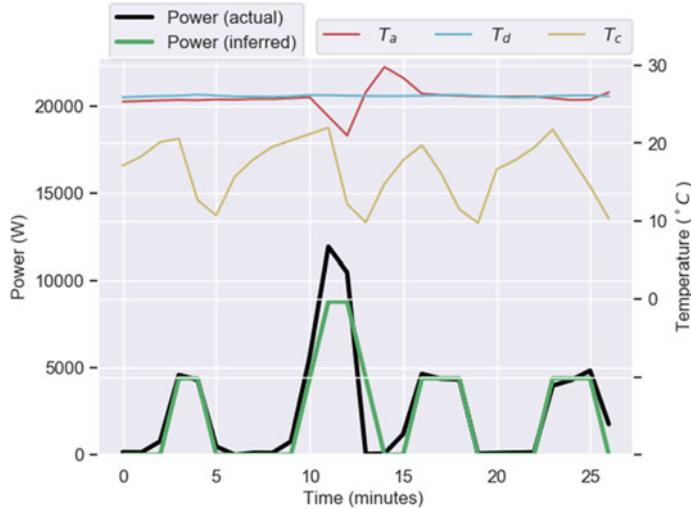


Fig. 4 Inferred and actual power consumption of the HVACs

is always above U_{limit} , which indicates that the HVAC power does not increase. This indicates that HVAC d is also defective, since the HVAC is turned on but cannot cool or heat the surrounding area.

- The root mean square (RMS) difference between the actual power consumption and inferred power consumption for HVAC a is about 2.5 kW. Similarly, the RMS difference for HVAC c is about 550 W. For δ^p as 15% of Δ^{rise} , the difference in inferred and actual power consumption of HVAC a is very high and is thus, labelled as faulty. Since the difference for HVAC c is less than δ^p , therefore it is labelled as non-faulty.

3.2 Energy Savings

As indicated, three of the four HVAC systems under evaluation are defective. A comparison is made between the hourly energy usage of these four HVACs (three faulty + one non-faulty) and that of four HVACs that are not faulty. Table 1 displays energy usage reductions.

According to Table 1, the total daily energy usage for malfunctioning HVAC units is 2146 kWh. When all HVAC units are functional, the daily energy consumption is only 1475 kWh, which is approximately 45.5% (671 kWh) less than when three HVAC units are faulty and one HVAC unit is problematic throughout a 24 h period.

Additionally, the built framework was evaluated on an application area of 48 air conditioners. There were a total of thirty-one defective air conditioners detected. The energy consumption of forty-eight non-faulty HVAC units is compared to the energy

Table 1 Per hour energy consumption of faulty and non-faulty HVACs for a day

Hour of the day	Total energy consumption (kWh)		Energy savings	
	All HVACs are functioning	Number of malfunctioning HVACs is one and other three are functioning	Absolute (kWh)	Percentage
0	59.07	93.36	34.29	58.05
1	61.89	93.04	31.15	50.33
2	64.51	88.12	23.61	36.6
3	59.15	90.74	31.59	53.41
4	61.68	89.3	27.62	44.78
5	58.58	94.73	36.15	61.71
6	60.81	92.23	31.42	51.67
7	64.07	91.31	27.24	42.52
8	63.49	87.27	23.78	37.45
9	59.33	88.67	29.34	49.45
10	66.38	90.63	24.25	36.53
11	61.03	91.38	30.35	49.73
12	62.52	90.2	27.68	44.27
13	66.63	89.63	23	34.52
14	62.01	88	25.99	41.91
15	64.09	89.27	25.18	39.29
16	60.55	88.62	28.07	46.36
17	58.67	92.09	33.42	56.96
18	62.91	86	23.09	36.7
19	63.18	88.54	25.36	40.14
20	60.9	88.04	27.14	44.56
21	58.96	91.03	32.07	54.39
22	64.02	91.45	27.43	42.85
23	63.77	85.21	21.44	33.62
Total	1488.2	2158.86	670.66	45.325

consumption of thirty-one faulty and seventeen non-faulty HVAC units. This is seen in Fig. 5.

Figure 5 demonstrates that when thirty-one air conditioners are defective, the overall monthly energy consumption is approximately 651 million Wh, compared to around 437 million Wh when all air conditioners are functional. Therefore, improper conduct results in an additional energy consumption of around 213 million Wh, or approximately 48% higher. Using the created framework, it is possible to conserve the 48% of energy that is lost due to HVAC malfunctions.

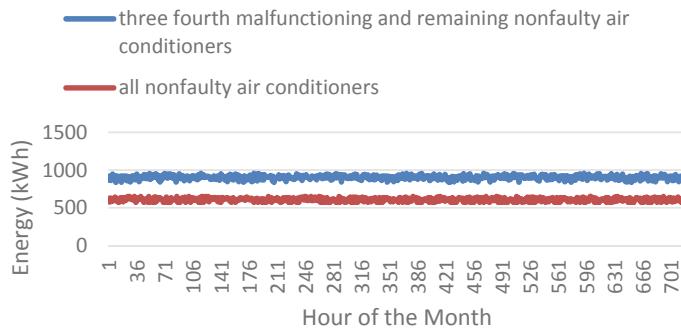


Fig. 5 Energy consumption by malfunctioning and functioning HVACs for one month

4 Conclusion

In a building, it is crucial that all HVAC systems operate at peak efficiency, since they are the equipment that consumes the most energy and whose malfunction results in significant energy waste. To address this issue, a new framework for detecting defective HVAC systems has been devised. This framework is applied to an existing building with forty-eight HVAC units, of which thirty-one were found to be defective. Comparing the energy usage of these HVACs to that of forty-eight HVACs without defective behaviour revealed that about 48% of energy is lost owing to improper operation. Through the deployment of this newly established framework, it is possible to quickly reduce energy waste by 48%. Thus, it can be stated that the developed framework is applicable and should be implemented in buildings in order to reduce energy waste.

References

1. International Energy Agency (2020) India 2020 energy policy review. https://webstore.iea.org/download/direct/2933?fileName=India_2020-Policy_Energy_Review.pdf
2. Agarwal A, Ramamritham K (2021) A novel approach for deploying minimum sensors in smart buildings. ACM Trans Cyber-Phys Syst 6(1):29, Article 2. <https://doi.org/10.1145/3477929>
3. Agarwal A, Ramamritham K (2020) Sensor minimization for energy management in smart buildings. In: 2020 IEEE first international conference on smart technologies for power, energy and control (STPEC), pp 1–6. <https://doi.org/10.1109/STPEC49749.2020.9297755>
4. Al-Turjman F, Altrjman C, Din S, Paul A (2019) Energy monitoring in IoT-based ad hoc networks: An overview. Comput Electr Eng 76:133–142
5. Mason K, Grijalva S (2019) A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 78:300–312
6. Perez-Lombard L, Ortiz J, Pout C (2008) A review on buildings energy consumption information. Energy Buildings 40(3):394–398
7. Downey T, Proctor J (2002) What can 13,000 air conditioners tell us. In: Proceedings of the 2002 ACEEE summer study on energy efficiency in buildings, vol 1, pp 53–67

8. Beghi A, Brignoli R, Cecchinato L, Menegazzo G, Rampazzo M, Simmini F (2016) Data-driven fault detection and diagnosis for HVAC water chillers. *Control Eng Pract* 53:79–91
9. Gao D, Wang S, Shan K, Yan C (2016) A system-level fault detection and diagnosis method for low delta-T syndrome in the complex HVAC systems. *Appl Energy* 164:1028–1038
10. Katipamula S, Brambley MR (2005) Review article: methods for fault detection, diagnostics, and prognostics for building systems a review, part I. *HVAC&R Research* 11(1):3–25
11. Frank PM (1990) Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: a survey and some new results. *Automatica* 26(3):459–474
12. Isermann R (1984) Process fault detection based on modeling and estimation methods a survey. *Automatica* 20(4):387–404
13. Li S, Wen J (2014) A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. *Energy Buildings* 68:63–71
14. Karmakar G, Arote U, Agarwal A, Ramamritham K (2018) Adaptive hybrid approaches to thermal modeling of building. In: Proceedings of the ninth international conference on future energy systems (e-Energy '18), ACM, New York, USA, pp 477–479
15. Bonvini M, Sohn MD, Granderson J, Wetter M, Ann Piette M (2014) Robust on-line fault detection diagnosis for HVAC components based on nonlinear state estimation techniques. *Appl Energy* 124:156–166
16. Agarwal A (2022) A novel approach to save energy by detecting faulty HVACs. *J Inst Eng India Ser B* 103:305–311. <https://doi.org/10.1007/s40031-021-00666-7>
17. Dey D, Dong B (2016) A probabilistic approach to diagnose faults of air handling units in buildings. *Energy Buildings* 130:177–187

Knowledge Representation and Information Retrieval from Ontologies



Azra Bashir, Renuka Nagpal, Deepti Mehrotra, and Manju Bala

Abstract The Ontology is an emerging domain that assists in intelligent decision making and connects various users, thereby facilitating a way of presenting the information on a common platform that could be used by the users for decision making. The Ontology leads to structuring of the unstructured data and retrieval of information from the system. The Ontology allows to share and reuse the data and its related concepts in a homogenized manner so as to provide a single nomenclature to the unstructured data that has been gathered from different sources and paving a way for elucidating the identical data easily. As the times are progressing to the fourth business revolution, absolute use of the artificial intelligence authorized ontologies leads to the decision support in many real-time applications. There are many methodologies by which an Ontology can be designed and the data can be structured and retrieved from the Ontology. This research paper aims to present a survey of the existing literature present in the domain of Ontology that would help in understanding the knowledge representation and information retrieval from the Ontologies created using various techniques.

Keywords Ontology · Semantic web · OWL · RDF · SPARQL

1 Introduction

In this piece of work, we present a review of recent published research on Ontologies created in different domains. A detailed review of the techniques and methods used for representing the knowledge into the Ontologies and the retrieval of information from the same is being presented in the paper. Semantic web is an integral component of an Ontology.

A. Bashir (✉) · R. Nagpal · D. Mehrotra
Amity University, Uttar Pradesh, Noida, India
e-mail: zarabashir93@gmail.com

M. Bala
Indraprastha College for Women, Delhi University, New Delhi, India

Semantic web: The word semantic [1] introduces a series of symbols that are used to impart denotation and this interaction can then affect the attitude in different ways.

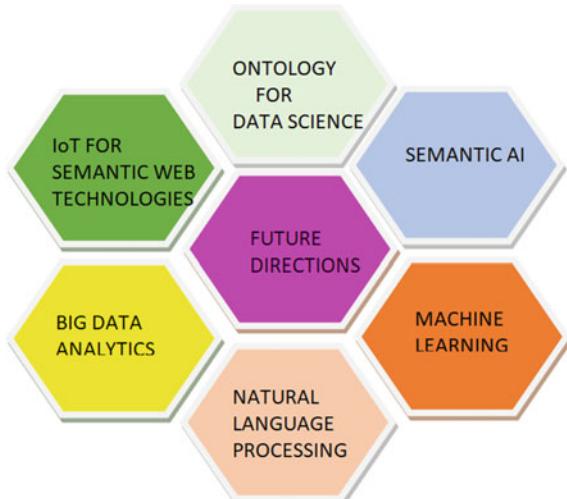
It has led to the generation next web known as semantic web. The web is focussed upon its role for automation of approaches that could exploit resources over the web. The semantic web allows construction of a system followed by a logic that is made by considering the Ontology structure. Sharing of semantic information is done across the automated systems in semantic web, and the same is partially conceptualized by an Ontology.

Its development has been dealing enormous number of users and the content in order to ensure trust at every step/stage in order to elevate the efficiency and capability of machines for the better understanding of the available existing information which is not currently the scenario over web because of a deficient universal format over the web. It refers to a web capable of processing information both for machines and for humans so that data can be interpreted by a machine, exchanging of information over web independently without any human interruption and production of data that is more relevant. It focusses on bringing a change in web development so that a machine is sensible enough to catch the words that are presented on a web page, relation between those words, ultimately producing more relevant information. It is focussed on describing things in a manner, easily understandable by different applications and web services. Few future aspects and research areas that rely on the semantic web are shown in Fig. 1.

Ontology: An Ontology refers to a descriptive or explanatory framework of a system, used for representing the entities and their relationships [2].

It serves as the semantic basis for the subjects (man, software machine, etc.) to inter communicate within a system. It forms the basis for sharing knowledge. It essentially provides means for machines to understand which lack the ability to comprehend

Fig. 1 Future directions for semantic web and Ontology



the semantics of human language, as the computer is only capable of dealing with text in the form of string data. Thus, Ontology represents a device comprehendible, meticulously specified knowledge that aids in describing the meta knowledge. It is contemplated as a potent aid in handling the knowledge.

Ontology is used for the representation and recognition of the jargon in a specified domain and clearly define the terms and their associations at various stages of a model for instance, for defining terms and the relationships among them in various medical fields. Ontologies in clinical field have been formed for supporting different areas in medicine. The examples of medical ontologies include systemic nomenclature of medicine (SNOMED), unified medical language source (UMLS) [3], etc. [4] Ontologies assist in the retrieval of information for providing query expansion terms which is defined as the inclusion of further relatable phrases and terms to the already existing list of words in retrieving by the application of computer semantics and the other technologies for generating the novel and more precise query terms, used in retrieving the information, enhancing the precision of retrieving the information and resolving any potential mismatch among the terms and thus furnishing adequate information for the user.

The Ontology once designed should be capable of decision support and should be easily accessible by user. Various methods are identified in which the knowledge is represented, i.e. stored in an Ontology and retrieval of information from the same.

2 Related Work

Several studies were conducted in various domains, and Ontologies for different sections were designed like in the field of medicine, healthcare, business, education, marketing, etc. Those studies carried out the process of designing Ontology in different ways, the knowledge representation and information retrieval was done in various ways by each one. According to the study by Munir [5], various ways and methods can be followed for retrieval of information from an Ontology by considering its outlook to model, process and translate the knowledge into the search requests for a given database. Various strategies for evaluation of Ontologies published till 2017 have been compiled by Degbelo [6] who classified them according to the similarity in their design of the way those Ontologies have been implemented. An Ontology for research paper selection has also been proposed by Balamurugan and Iyswarya [7] in which they categorized the concept terms in various fields of research and described a relation between them.

Various languages have been used while developing an Ontology, common being used is based on XML that makes it machine interpretable. Few examples are RDF and RDF Schema and the Ontology web language. OWL consists of other alternate languages such as OWL DL, OWL Lite and OWL Full. Few rules are added to OWL DL that forms semantic rule language (SWRL). The selection among those could

be made depending upon the type of task and the structure of the Ontology to be formed. A deep insight of the security of web services by using OWL frameworks is discussed by Pascal [8].

Rubin [9] et al. designed an Ontology using Protégé OWL providing a lot of features making it useful for designing Ontologies in OWL. It also supported design of intelligent applications using those ontologies. An overview of the research work for the usage of ontologies to access heterogeneous and incomplete data is given and described by Schneider et al. [10] (Sep 2020) [11] (Dec 2020).

Various researches have discussed the evolution of Ontology, one being Alqawasmeh [12] who discussed the evolution of Ontology and identified the loopholes Ontology networks. Another being Abdel-Qader et al. [13] who discussed the change in term and how those changes could be adopted while the Ontologies are evolving. The work further discussed the addition and deletion of those terms from the Ontologies selected from LOV.

Malviya et al. [14] designed an Ontology using Protégé 4.1 beta versions for designing an Ontology for RGPV University Bhopal. The Ontologies were extended to fuzzy approaches using Protege plug in. A fuzzy Ontology was created by Bobillo and Straccia [15] using Protégé which consisted of a core Ontology developed by using an Ontology editor that supported OWL2. A GUI was developed that made encoding of the annotation features visible to the user.

Middleware was also introduced for searching from the Ontologies. An OWL search Middleware [16] was designed for demonstration of various functions of the middleware that is capable of searching various ontologies at similar time. The search supported complex queries and performed well without having in depth knowledge of the structure of the Ontology and the querying language, allowing the users to search by just inputting the keywords for their search. A graphical Ontology web language was presented by Héon and Paquette [17] to model and visualize OWL2 and ontologies created using RDFs. G-OWL simplified the use and interpretation of the ontologies by using semantic and syntactic rules. Furthermore, extensible and flexible RDFS were designed by Staab [18] to exchange and access the axioms in RDFs.

Narayanasamy et al. [19] proposed a detailed review of the advances in the field of semantic web and Ontology.

3 Methods for Representation and Retrieval of Knowledge

Intuitive structures based on Ontology are questioning refinement frameworks for data sets using visual portrayals to impart relevant information request. The systems had the option to change Ontology for database inquiry ploys to work on practicality of the associations of human and PC. Various frameworks are identified in the research literature [20], e.g. TAMBIS, the GRQL, the SEWASIE, the Ontogator, the OntoViews, the OntoQF, the VISAGE, the Smartch, semantic-based and various others. The TAMBIS framework [21] reinforces the specialization or hypothesis of

Table 1 Knowledge Representation in Ontology

Concepts	OWL 1	OWL 2	RDFs
Class definitions	✓	✓	✗
Cardinality restrictions	✓	✓	✗
Constraint	✓	✓	✓
Disjoint properties	✗	✓	✓
Equivalence	✓	✓	✓
Enumerations	✓	✓	✓
Formal semantics	✓	✓	✓
Inference	✓	✓	✓
Qualified cardinality restrictions	✗	✓	✓
Property chains	✗	✓	✓

the base or filler Ontological ideas to develop data base unequivocal inquiries instinctively. One more similar system subject to ontological graph configuration questions [22] is given in GRQL and the Knowledge Sifter. The GRQL depends entirely upon the RDF/S data model and gives a graphic user interface to form requests depending on the Ontological course. This framework is focussed on the questions that are created by the graphical investigation through the individual classes for RDF/S and property definitions.

In SEmantic Webs and AgentS in Integrated Economies (SEWASIE) [23], the basis of organizing and development of a productive query interface based on Ontology are presented. Contrary to the other existing approaches, OntoQF is used that involves the usage of both the substitution of database to Ontology and maps for empowering of the detailed query process that has been programmed to assist in creation of the processes for database. A question and extraction of information board has been introduced in a draft of the EU Translational study and Safety of the Patient in Europe named as TRANSFoRM [24]. The TRANSFoRM provides an interface to the designer, store and share queries of the clinical knowledge for the recognition of the subjects for their medical investigations.

Based on the literature discussed above, we have identified and discussed various methods to access the Ontology. First step being representing the knowledge in Ontology, followed by retrieving the information from the designed Ontology. The knowledge can be stored in the Ontology by using RDF, OWL 1 and OWL 2 as shown and discussed in Table 1. The information can be retrieved by various techniques as shown and discussed in Table 2.

3.1 Knowledge Representation

The RDF and OWL1 and OWL2 that show various concepts for the formulation of an Ontology are shown Table 1.

Table 2 Information retrieval from Ontology

Methods	Query assistance by text	Query assistance by semantic clause	Supports multimedia database	Duplication of data not needed in Ontology	Heterogeneous data sources support	Natural language processing
TAMBIS	✓	✓	✗	✗	✗	✗
SEWASIE	✓	✓	✗	✓	✓	✗
GRQL	✓	✓	✗	✗	✗	✗
OntoViews	✓	✓	✓	✗	✗	✗
Ontogator	✓	✓	✓	✗	✗	✗
OntoQF	✓	✓	✗	✓	✓	✗
TRANSFoRM	✓	✓	✗	✓	✓	✗
Smartch	✓	✓	✗	✗	✗	✗
VISAGE	✓	✓	✗	✓	✓	✗
CROEQS	✓	✓	✓	✓	✗	✗
KIRA	✓	✓	✓	✓	✓	✗
Ontop	✓	✗	✗	✓	✓	✗
Optique	✓	✗	✗	✓	✓	✗
Ontology and natural language	✓	✗	✗	✓	✗	✓
querying via OWL 2 querying language	✗	✓	✗	✓	✗	✗
Using Ontology SPARQL	✓	✗	✗	✓	✗	✓

Concluding the comparison, it could be seen that OWL is more powerful than RDFs with increased machine interpretability and vast vocabulary and a stronger syntax than RDFs.

3.2 Measures for Information Retrieval

There are various measures that can be used to access the Ontology that has been developed such as TAMBIS, OntoViews, GRQL, VISAGE, SEWASIE, Ontogator, OntoQF, semantic-based and Smartch.

TAMBIS builds database queries based on filler Ontology that is supported by its generalization or specialization. GRQL depends on the RDFs data design and forms a graphic user interface for query building that is based navigation of the Ontology through class and property representations. In SEMantic Webs and AgentS in Integrated Economies (SEWASIE), the designing and developing protocols to form an Ontology are shown. In OWL DL Ontology, information is retrieved by automatic calling of relational database queries. Ontogator and OntoViews are image retrieval systems. Table 2 represents a contrast of several major Ontology-based query designing tools and methods.

3.3 *Ontology-Based Information Retrieval*

Retrieval of information is the pursuit for knowledge in the Ontology. The necessity for successful approaches to retrieve the data/knowledge has developed in abundance on account of elevating the amount of both, structured and unstructured data. Various Ontology languages with significant semantics have been developed. Over the years, various visual IE [5, 20] methods existed that were intended to end the exertion of a user associated with the databases. The query optimizations [25] based on ER could not form a strong foundation that could depend on the exhaustiveness for communication with low level queries. So to overcome this deficiency, various tools for Ontology were introduced that were capable of search operations in the given semantics, shown in the Table 3.

There are various domains in which the Ontology can be used and applied. One important domain identified that needs attention is health care. A lot of advances have been done in the Ontology in the domain of health care, providing decision support to the medical experts and building intelligent decision support system by incorporating artificial intelligence and Ontology. The same is explained in the section below.

Table 3 Standard tools for semantic search from Ontology

Ontology tool	Licensing	Language used
Apache	Apache License	Java
Fuseki	W3C	Java
Virtuoso	GPL	C
Sesame	BSD	Java
Blazegraph	GPL	Java
Protege	BSD 2-clause	Java

4 Standard Ontologies in Healthcare Domain

Over the years, the data produced in the biomedical research has been increasing enormously, and the structuring of this data becomes a challenge. The medical data is being generated by various origins such as hospitals, research centres, clinical images, pathological reports, and medical records gathering and storage of the data is not much of a bigger challenge as various technologies for the same are present. Moreover, the technologies provided a way for storage of information in a transparent manner that can be accessed by the user [25]. However, storage of the data and its retrieval is a matter of concern in case of some distributed systems and the major issues arise in interoperability of such data between systems. The medical data being generated by various sources leads to a cumbersome process for its unification on a single platform. To address the same issue, Ontologies come into play. The data from various sources is structured in a single system, and the information is retrieved from the same with the help of various methodologies as described in the above sections.

Some of the basic methods used for the utilization of the medical data are the Ontologies. In order to assist the Ontologies in medical realms, few strides have been followed for better search and the integration of data:

- (a) The medical data is converted to suitable RDFs by identifying the capable medical structures that have been identified by the Ontology.
- (b) Proper rules have been set by using a suitable RDF schema and the hierarchy for data integrations is included. Also, the original medical data is transformed into unified RDF representations.
- (c) The refined data is stored in the Ontology using OWL.
- (d) The retrieval of RDF data and information is done semantically using SPARQL query.

Some of the conventional Ontologies that have been used in the medical domain are listed in Table 4.

5 Conclusion and Future Work

The contribution of semantic web has been growing rapidly with the usage of Ontologies and other modelled application of Ontology. The Ontologies have provided many ways to change the unstructured data or the semi-structured data into the standard and structured data by using RDF schema and OWL frameworks. In this research survey, we focussed upon the methods and techniques used for creation of Ontology and accessing the same. After identifying, a number of methods as discussed in the paper it has been concluded that for storing the data into the Ontology, OWL 1 and OWL2 are most commonly used and the information retrieval is majorly done using NLP techniques and SPARQL.

Table 4 Usage of conventional Ontologies in healthcare systems

References	Healthcare Ontology	Description	Format	Total classes
Satija et al. [23]	ATC	The Ontology is capable of gathering details that are related to the medicine ingredients of organ and its classification by using the chemical properties	UMLS	6358
Damjanović [21]	DOID	The Ontology is related to general human sickness and segregation of the details of illness on few medical properties	OBO	12,694
Robson and Barr [22]	HP	Classification of the monogenic illness and sequence the medical vocabulary for the grammatical highlights	OBO	18,407
Stutt and Motta [24]	MeDRA	Classification of drug and discovering consequences on health	UMLS	73,429
Zhu et al. [26]	PMR	Rehabilitation data for patients	OWL	1597

After studying, the techniques used or the languages used for retrieval of information from a given Ontology SPARQL is considered as a language to construct Ontology for our desired work due to the following reasons:

- User friendly interface
- Open source
- Easy to learn and easy to use
- Querying in the form of text
- NLP techniques
- Manageable

The future work would be the construction of an Ontology in OWL using Protégé tool editor and the information retrieval from the Ontology designed using SPARQL language in healthcare domain in real-time applications.

References

1. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci Am* 284(5):34–43
2. Smith B (2012) Ontology. In: The furniture of the world 2012. Brill, pp 47–68, 1 Jan 2012
3. Gruber TR (1993) A translational approach to portable ontology specifications. *Knowl Aquis* 5:199–220
4. OWL. <http://www.w3.org/TR/owl-features/>
5. Munir K, Anjum MS (2018) The use of ontologies for effective knowledge modelling and information retrieval. *Appl Comput Inform* 14(2):116–126
6. Degbelo A (2017) A snapshot of ontology evaluation criteria and strategies. In: Proceedings of the 13th international conference on semantic systems, pp 1–8, 11 Sep 2017
7. Balamurugan M, Iyswarya E (2018) Construction for research paper selection ontology using Protégé. *Int J Appl Eng Res* 13(9):6989–6993
8. Pascal H (2021) A review of the semantic web field. *Commun ACM* 64:76–83
9. Rubin DL, Knublauch H, Fergerson RW, Dameron O, Musen MA (2005) Protege-owl creating ontology-driven reasoning applications with the web ontology language. In: AMIA annual symposium proceedings 2005 vol 2005, p 1179. American Medical Informatics Association
10. Schneider T, Šimkus M (2020) Special issue on ontologies and data management: part I. *KI-Künstliche Intelligenz*. 34(3):287–289
11. Schneider T, Šimkus M (2020) Special issue on ontologies and data management: Part II. *KI-Künstliche Intelligenz* 34(4):439–441
12. Alqawasmeh O, Towards a collaborative framework for ontology engineering: Impact on ontology evolution and pitfalls in ontology networks and versioned ontologies (Doctoral dissertation, Lyon)
13. Abdel-Qader M, Vagliano I, Scherp A (2019) Analyzing the evolution of linked vocabularies. In: International conference on web engineering. Springer, Cham, pp 409–424, 11 Jun 2019
14. Malviya N, Mishra N, Sahu S (2011) Developing university ontology using protégé owl tool: process and reasoning. *Int J Sci Eng Res* 2(9):1–8
15. Bobillo F, Straccia U (2009) An OWL ontology for fuzzy OWL 2. In: International symposium on methodologies for intelligent systems. Springer, Berlin, Heidelberg, pp 151–160, 14 Sep 2009
16. Su X, Meng X, Hu H (2013) Research on development and application of OWL ontology search middleware. *Math Comput Model* 58(3–4):614–618
17. Héon M, Paquette G (2020) G-OWL: A complete visual syntax for OWL 2 ontology modeling and communication. *Semant Web J*
18. Staab S, Erdmann M, Maedche A, Decker S (2002) An extensible approach for modeling ontologies in RDF (S). In: Knowledge media in healthcare: opportunities and challenges 2002. IGI Global, pp 234–253
19. Narayanasamy SK, Srinivasan K, Hu YC, Masilamani SK, Huang KY (2022) A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and ontology-based information processing systems. *Electronics* 11(3):453
20. Imhof M, Braschler M (2018) A study of untrained models for multimodal information retrieval. *Inf Retr J* 21:81–106
21. Damjanović B (2017) SOA and services orchestration: History, role and open source technologies. *Info M* 2017 16:16–26
22. Robson R, Barr A (2018) Learning technology standards-the new awakening. In: Proceedings of the sixth annual GIFT users symposium, Orlando, FL, USA, vol 6. Army Research Laboratory, Adelphi, MD, USA, pp 1–9, 30 May–2 June 2018
23. Satija MP, Bagchi M, Martínez-Ávila D (2020) Metadata management and application. *Libr Her* 58:84–107
24. Stutt A, Motta E (2004) Semantic webs for learning: A vision and its realization. In: Proceedings of the international conference on knowledge engineering and knowledge management, Whittlebury Hall, UK. Springer, Berlin/Heidelberg, Germany, pp 132–143, 5–8 Oct 2004

25. Cano AE, Rizzo G, Varga A, Rowe M, Stankovic M, Dadzie AS (2014) Making sense of micro-posts:(# microposts2014) named entity extraction & linking challenge. In: CEUR workshop proceedings, vol 1141. CEUR-WS.org, Seoul, Korea, pp 54–60
26. Zhu Y, Yan E, Song IY (2017) The use of a graph-based system to improve bibliographic information retrieval: system design, implementation, and evaluation. *J Assoc Inf Sci Technol* 68:480–490
27. Caldarola EG, Rinaldi AM (2016) An approach to ontology integration for ontology reuse. In: Proceedings of the 2016 IEEE 17th international conference on information reuse and integration (IRI) IEEE, Pittsburgh, PA, USA, pp 384–393, 28–30 Jul 2016
28. Sanjana M, Badar P (2019) The utilization of ontologies for knowledge model and data recovery. *Perspect Commun Embed Syst Signal-Process-PiCES* 3:5–8
29. Selvalakshmi B, Subramaniam M (2019) Intelligent ontology based semantic information retrieval using feature selection and classification. *Clust Comput* 22:12871–12881
30. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, Ethier JF, Kostopoulou O, Ku-chinke W, McGilchrist M et al (2015) Translational medicine and patient safety in Europe: TRANS-FoRm—architecture for the learning health system in Europe. *BioMed Res Int* 2015:961526

Design and Analysis of GIGA Fiber like Connectivity of 5G Technology Using 60 GHz Band



Bhanu P. Singh and Anand Agarwal

Abstract With the advent of the 5G era, high end intelligent devices have shown tremendous growth in data traffic and acquired an ultra-high transmission rate for the various emerging applications. Unfortunately, researchers around the world have observed that the present 5G cellular technology is unable to attain the set objectives, such as fiber like connectivity, high end service quality and its security, interoperability among the existing & evolving set of RF/communication standards and others. With the emphasis, massive multi–input–multi–output (mMIMO) aided GIGA fiber using PSO-NN scheme has been investigated. Additionally, MATLAB simulation of various performance measures of the mMIMO-PSO-NN system at 60 GHz spectrum have been evaluated and compared with the reported results.

Keywords 5G · GIGA fiber · Channel optimizations

1 Introduction

With the advancement of milli wave and optical communication era, there have been several different applications due to offering advantages such as low-power consumption using, high-mechanical stability, low footprint, small dimension, enhanced functionalities and ease of complex system architectures. Additionally, to support the anticipated massive devices, there has been general consensus that the fifth-generation (5G) wireless communication system is the viable and promising solution. Meanwhile, massive multiple-input multiple-output (M-MIMO) antenna and millimeter-wave (mm-wave) technologies are anticipated to be integrated into the 5G networks, so as to enhance the wireless system bandwidth. Today, optics empowered visual correspondence is still generally utilized in our day-to-day routines. For

B. P. Singh (✉)

ECE, Lakshmi Narain College of Technology, Bhopal, Madhya Pradesh, India
e-mail: bhanuprataps@lnct.ac.in

A. Agarwal

ECE, Indian Institute of Information Technology, Kota, India
e-mail: anand.ece@iitkota.ac.in

instance, airplanes utilize the arrival lights at air terminals to land securely, particularly around evening time and under enemy climate conditions. Airplanes setting down on a plane carrying warship utilize a comparative framework to land accurately on the transporter deck. Transports regularly utilize a sign light to flag in Morse code or utilize global oceanic sign banners to trade messages. Trouble flares are utilized by mariners in crises while beacons and route lights are utilized to impart route risks [1, 2].

As a general rule, an optical correspondence framework comprises of a transmitter or a light source which encodes a message into an optical sign, a channel or a waveguide which conveys the sign to its objective, and a beneficiary which recreates the message from the got optical sign. The cutting edge directed optical correspondence depends on the light's absolute inner reflection guideline, clarified by the Snell's Law, which has been known for a really long time and was utilized to enlighten floods of water in intricate public wellsprings in Victorian occasions. The advancement of present-day optical correspondence is obliged to innovative forward leaps and upgrades acknowledged in various regions: light source—Drove and laser, materials and the assembling of low-misfortune light wave guides—optical strands, and different parts and gadgets fundamental for compelling optical transmission and correspondence, just as modern hardware and sign handling procedures.

1.1 Evolution of Optical-Fiber Transmission System

The development of correspondence organization, which has and will keep on delivering a broad social advancement from the previous a long time to this new thousand years, owes particularly to the new advancement in the correspondence advances.

Figure 1 shows the optical communication with different technologies. Most importantly, the improvement of Optical-fiber correspondence innovation, which permits communicating an enormous amount of data over longer distance at decreased expense, has seriously advanced the advancement. Early frameworks of late 1970s through the mid-1980s involved LEDs or MLM laser transmitters in the 0.8 and 1.3 μm frequency groups and multi-mode filaments, empowering the sign to be communicated for a sensible distance before the sign should be recovered each couple of kilometers (e.g., 10 km) through an optical-electro-optical recovery process.

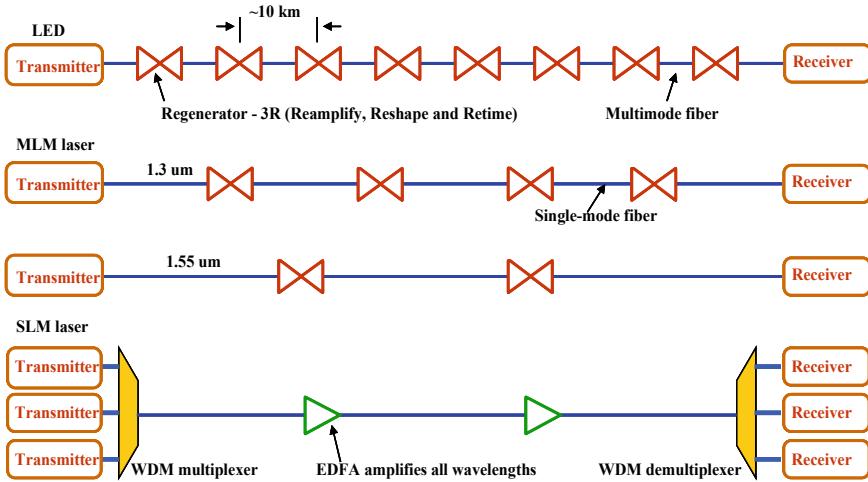


Fig. 1 Evolution of optical-fiber transmission systems

2 Methodology

2.1 Particle Swarm Optimization (PSO)

PSO is a developmental algorithm. At each iteration, every particle is update or change its velocity (V_{knew}) and position (X_{knew}) by using Eq. (1) and Eq. (2). Particle place or position in a swarm found through the Fitness function. It describes best position (called P_{best}) and quality of each particle. The velocity of each particle is compared continuously with its current value and previous value, until reach the optimum value. If previous value is better than present value than it changed its value with own best previous position (P_{best}) that, best position found by particle itself (fitness value). PSO used for channel selection in CRN other than using other stochastic algorithms. Suppose, X_k represent the position of the k th particle and V_K represent the velocity of k th particles.

$$V_k^{new} = W * V_K + C_1 r_1 (p_{best_k} - X_k) + C_2 r_2 (n_{best_k} - X_k) \quad (1)$$

$$X_k^{new} = X_k * V_k^{new} \quad (2)$$

where W = inertial weight, n = number of particles in the swarm, $c1$ and $c2$ = acceleration constants and $r1$ and $r2$ = random numbers distributed in $[0, 1]$. In which, $W, c1$ and $c2$ represents the changed of their own previous velocities, personal best position and its neighbor's best position by the new velocity, respectively. At

last, the swarm will converge to the optimal location, as it is driven by individual particle experience and global experience.

2.2 Neural Network

The corresponding conditional probability of the occurrence to be offered is

$$\hat{y} = P(\text{decision} = 1|w) = g(w^T f) \quad (3)$$

$$g(a) = \frac{e^a}{1 + e^a} \quad (4)$$

where g represents the best function evaluated at activation a . Let w denote weight vector and f the column vector of the importance functions: $f^T = [f_1, \dots, f_5]$. Then the “decision” is generated according to the model.

2.3 Parallel Implementation of the PSO-BP Neural Network Algorithm

The arrival of the big data era poses a challenge to traditional machine learning algorithms. In information technology, big data is a collection of datasets so large and complex that they become difficult to process using available database management tools or traditional data processing applications. Big data is usually composed of datasets with sizes beyond the ability of commonly used software tools, which are unable to capture, curate, manage, or process such data within a reasonable elapsed time. The challenges include capture, duration, storage, search, sharing, analysis and visualization. Therefore, both the time and space efficiency of traditional algorithms decrease dramatically when addressing big data. Most frameworks utilize a deterministic dynamic directing calculation: When a gadget picks a way to a specific last goal, that gadget dependably picks a similar way to that goal until the point that it gets data that influences it to think some other way is better. A couple of directing calculations don't utilize a deterministic calculation to locate the “best” connection for a parcel to get from its unique source to its last goal. Rather, to maintain a strategic distance from blockage in exchanged frameworks or system problem areas in parcel frameworks, a couple of calculations utilize a randomized calculation—Valiant's worldview—those highways a way to an arbitrarily picked middle of the road goal, and from that point to its actual last destination [2, 3]. In numerous early phone switches, a randomizer was frequently used to choose the beginning of a way through a multistage exchanging texture. Contingent upon the application for which way determination is performed, diverse measurements can be utilized. For instance,

for web demands one can utilize least inactivity ways to limit site page stack time, or for mass information exchanges one can pick the slightest used way to adjust stack over the system and increment throughput. A mainstream way choice goal is to lessen the normal fruition times of movement streams and the aggregate system data transmission utilization which essentially prompt better utilization of system limit. As of late, a way choice metric was suggested that figures the aggregate number of bytes booked on the edges per way as determination metric [4]. An observational investigation of a few way choice measurements, including this new proposition, has been made accessible.

3 Results and Discussion

The MATLAB software is used for the implementation and the simulation of the proposed research. The communication library function and commands are used for the MATLAB script. The Neural Network iteration method is capable of calculating the maximum of the thousands of the nodes. In this each node depends on the previous node for the optimal power allocation. For example, consider a single node and let us assume 10 s as duration for each node and if we consider for the 10 nodes then the time should be of 100 s. So, in this it calculates for each and every node and assumes the value itself when there are the multiple numbers of the nodes.

In the random instances method, the calculation will be done completely in the random manner and finally from the above graph the optimal power allocation is obtained.

Figure 2 is showing continuation of simulation of proposed approach. It is clear that iteration is running till achieve P_{best} value.

In Table 1, showing initial value that is called weight before then after simulation and first iteration optimize value or best value (P_{best}) generate. Figure 3 showing graphical representation of initial value that is called weight before then after simulation.

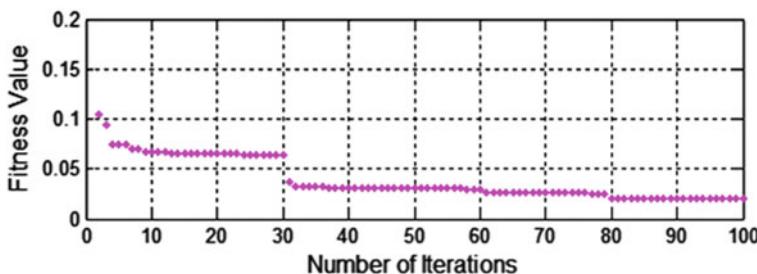


Fig. 2 Simulation of PSO + NN iteration

Table 1 Weight before and weight after-I

Weight before	Weight after-I
0.1991	0.1783
0.1821	0.1824
0.1614	0.1789
0.1625	0.1744
0.1529	0.1608
0.1420	0.1252

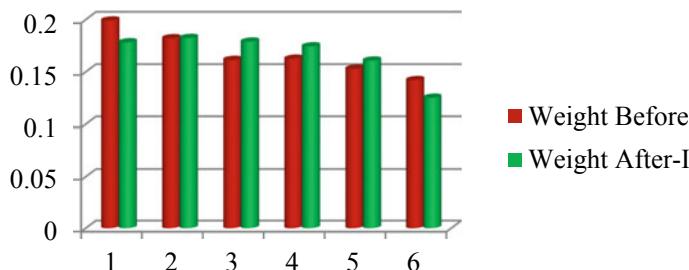
**Fig. 3** Weight before and weight after-I

Figure 4 is showing graphical representation of simulation result of first iteration. Iteration is continuing till 100 and during iteration, some parameters value is generating in case of weight before and weight after-I. The parameters are channel Signal strength, Bandwidth, Power, speed and connectivity are including.

In Fig. 5, Fast path is a term utilized as a part of software engineering to portray a way with shorter direction way length through a program contrasted with the “ordinary” way. For a quick way to be successful it must deal with the most ordinarily happening undertakings more productively than the “typical” way, leaving the last to deal with phenomenal cases, corner cases, mistake taking care of and different oddities. Quick ways are a type of enhancement.

Figure 6 is showing graphical representation of simulation result of iteration-I to iteration-V. This graph also shows value of global best in each iteration.

Table 2, present parameter value or defined model. In which take more population size for PSO and NN than previous work. Table 3 is showing comparison of previous work with proposed work. So, it can be seen that previous approach used PSO and GA for optimization of best value while proposed approach use PSO and NN. Due to this iteration count is reduced so achieved optimize value in low time. Average standard deviation value is also reduced. This approach is also used for more population size. Therefore, proposed research provides optimal solution for GIGA fiber network channel optimization and handover route.

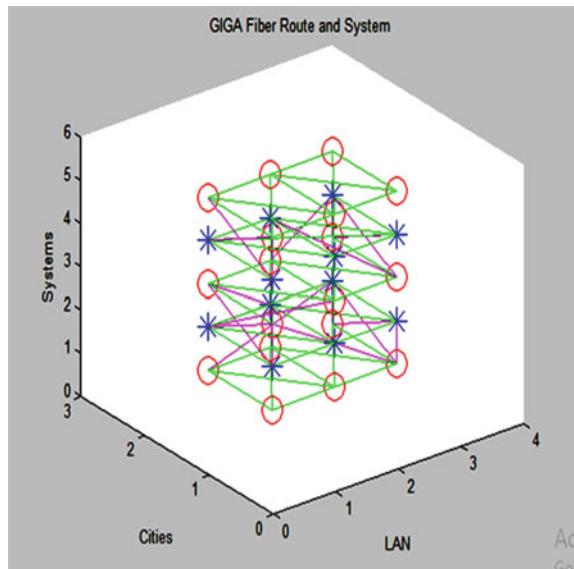


Fig. 4 GIGA fiber route

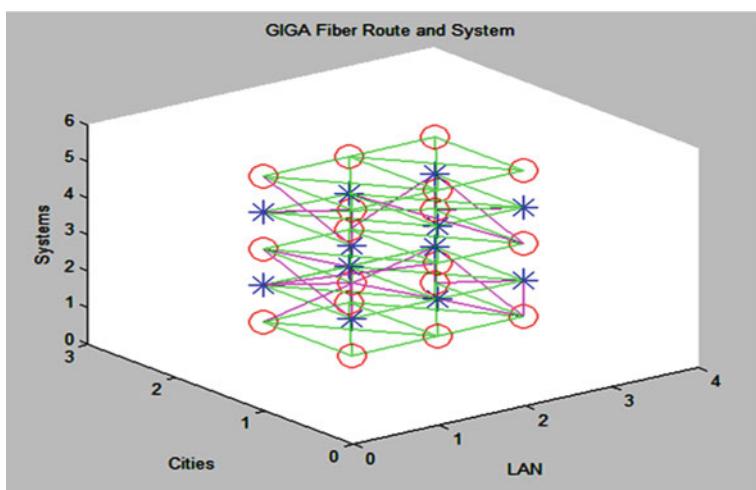


Fig. 5 To identify fast delivery handover route

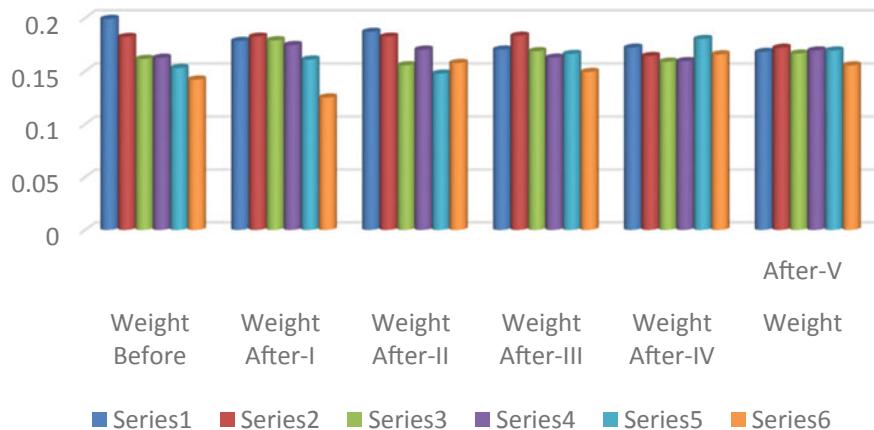


Fig. 6 Weight Before and weight after-I to V-1

Table 2 Calculated parameters for proposed work

Sr. No.	Parameters	Proposed work
1	Method	PSO + NN
2	Iteration	100
3	Population size	PSO-30 and NN-30
4	Inertial weight	1–6
5	Cross over rate	0.4

Table 3 Comparison of results parameters

Parameters	Previous work [5, 6]	Proposed work
Method	PSO + GA	PSO + NN
Iteration	100–3000	100
Population size	PSO-12 and GA-12	PSO-30 and NN-30
Fitness value	0.00145–0.00175	0.05–0.15
Standard deviation	2.5727×10^{-5} and 5.389×10^{-5}	2.389×10^{-3}
Inertial weight	0.6	1–6
Cross over rate	0.5	0.4

4 Conclusions

The GIGA fiber-optical communication is a recent technology to provide Internet services with higher speed and quality of services. The network channel optimization with improvement of speed, connectivity, signal strength and bandwidth is the key challenges of this technique. The optimization technique provides the good values for implementation and performance enhancement. This paper proposed the particle swarm optimization (PSO) with neural network (NN) approach for implementation of GIGA fiber-optical channel using 5G network. The simulation is performed using MATLAB software. The network selection and channel optimization are great challenges in the free space fiber optical communication. The various soft computing techniques like neural network, genetic algorithm, particle swarm optimization can provide the significant better results.

Acknowledgements We would like to thank SERB (Science and Engineering Research Board) for the funding and provide this wonderful research opportunity.

References

1. Wang G et al (2021) Stable and highly efficient free-space optical wireless communication system based on polarization modulation and in-fiber diffraction. *J Lightwave Technol* 39(1):83–90. [https://doi.org/10.1109/JLT.2020](https://doi.org/10.1109/JLT.2020.3000000)
2. Velandia JG, P. López KC, Ortiz LAG (2021) Linear and non-linear effects in fiber optic transmission. In: 2021 IEEE Colombian conference on communications and computing (COLCOM) pp 1–6
3. Tian X, Hu L, Wu G, Chen J, Hybrid fiber-optic radio frequency and optical frequency dissemination with a single optical actuator and dual-optical phase stabilization. *J Lightwave Technol* 38(16), pp 4270–4278
4. Umezawa T, Dat PT, Kanno A, Yamamoto N (2020) Fiber wireless and optical wireless communications using high-speed photonic devices. *Int Conf Adv Technol Commun (ATC)* 2020:1–4
5. Sharma V, Sergeyev S, Kaur J (2020) Adaptive 2×2 MIMO employed wavelet-OFDM-radio over fibre transmission. *IEEE Access* 8:23336–23345
6. Delmade A et al (2021) Optical heterodyne analog radio-over-fiber link for millimeter-wave wireless systems. *J Lightwave Technol* 39(2):465–474

Coronavirus Herd Immunity Optimization-Based Control of DC-DC Boost Converter



Manoj Sai Pendem, Tousif Khan Nizami, Priyanka Singh, and Mohamed Shaik Honnurvali

Abstract This paper presents a novel coronavirus herd immunity optimization (CHIO) algorithm for tuning the proportional-integral-derivative (PID) controller for the DC-DC boost converter. The closed-loop control action using the PID controller is designed to regulate the output voltage of DC-DC boost converter across the load end. CHIO is a nature-inspired meta-heuristic optimization algorithm formulated based on the way humankind handled the coronavirus pandemic (COVID-19) in recent years. This optimization algorithm exploits the herd immunity and social distancing concepts. The optimization algorithm has been developed on MATLAB/Simulink software for obtaining the optimum PID controller gains. Extensive simulations are conducted under (i) start-up response, (ii) reference voltage change (iii) load resistance change, and (iv) input voltage change to find the performance of the proposed controller. The obtained results indicate a successful convergence and satisfactory dynamic response of the output voltage under wide variation in the operating points.

Keywords DC-DC boost converter · Proportional-integral-derivative (PID) controller · Herd immunity · COVID-19

1 Introduction

The widespread usage of battery-operated systems and renewable energy systems has raised the demand for DC-DC converters [1]. One of the most popular converters that can increase the input voltage across its output terminal is the DC-DC boost converter. Since the boost converter is a non-linear system, its performance and efficiency are decreased by increasing the voltage gain. This is due to an increase

M. S. Pendem · T. K. Nizami (✉) · P. Singh
SRM University-AP, Guntur, Andhra Pradesh, India
e-mail: tousif.k@srmmap.edu.in

M. S. Pendem
e-mail: pendem_manoj@srmmap.edu.in

M. S. Honnurvali
A'Sharqiyah University, Ibra, Sultanate of Oman

in converter switching losses when the duty cycle increases [2]. To enhance the performance and ensure constant voltage regulation regardless of load disturbances, closed-loop control action is necessary. Different types of controllers can be used to implement closed-loop control. Due to its ease of use and cost-effectiveness, the proportional-integral-derivative (PID) controller are the most popular controllers in recent decades [3].

In the literature, numerous ways to tune the PID controller have been proposed. There are two primary categories of tuning procedures: conventional methods and intelligent methods. The conventional approaches, such as the Ziegler-Nichols method, the good gain method, the damped oscillation method [4], etc. These techniques are suitable to tune the gain parameters for linear, steady systems. Intelligent approaches are fuzzy logic, artificial neural network (ANN), adaptive neuro-fuzzy interference system (ANFIS), meta-heuristic algorithmic methods [5], etc. These methods can deliver the better performance compared to conventional ways of tuning. Vikram Chopra et al proposed the PID controller tuning rules using intelligent techniques based on fuzzy logic, artificial neural network (ANN), adaptive neuro-fuzzy interference system (ANFIS), and genetic algorithm (GA) [6]. The authors compared the performance specification like rise time, peak overshoot, and steady-state error with the conventional Ziegler-Nichols tuning technique. Finally, the authors concluded the intelligent method outperformed the conventional method of tuning. Amongst all the intelligent methods, meta-heuristic methods are one of the efficient ways to find a feasible solution for a complex problem in a reasonably lesser time [7]. Similarly, other methods proposed include particle swarm optimization, cuckoo search algorithm [8], and Bat algorithm [9] amongst meta-heuristic techniques in the literature.

In this paper, a coronavirus herd immunity optimizer (CHIO) is used to tune the PID controller for the DC-DC boost converter. CHIO is a novel nature-inspired meta-heuristic optimization algorithm and it is formulated in the ways in which humankind handled the the coronavirus pandemic (COVID-19) in recent years [10]. This optimization technique is based on herd immunity and social distancing concepts. The numerical simulations are carried out in MATLAB/Simulink environment.

The paper is organized as follows, the boost converter modeling and specifications were discussed in Sect. 3 and the optimization process and algorithm is discussed in Sect. 4. Section 5 discusses results and observation and the work is concluded in Sect. 6.

2 Problem Statement

The DC-DC boost converter is expected to maintain a higher level of output voltage then at its source. Due to input and load uncertainties, the open loop DC-DC boost converter do not produce the desired output voltage and instead exhibits poor dynamic and steady-state behavior. Hence, PID controller is adopted to yield an enhanced

transient and dynamic response. The objective function considered for this problem is based on sum squared error, as follows:

$$f(x) = \sum_{t=1}^{\infty} (V_{\text{ref}} - V_o(t))^2 \quad (1)$$

where V_{ref} represents the reference voltage and V_o is the actual voltage at any given instant of time.

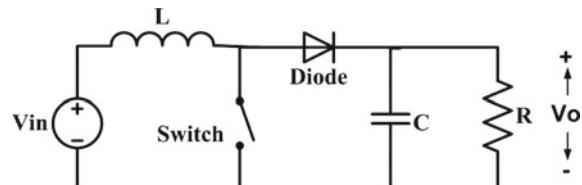
3 DC-DC Boost Converter

The DC-DC boost converter increases the DC input voltage at the output load terminal. This converter circuit comprises an inductor, a capacitor, a diode, and a switch coupled as shown in Fig. 1. The switching mechanism determines which of two modes the boost converter operates. Mode 1 is active when the switch is closed. At this point, the input circuit is completely disconnected from the load component. The magnetic field of the inductor will start to store energy as it starts to charge leading to increase in output voltage. When the switch is closed, mode 2 is activated. The capacitor will begin charging at this time and load will receive power from both the input source and the inductor. The state space model of boost converter is shown in Eqs. 2 and 3, where duty cycle \hat{d} has to be regulated using closed-loop control section.

$$\begin{bmatrix} \frac{d\hat{i}_L}{dt} \\ \frac{d\hat{V}_c}{dt} \end{bmatrix} = \begin{bmatrix} 0 & -(1-D) \\ \frac{(1-D)}{C} & \frac{L}{RC} \end{bmatrix} \begin{bmatrix} \hat{i}_L \\ \hat{V}_c \end{bmatrix} + \begin{bmatrix} \frac{1}{L} & \frac{V_c}{L} \\ 0 & \frac{-1}{C} \end{bmatrix} \hat{V}_{\text{in}} \hat{d} \quad (2)$$

$$v_o = [0 \ 1] \begin{bmatrix} \hat{i}_L \\ \hat{V}_c \end{bmatrix} \quad (3)$$

Fig. 1 Boost converter circuit



4 Coronavirus Herd Immunity Optimization Algorithm

4.1 Background

CHIO is a novel nature-inspired human-based meta-heuristic optimization algorithm [10]. It originated from a way, we handled the coronavirus pandemic (COVID-19) in recent years. Recently, the novel 2019 coronavirus spread globally. This virus has a huge impact due to its high reproduction rate of 2–2.5 and also fast transmission speed of 3–5 days. Normally, the virus spread and evolves very quickly among individuals in the population. Vaccines are made to build immunity to fight against the virus. However, new viruses need a period for preparing the vaccines. To control this outbreak during this time, the World Health Organization proposed herd immunity and social distancing techniques. These proposed techniques work to protect the community from disease transmission by driving them into herd immunity. Herd immunity is the state of the situation when most people get into an immune state and this leads to avoiding the transmission of disease. The population has been divided into three groups during this process.

- **Immune population:** These people received the virus and recovered from it. These individuals won't carry the disease ever again.
- **Susceptible population:** However not affected by the virus, these people lacked immunity. Therefore, there is a chance that those who come in contact with infected cases will become infected.
- **Infected population:** The virus was present in these people, and they may spread it to an additional susceptible case.

The contrast between a population with herd immunity and without is shown in Fig. 2. It has been observed that herd immunity reduces the probability of disease transmission in a community by indirectly protecting the susceptible cases. As a result, several medical professionals have recommended that herd immunity is one method for preventing the spread of disease. The disease will first infect a portion of the population, and those who have the strength to fight it will recover and enter a condition of immunity, eliminating the fewest number of people who died. And as a result, herd immunity develops among the population. Accordingly, based on the survival of the fittest principle, people shift from being susceptible to becoming infected and again from being infected to becoming immune.

Fig. 2 Population with and without herd immunity

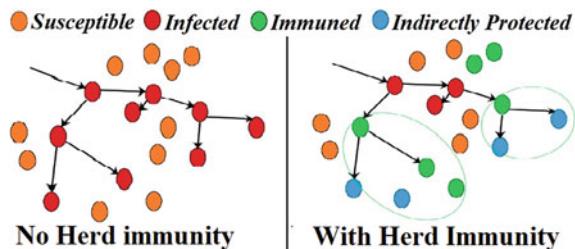
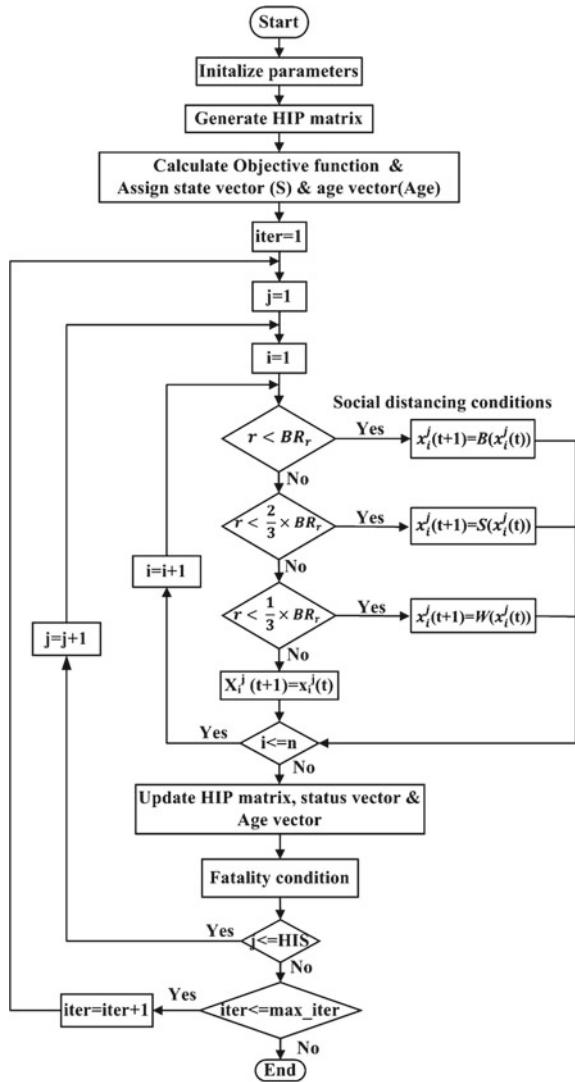


Fig. 3 CHIO algorithmic flow chart



4.2 Optimization Process

According to these herd immunity strategies, the optimization algorithm is modeled. The algorithm flow chart is shown in Fig. 3. This optimization process consists of six steps and it has been described as follows:

Step 1: Initializing the parameters,

- Co indicates the number of initial infected cases.
- HIS indicates the population size.
- The HIP is the matrix consisting of the population.
- n is the problem dimensionality.

There are two control parameters,

- Basic reproduction rate (BRr): Basic reproduction rate or spreading rate of the virus depends upon several factors. It varies from place to place.
- Max age (Maxage): it is the max-age limit for any individual. If the person is not recovered after this age limit, then the individual will be considered a fatality case.

Step 2: Generate herd immunity population. The HIP matrix consists of population, the size of this matrix is defined by HIS and n .

$$\text{HIP} = \begin{bmatrix} x_1^1 & \dots & x_1^n \\ \vdots & \ddots & \vdots \\ x_{\text{HIS}}^1 & \dots & x_{\text{HIS}}^n \end{bmatrix} \quad (4)$$

$$x_i^j = \text{lb}_i + (\text{ub}_i - \text{lb}_i) \times r$$

where,

$$i = 1, 2, \dots, n$$

$$j = 1, 2, \dots, \text{HIS}$$

R is the random number between 0 and 1.

lb and ub are the lower bound and upper bounds.

Every individual in the HIP matrix is formed using Eq. 4. For every solution set, calculate the objective function and store them in the fitness vector and assign the status vector and age vector depending on it. The status vector shows the status of every individual. Initially, the status vector initialization depends on Co and further, it will be updated as 1 for the infected (worst case), 2 for the immune (best case), and 0 for the susceptible case.

Step 3: Herd immunity evolution.

This is the main improvement loop, every individual element x_i^j will be updated or remain the same based on social distancing conditions. These social distancing conditions are formulated as follows:

$$x_i^j(t+1) = \begin{cases} W(x_i^j(t)), & 0 \leq r < \frac{1}{3} \times \text{BR}_r \\ S(x_i^j(t)), & \frac{1}{3} \times \text{BR}_r \leq r < \frac{2}{3} \times \text{BR}_r \\ B(x_i^j(t)), & \frac{2}{3} \times \text{BR}_r \leq r < \text{BR}_r \end{cases}$$

The individual will remain the same if $r \geq \text{BR}_r$. Where, r is a random number between 0 and 1.

Infected case: if r is in the range of $r \in [0, \frac{1}{3} \times \text{BR}_r]$, then the new updated elements will depend on the previous infected case. It resembles, the individual is effecting by the infected case.

$$W(x_i^j(t)) = x_i^j(t) + r \times (x_i^j(t) - x_i^c(t)) \quad (5)$$

where, $x_i^c(t)$ is the randomly selected infected case based on status vector, $c = i|Si = 1$.

Update the variable $\text{is}_{\text{corona}}$ as 1, if the individual is updated using this condition. It will indicate that the virus is spreading by this individual.

Susceptible case: if r is in the range of $r \in [\frac{1}{3} \times \text{BR}_r, \frac{2}{3} \times \text{BR}_r]$, then the new updated elements will depend on the previous susceptible case. It resembles, the individual is effecting by the susceptible case.

$$S(x_i^j(t)) = x_i^j(t) + r \times (x_i^j(t) - x_i^m(t)) \quad (6)$$

where, $x_i^m(t)$ is the randomly selected susceptible case based on status vector, $m = i|Si = 0$.

Immune case: if r is in the range of $r \in [\frac{2}{3} \times \text{BR}_r, \text{BR}_r]$, then the new updated elements will depend on the previous immuned case. It resembles, the individual is effecting by the immuned case.

$$B(x_i^j(t)) = x_i^j(t) + r \times (x_i^j(t) - x_i^v(t)) \quad (7)$$

where, $x_i^v(t)$ is the randomly selected susceptible case based on status vector, $v = i|Si = 2$.

Again calculate the objective function values for the newly updated individuals, $f(x_i^j(t + 1))$.

Step 4: Update the herd immunity population matrix and status vector.

If $f(x_i^j(t + 1)) < f(x_i^j(t))$ then, the newly generated cases $x_i^j(t + 1)$ are replaced in $x_i^j(t)$. If the above condition is not satisfied and also if $S_j = 1$ then, it is observed that the individual is not getting better even after social distancing. Hence, the age vector should increase by 1. The status vector should be updated as follows:

$$S_j = \begin{cases} 1, & f(x_i^j(t + 1)) < \frac{f(x_i^j(t+1))}{\delta f(x)} \text{ and } S_j = 0 \text{ and } \text{is}_{\text{corona}}(x^j(t + 1)) == 1 \\ 2, & f(x_i^j(t + 1)) > \frac{f(x_i^j(t+1))}{\delta f(x)} \text{ and } S_j = 0 \end{cases}$$

where $\delta f(x)$ is the mean of objective function values of all the solution sets such as $\frac{f(x_i)}{\text{HIS}}$. Sometimes, the individual $x_i^j(t + 1)$ may perform well compared to $x_i^j(t)$ but

Table 1 Considered constraints for gain parameters

Constraint	Value
$[K_{p\max}, K_{p\min}]$	[0.0001, 0.01]
$[K_{i\max}, K_{i\min}]$	[0.0001, 0.01]
$[K_{d\max}, K_{d\min}]$	[0.00001, 0.001]

if is_{corona} is one, it should be considered as the infected worst case as it's updated from an infected case. So, S_j is updated as one and also age vector is increased by one simultaneously.

Step 5: Fatality condition.

If an individual $f(x_i(t + 1))$ of the current infected case ($S_j == 1$) and couldn't be improved after a certain limit of iteration as specified in Max_{age} . then that case is considered as died and removed from the HIP matrix. Another new value will be generated using Eq. 4 and replaced in place of the dead case and S_j and age_j vectors are set to be zero. This fatality condition will help to escape from the local optima.

Step 6: Stop criterion

CHIO will repeat step 3 to step 6 until the maximum iteration limit reaches. And at the stage of herd immunity, the infected cases will disappear and immune and susceptible cases dominates the overall population.

4.3 CHIO Application on Boost Converter

The gain parameters (K_p , K_i , and K_d) should be determined in order to fine-tune the controller. Hence, these parameters are considered as the population of individuals. The considered lower and upper bounds are shown in Table 1.

After every iteration, the minimum value obtained among all the solution sets will be considered the best value (as in the immune case). The maximum value will be considered the worst case (like the infected case). The population size is considered as 20 and maximum age limit is set to 15 and the basic reproduction rate is to 0.3 to get the better exploration.

5 Simulation Results and Discussion

The boost converter specifications are given in Table 2.

Table 2 Designed boost converter parameters

Parameters	Value
Input voltage (V_{in})	24
Output voltage (V_o)	48
Duty cycle (D)	0.5
inductance (L)	48 mH
Capacitance (C)	520 μ F
percentage of inductor current ripple	1%
percentage of capacitor voltage ripple	0.5%
Switching frequency (F_{sw})	10 kHz
Load resistance (R)	19.2 Ω

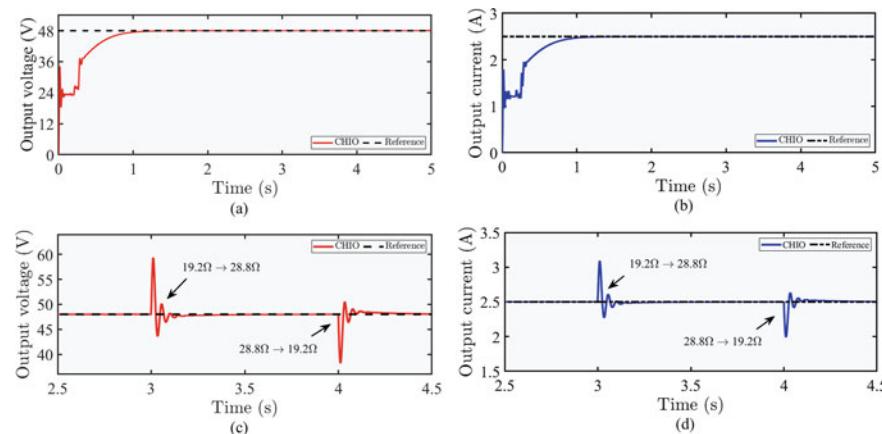


Fig. 4 **a** Output voltage response, **b** output current response, **c** output voltage response during load variations, **d** output current response during load variations

The simulation work is carried out in Matlab/simulink platform. The obtained PID gain parameters from CHIO are $K_p = 0.0073$, $K_i = 0.5113$ and $K_d = 8.0336e - 04$. The output voltage and current responses of converter during start-up is shown in Fig. 4a, b, respectively. It is observed that, the converter reached the desired output responses in approximately 1 s. Several tests were conducted to evaluate the tuned control action. The 50% of load is added to the converter after the system gets to stability. This creates transients in response and reaches desired output in 0.2 s as shown in Fig. 4c, d. To understand the controller performance during input change and reference change. The input is varied from 24 to 30 V. during this time, the output response reaches the desired voltage in 0.95 s and current response in 0.86 s as shown in Fig. 5a, b. The reference voltage is varied from 48 to 60 V. the output voltage and output current in Fig. 5c, d clearly indicates that the response reached the reference voltage successfully in 0.5 s and 0.6 s, respectively.

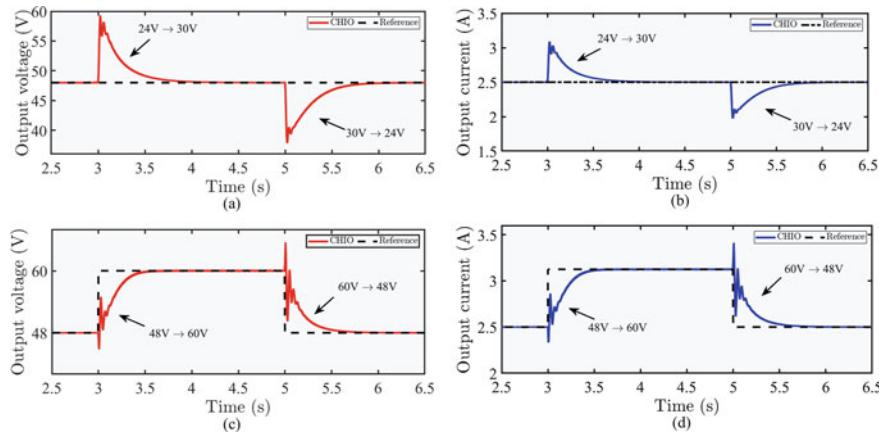


Fig. 5 **a** Output voltage response during input voltage variations, **b** output current response during input voltage variations, **c** output voltage response during reference voltage variation, **d** output current response during reference variation

6 Conclusion

In this paper, PID controller for a DC-DC boost converter is tuned using a new human-based meta-heuristic technique called coronavirus herd immunity algorithm. The obtained results are the evident that the converter shows satisfactory dynamic response of output voltage under wide variation in load, input and reference voltage with less settling time.

References

- Rai JN (2016) Design and analysis of DC-DC boost converter. *IJARI* 4:499–502
- Andrade AMSS, Schuch L, da Silva Martins ML (2018) Analysis and design of high-efficiency hybrid high step-up DC-DC converter for distributed PV generation systems. *IEEE Trans Ind Electron* 66(5):3860–3868
- Adnan M, Oninda M, Nishat M, Islam N (2017) Design and simulation of a DC-DC boost converter with PID controller for enhanced performance. *Int J Eng Res Technol (IJERT)* 6:27–32. <https://doi.org/10.17577/IJERTV6IS090029>
- Abushawish EA, Hamadeh M, Nassif A (2020) PID controller gains tuning using metaheuristic optimization methods: a survey. *Int J Comput* 14. <https://doi.org/10.46300/9108.2020.14.14>
- Pareek S, Kishnani M, Gupta R (2014) Optimal tuning of PID controller using meta heuristic algorithms. In: 2014 international conference on advances in engineering and technology research (ICAETR-2014), pp 1–5. <https://doi.org/10.1109/ICAETR.2014.7012816>
- Chopra V, Singla S, Dewan L (2014) Comparative analysis of tuning a PID controller using intelligent methods. *Acta Polytech Hung* 11:235–249
- Gandomi A, Yang X-S, Talatahari S, Alavi A (2013) Metaheuristic algorithms in modeling and optimization. <https://doi.org/10.1016/B978-0-12-398364-0.00001-2>

8. Jagindar Singh KSM, Elamvazuthi I, Shaari KZK, Lima FV (2016) PID tuning control strategy using Cuckoo search algorithm for pressure plant. In: 2016 6th international conference on intelligent and advanced systems (ICIAS), pp 1–6. <https://doi.org/10.1109/ICIAS.2016.7824127>
9. Singh K, Vasant P, Elamvazuthi I, Kannan R (2015) PID tuning of servo motor using bat algorithm. *Procedia Comput Sci* 60:1798–1808. <https://doi.org/10.1016/j.procs.2015.08.290>
10. Al-Betar MA, Alyasser ZAA, Awadallah MA et al (2021) Coronavirus herd immunity optimizer (CHIO). *Neural Comput Appl* 33:5011–5042. <https://doi.org/10.1007/s00521-020-05296-6>

Improvisation in Opinion Mining Using Negation Detection and Negation Handling Techniques: A Survey



Kartika Makkar , Pardeep Kumar , and Monika Poriye 

Abstract Negation handling is one of the challenges of natural language processing. Nowadays, negation detection and negation handling have significance in numerous fields like sentiment analysis, chat-bots, etc. In this digital era, a huge amount of data is generated from e-commerce websites, social media, etc. In order to perceive the sentiments from this data, an accurate classification of data is required, but in the opinion mining, stopwords are removed during preprocessing step and most of the negative words are removed from the text. Negation words are mostly responsible for flipping the polarity of sentences and if removed in preprocessing may cause incorrect classification of sentences or documents and result in inaccurate predictions. Negation handling is used in many domains and among all medical domains is a crucial domain. If it is not used, in the medical domain may produce wrong predictions about the disease of the patient and may prove to be fatal. Before the era of machine learning and deep learning, mainly rule based and lexicon approaches were used and these approaches were domain dependent. With the advent of machine learning and deep learning, more emphasis has been exercised on negation handling. Negations are mainly of two types implicit and explicit, and these negations can be handled using various approaches such as rule based, lexicon based, machine learning, deep learning, and hybrid. This article presents an overview of various negation detection and negation handling techniques existing in literature to get an insight into negation detection and negation handling.

Keywords Support vector machine (SVM) · Naïve bayes (NB) · Decision tree (DT) · Conditional random fields (CRF) · Bidirectional long and short-term memory (BiLSTM) · Multitask learning (MTL) · Single task learning (STL) · Negation cues and scope (NCS)

K. Makkar  · P. Kumar · M. Poriye

Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India

e-mail: sonikartika19@gmail.com

P. Kumar

e-mail: pmittal@kuk.ac.in

M. Poriye

e-mail: monikaporiye@kuk.ac.in

1 Introduction

Negation is a lingual phenomenon that flips the polarity of words or sentences in text. Due to these negations, the polarity of sentiments expressed using data changes the meaning of data. Negations are mainly of two types, viz., implicit and explicit. Explicit negations are morphological, syntactic, diminishers, and double negations as displayed in Fig. 2. While in implicit negation, it is not necessary that negation word is present but the meaning of sentence is negative. For example, “due to this act, it will be your last movie”, is an example of implicit negation as there is no negative word but the meaning of sentence is still negative. On the contrary “this car is not good” is an example of explicit negation [6]. In order to work with these negations, there are many approaches such as rule based, lexicon based, machine learning, deep learning, and hybrid. In prior research, mostly rule-based approach was used which is useful when the negations are static. But if the negations are dynamic, then this rule-based approach is not appropriate as these rule needs to be changed according to the negations. Rule-based approach is domain specific and it is more successful in medical domain, some of the rule-based algorithms for medical domain are NegEx [2], DEEPEN [13], etc. Another approach is lexicons based that is also used to assign polarity to words. Some of the lexicons are WordNet, SentiWordNet, etc. Recently, machine learning and neural network-based approaches are mostly used due to linguistic complexity. However, these approaches require less human intervention but pre-annotated data for negation detection and handling [18]. Main motive to write this article is to take an overview of negation detection and negation handling techniques so that a suitable model for negation detection and negation handling can be prepared by considering pros and cons of all the present techniques in literature. Negation detection is mainly the part of preprocessing step. In this step, various preprocessing techniques are applied such as number and alphabets removal, special character removal, punctuation removal, stopwords removal, case normalization, spelling correction, POS tagging, stemming, or lemmatization. When stopword removal is applied in data, then most of the negation words are removed. So, before applying negation handling, negative stopwords should be removed from the stopwords list. In preprocessing step, negations are also detected by using any of the approaches shown in Fig. 1. If negations are detected, then negation handling can also be performed otherwise feature engineering, train test split, classification, and evaluation can be performed on preprocessed data as shown in Fig. 1.

The whole paper scrutinizes into five sections, wherein Sect. 1 introduction part is given, Sect. 2 is the related work that gives an insight into work already done, Sect. 3 is further divided into three parts contains general framework, various types of negations and approaches for negation detection and handling, Sect. 4 includes research gaps, and finally, Sect. 5 gives the conclusion that provides the summary of survey and future work.

2 Related Work

Negation handling is one of the challenges of sentiment analysis that impacts the polarity of a sentence. Various traditional approaches discussed in the literature were unable to properly handle this challenge. One of the traditional approaches such as lexicon-based approach was unable to find proper negation window. To overcome this limitation, a new approach was introduced to determine the scope of negation. After determining the scope of negation, negation handling technique was introduced. This negation handling algorithm was based on dependency-based parse tree and it also takes into consideration the grammatical relation among the words and determines the words affected by negation. In order to determine the sentiment of sentence, this algorithm uses semantic disambiguation technique. In the proposed algorithm, firstly the parse tree was generated by considering the grammatical relations then the sentiment scores are assigned to the words generated by the parse tree using SentiWordNet. Thereafter, this tree is traversed using DFS and the negation words are identified at the leaf node and words affected by negation words. Then, score 0 is assigned to negation words and -1 to the words affected by negation, and finally, the polarity is calculated by adding all the polarities of words in a sentence. And the results depict that sentiment analysis after negation handling with word sense disambiguation gives 62.57% accuracy and the standard approach gives 56.35% accuracy [4].

Another similar approach for negation handling in unstructured data was introduced and this approach gives better accuracy as compared to the prior approach with an accuracy of 92%. In this research, morphological and syntactic negations were handled using prefix algorithm and dependency-based parse tree (DBPT). Here, the sentence is parsed using a statistical dependency parser then the dependency tree was generated using the Sanford parser. Thereafter, negation is explored using depth first search algorithm. If negation is detected in one of the nodes, then it inverts the polarity of the root word. Finally, the polarity of a full sentence is calculated using the mean of all the words in that sentence [16].

Most of the traditional approaches [9, 19] use punctuation marks and static windows to identify the scope of negations but these approaches were unable to handle negations properly due to variation in negation scope, linguistic features and word sense disambiguation. These challenges were identified by a new approach [7] that improvises the negation scope identification. Further, a linguistic feature-based negation handling technique was also proposed and the results depict that these approaches improve the negation scope detection and overall polarity of sentences. In this article, existing approaches such as static window for $k = 1$ to 5, i.e., (SW 1, SW 2, SW 3, SW 4, SW 5), punctuation marks based, and proposed methods were used to determine the scope of negations, and the results depict that proposed approach outperformed the existing methods with 83.3% accuracy.

A deep learning-based negation handling (NH) approach was introduced to get accurate results of sentiment analysis. Most of the NH [2, 13] are rule based (RB) and domain specific. Due to the complexity of negations, it is hard to create a general purpose model to handle these negations. Negations are of an explicit type or implicit

type and this article mainly works with an explicit negations. This article mainly deals with two components cue and scope, a cue is the negation word and scope is the parts of sentences affected by cue. In this model firstly, a cue is identified then the relationship between the cue and other words in a sentence is identified using BILSTM. Results depict that the BILSTM outperforms the SVM, HMM, CRF, and RB approaches with an *F1* score of 93.34% [18].

In another research, one more approach was introduced to identify the explicit negations. Here, a preprocessing algorithm was designed to determine the negation words. Then, the lemma of that word was identified and the _Neg word was combined with the lemma of negation. Thereafter, SA was done using various ML algorithms such as NB, SVM, Artificial Neural Network (ANN), and Recurrent Neural Network (RNN). The results depict that RNN gives the best accuracy of 0.9567. Here, various negations such as morphological, syntactical, and double negations were handled using the same approach [14].

Another sentiment analysis approach was proposed using Twitter dataset with improved NH. Various features such as morphological, lexicon-based (LB), n-gram, and POS-based features were used to train the classifiers, and mainly three classifiers (SVM, NB, and DT) were trained using different feature groups. Here, the static window approach for five words was used with linguistic features and, “_NEG” tag is attached to words that are under the scope. However, this tag is attached to only nouns, adjectives, adverbs, and verbs because negations do not affect all the words inside the scope. In order to handle these words, some negation exception rules were also proposed to get accurate classifications of twitter data. Finally, results depict that the performance of the SVM is the best among all classifiers with an average recall score of 67.5% and *F1* score of 69.5% [8].

In sentiment analysis (SA), polarity shifting is one of the factors responsible for the misclassification of text data. To overcome this problem, a three-stage approach named polarity shift detection, ensemble, and elimination (PSDEE) was introduced. A hybrid approach was proposed that was based upon RB and statistical approach for the detection and elimination of polarity shifts. Here, RB approach was used to detect negations, explicit, contrast, and a statistical approach was used to detect implicit negations. After the detection of various negations, a polarity shift elimination technique was applied, and finally, a polarity shift-based ensemble model was proposed that was the combination of several base classifiers and gives an improved output of 81.4% for unigram features and 82.9% for bigram features [20].

Negation is one of the important features while performing sentiment analysis on text data that can change the polarity of each sentence, which causes wrong predictions of sentiments in order to overcome this, NH function was created based on antonym of word in the WordNet using IMDB movie review. And the results proved that this approach increases the accuracy by 4–5% as compared to n-grams and deep learning (DL) methods [11]. Further, to enhance the accuracy of lexicon-based (LB) approach, an algorithm was proposed that was based on frequency score generation. In this research, some rules for negation and intensifiers were also implemented, and the results depict that the performance of SA by using this LB approach was enhanced [10] (Table 1).

Table 1 Summarized findings

Paper	Negation handled	Proposed technique	Results	Limitation
Barnes et al. [1]	Explicit negations	MTL approach consists of BiLSTM to deal with ncs	MTL approach outperforms STL	No confirmation about prediction of negative sentences
Cruz et al. [3]	Explicit negations	CS-SVM	CS-SVM outperforms NB and baseline with an <i>F1</i> score of 92.37% and 89.64%	Wrong scope detection
Diamantini et al. [4]	Explicit negations	Negation handling technique based on DBPT coupled with semantic disambiguation technique	The proposed algorithm with word sense disambiguation and negation handling gives 67% accuracy	Unable to work with sentences having intensifiers
Farooq et al. [7]	Explicit negations	NH method based on linguistic features	83.3% accuracy as compared to the existing approach based on the static window and punctuation marks	
Gupta and Joshi [8]	Explicit negations	A feature-based approach with corpus-based approach	SVM gives the best accuracy with feature-based model with 67.5% recall and 69.5% <i>F1</i> score	Unable to work with various explicit negations
Lilian et al. [12]	Explicit negations	Anti-negation algorithm	Anti-negation algorithm gives 96.2% accuracy	It works with limited languages, negations
Mehrabi et al. [13]	Explicit negations	DEEPEN	For the IU dataset and Mayo clinical dataset, <i>F1</i> score is 0.96%, 0.837%	DEEPEN is unable to address the concept affirmed by NegEx
Mukherjee et al. [14]	Explicit negations	Negation marking algorithm	Here, ANN gives 95.67% accuracy with the negation algorithm	
Singh and Paul [18]	Explicit negations	BiLSTM-based NH technique	Here, BiLSTM-based approach gives an accuracy of 93.34% accuracy as compared to ML approach	It requires pre-annotated dataset with negations and no sentiment prediction
Xia et al. [20]	Explicit, implicit	RB, statistic-based methods for PDEE as base classifiers for SA	Ensemble classifier gives an improved output for unigram and bigram features	
Zirpe and Joglekar [21]	Implicit, explicit	Stacking ensemble method	Stacking ensemble with NH gives 0.97%	Worked only with simple polarity shift structures

3 General Framework and Approaches for Detection and Handling of Various Types of Negations

3.1 General Framework for Negation Detection and Handling

In this framework, first data collection and preprocessing are accomplished and in order to enhance the accuracy of opinion mining negation detection (ND), then NH is performed by using various approaches as shown in Fig. 1 and then feature engineering, train test split, classification, and evaluation steps are performed to perform the SA.

3.2 Types of Negations

Negations are of two types explicit, and implicit. Explicit negations are further divided into syntactic, diminishers, morphological, conjunctions, and double negations. On the other hand, implicit negations are not having negative words but still the meaning of the sentence is negative shown in Fig. 2.

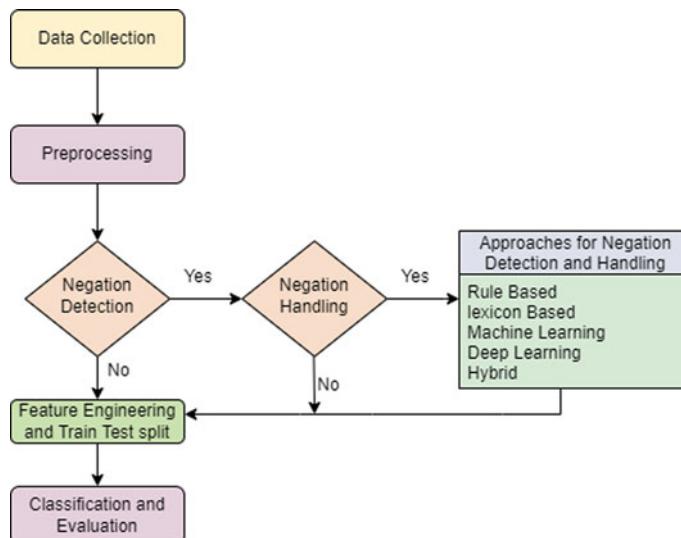


Fig. 1 General framework and approaches for detection and handling of various types of negations

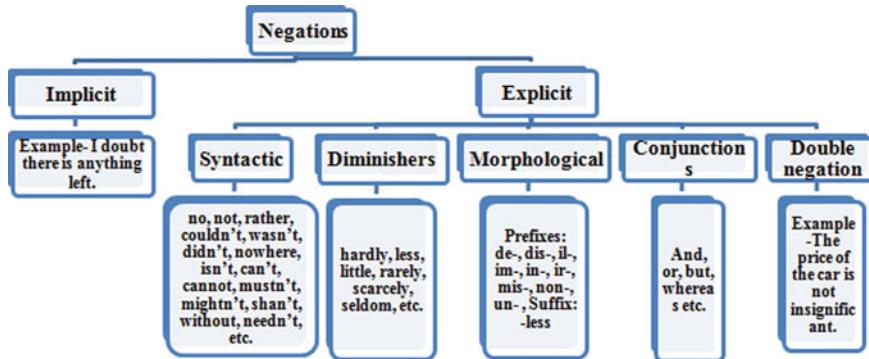


Fig. 2 Types of negations

3.3 Approaches for Negation Detection and Handling

Rule Based RB approach for negation detection handling (NDH) is implemented using regular expressions and requires human intervention for creation. These rules are created in such a manner that it covers maximum possible negations. But, still RB approach is unable to handle all types of negations as this approach is static in nature and negations are dynamic in nature. NegEx is one of the RB approaches used to detect negations in data. This algorithm is mostly used to detect negations in clinical data of NLP. But in some cases, this approach gives inaccurate output due to complex structure of some sentences. Due to this, NegEx is unable to find the contextual relationships among all the words in a sentence, which causes inaccurate detection of negations, and hence patients data. To overcome this challenge, a new algorithm named DEEPEN was proposed that takes into consideration dependency relations between negation words using the stanford dependency parser (SDP) and thus the performance of NegEx was improved [13].

Lexicon-Based Approach Lexicon-based (LB) approach uses predefined dictionaries to assign scores to a different word in a document. These dictionaries are created by determining different positive, negative, and neutral words, and their corresponding scores are already defined in those dictionaries. SentiWordNet is one of the lexicons that is used to assign polarities to words in a document and can be used to find the sentiments of that document or sentences. This approach can also be used in NDH to find the NCS [9]. LB approach can also be used to prepare lexicons of different languages, one such lexicon was also created for the Urdu language. At last, sentiment analysis was performed using this lexicon and RB approach [15].

Machine Learning Approach The machine learning (ML) approach helps to overcome the shortcomings of the lexicon and RB approach. Both of these approaches require human intervention for creations and updates. But, the ML approach does not

require much human intervention. Here, a huge amount of labeled data is required for training the model. In the case of the NH approach, pre-annotated dataset proved very helpful to work with negations. Using these datasets, various ML classifiers such as SVM and the CRF are trained using various features and help to determine NCS in a sentence. Finally, these cues and scope can be used to find the sentiments of the whole sentence [3, 5, 17].

Deep Learning Approach ML models require less human intervention as compared to RB and LB approach but, due to complexity in syntactical structures, machine learning approaches are not so success in creating a general purpose NH model. Also, it needs structured data and takes time to extract features from data. To overcome these shortcomings, a DL approach is used. In this approach, feature extraction step does not require human intervention. Here, negation cues are identified using pre-annotated negation dataset, and then BILSTM is used to find out the negation scopes. This BILSTM is trained using word level features and the results depicts that it gives higher accuracy as compared to CRF, HMM, and SVM in NH [18].

Hybrid Approach It is the combination of any two or more approaches among RB, LB, ML, and DL. Hybrid negation handling techniques also improve the accuracy of sentiment analysis [21]. Here, explicit negations were detected using RB approach and implicit negations were detected using statistical-based approach. Moreover, negations were also eliminated using antonym dictionary and finally the sentiments were predicted using LR, SVM, NB, and stacking ensemble. The results depict that the ensemble approach outperforms other algorithms.

4 Research Gaps

Most of the researchers discussed various approaches for NDH and it was revealed that there is limited work on implicit negations. For NH, there is limited work using ML and DL approaches can be used to refine the results. In literature, scope of negation was determined using BIO label but these BIO labels were not used to find the final sentiments of the text. This work can be extended as future work. Further, there is only a few pre-annotated datasets for NDH.

5 Conclusion and Future Work

NH is one of the hardest problems in natural language processing due to complex linguistic phenomena. This paper consists of an overview of various articles to find the best approach of the NH model. It was revealed that the RB and the LB approach are successful in limited domains. On the other hand, DL and ML models are not so much domain-dependent but both of these approaches require structured data.

However, the DL approach is more successful and requires less human intervention as compared to ML approach. In literature, it was also found that there is more work in explicit negations as compared to implicit negations. Implicit negations are more difficult to detect and handle. So, implicit NDH can be a vital research area. There is limited work on negation scope detection and NH in sentiment analysis using ML and DL too. This can also be future work for research.

References

1. Barnes J, Velldal E, Øvreliid L (2021) Improving sentiment analysis with multi-task learning of negation. *Nat Lang Eng* 27(2):249–269. <https://doi.org/10.1162/coli.08-012-R1-06-90>
2. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG (2001) A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 34(5):301–310. <https://doi.org/10.1006/jbin.2001.1029>
3. Cruz NP, Taboada M, Mitkov R (2016) A machine-learning approach to negation and speculation detection for sentiment analysis. *J Assoc Inf Sci Technol* 67(9):2118–2136. <https://doi.org/10.1002/asi.23533>
4. Diamantini C, Mircoli A, Potena D (2016) A negation handling technique for sentiment analysis. In: 2016 international conference on collaboration technologies and systems (CTS). IEEE, pp 188–195. <https://doi.org/10.1109/CTS.2016.0048>
5. Enger M, Velldal E, Øvreliid L (2017) An open-source tool for negation detection: a maximum-margin approach. In: Proceedings of the workshop computational semantics beyond events and roles, pp 64–69. <https://doi.org/10.18653/v1/W17-1810>
6. Evans JSB, Clibbens J, Rood B (1996) The role of implicit and explicit negation in conditional reasoning bias. *J Mem Lang* 35(3):392–409. <https://doi.org/10.1006/jmla.1996.0022>
7. Farooq U, Mansour H, Nongaillard A, Qadir MA et al (2016) Negation handling in sentiment analysis at sentence level. <https://doi.org/10.17706/jcp.12.5.470-478>
8. Gupta I, Joshi N (2021) Feature-based twitter sentiment analysis with improved negation handling. *IEEE Trans Comput Soc Syst* 8(4):917–927. <https://doi.org/10.1109/TCSS.2021.3069413>
9. Hogenboom A, Van Iterson P, Heerschap B, Frasincar F, Kaymak U (2011) Determining negation scope and strength in sentiment analysis. In: 2011 IEEE international conference on systems, man, and cybernetics. IEEE, pp 2589–2594. <https://doi.org/10.1109/ICSMC.2011.6084066>
10. Kulkarni DS, Rodd SF (2022) Towards enhancement of the lexicon approach for Hindi sentiment analysis. In: IOT with smart systems. Springer, Berlin, pp 445–451. https://doi.org/10.1007/978-981-16-3945-6_43
11. Lal U, Kamath P (2022) Effective negation handling approach for sentiment classification using synsets in the wordnet lexical database. In: 2022 first international conference on electrical, electronics, information and communication technologies (ICEEICT). IEEE, pp 01–07. <https://doi.org/10.1109/ICEEICT53079.2022.9768641>
12. Lilian JF, Sundarakantham K, Shalinie SM (2021) Anti-negation method for handling negation words in question answering system. *J Supercomput* 77(5):4244–4266. <https://doi.org/10.1007/s11227-020-03437-1>
13. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, Beesley C, Dexter P, Schmidt CM, Liu H et al (2015) Deepen: a negation detection system for clinical text incorporating dependency relation into negex. *J Biomed Inf* 54:213–219. <https://doi.org/10.1016/j.jbi.2015.02.010>
14. Mukherjee P, Badr Y, Doppalapudi S, Srinivasan SM, Sangwan RS, Sharma R (2021) Effect of negation in sentences on sentiment analysis and polarity detection. *Procedia Comput Sci* 185:370–379. <https://doi.org/10.1016/j.procs.2021.05.038>

15. Mukhtar N, Khan MA (2020) Effective lexicon-based approach for Urdu sentiment analysis. *Artif Intell Rev* 53(4):2521–2548. <https://doi.org/10.1007/s10462-019-09740-5>
16. Pandey S, Sagnika S, Mishra BSP (2018) A technique to handle negation in sentiment analysis on movie reviews. In: 2018 international conference on communication and signal processing (ICCS). IEEE, pp 0737–0743. <https://doi.org/10.1109/ICCS.2018.8524421>
17. Reitan J, Faret J, Gambäck B, Bungum L (2015) Negation scope detection for twitter sentiment analysis. In: Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 99–108
18. Singh PK, Paul S (2021) Deep learning approach for negation handling in sentiment analysis. *IEEE Access* 9:102579–102592. <https://doi.org/10.1109/ACCESS.2021.3095412>
19. Wilson T, Wiebe J, Hoffmann P (2009) Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Comput Linguist* 35(3):399–433. <https://doi.org/10.1162/coli.08-012-R1-06-90>
20. Xia R, Xu F, Yu J, Qi Y, Cambria E (2016) Polarity shift detection, elimination and ensemble: a three-stage model for document-level sentiment analysis. *Inf Process Manag* 52(1):36–45. <https://doi.org/10.1016/j.ipm.2015.04.003>
21. Zirpe S, Joglekar B (2017) Negation handling using stacking ensemble method. In: 2017 international conference on computing, communication, control and automation (ICCUBEA). IEEE, pp 1–5. <https://doi.org/10.1109/ICCUBEA.2017.8463946>

Localization of the Closed-Loop Differential Drive Mobile Robot Using Wheel Odometry



Gurpreet Singh  and Vijay Kumar 

Abstract A differential drives mobile robot is designed and developed for an obstacle-free environment. Both motors are controlled individually to reach the desired location using inverse kinematics. The robot's current state is estimated by applying forward kinematics after getting feedback from the magnetic encoders. High-level architecture is built using a 64-bit microprocessor for processing, and low-level architectures are developed using 8-bit dedicated microcontrollers for individual motors. The trajectory was plotted on the Raspberry Pi user interface, and the trajectory data was also saved. The wheeled robot was driven on the floor, and the trajectory followed in actual is compared with the trajectory on the screen. Simulation and experimental results were compared and verified.

Keywords Localization · State estimation · Wheel encoders · Wheel odometry · Differential drive mobile robot · Inverse kinematics

1 Introduction

In the last two decades, autonomous vehicles (AV) and automated guided vehicles (AGVs) are the most promising area of research and will be an area of great interest for researchers for decades [1, 2]. Optimists believe that automated vehicles will soon be trustworthy and affordable enough to replace the majority of human driving, leading to huge savings and benefits, based on prior technological developments such as desktop computers, smartphones, and digital cameras. Mobile robots can solve many problems and can perform many useful tasks without any human interference like cleaning, grass cutting, inspection, space exploration, sanitizing, vehicle driving, etc. The differential drive mobile robot, in which two different motors are operated independently, is the most popular and efficient drive mechanism and structure that

G. Singh  · V. Kumar

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
e-mail: gurpreet.singh@chitkra.edu.in

has been presented by numerous researchers. Due to the perfect rolling limits on a wheel motion, mobile robots with such motion control are a prominent example of non-holonomic systems (no longitudinal or lateral slipping).

A dynamic and kinematic model is introduced and simulated to control a differential drive mobile robot (DDMR) [3], and mathematical model and control strategies for two-wheel DDMR were proposed [4]. Two independent sets of wheel and motor is placed in front or rear part of the robot chassis and it is the simplest and commonly used platform. Modelling and control of this DDMR has been investigated [5]. Additional free wheel was employed to maintain the stability of three-wheeled and four-wheeled mobile vehicle. [6]. A four-wheeled mobile robot with two castor wheel and two motorized wheels is introduced. The forces generated by the two drive wheels cause the castor wheels to automatically and freely align on the path [7]. The H-bridge driver circuit was used to control the speed of the motors independently [8]. A closed-loop control system was proposed to control the position and orientation of the mobile in tracing the given trajectory [9]. A prototype of a wheeled mobile robot was developed to move to a new position based on input commands. A closed-loop control system was implemented for various navigation, path planning, and obstacle anticipation [10]. A dynamic model of a mobile robot (MR) was proposed to improve the performance and stability of the robot with two motorized wheels by using an optimal control strategy [11]. Other approaches utilizing complaint linkage were utilized to create the kinematics and control of mobile robots with several degrees of freedom [12]. Inverse and forward kinematics was implemented on a specially designed and developed differential drive of the mobile robot to compare the simulation and experimental results [13].

In the present study, forward kinematics and reverse kinematics are implemented on laser cut differential drive mobile robot. Magnetic encoders are coupled with the motor to obtain the distance travelled by each motor for the localization of the mobile robot. The mobile robot can reach any position in an obstacle-free 2D Cartesian environment.

2 Experimental Setup

A differential derive mobile robot platform is designed and developed using a laser cutting machine to ensure the alignment of the motors and castor wheel. For the localization of mobile robot using wheel odometry, the low-level and high-level architecture hardware is assembled on the mechanical structure. The low-level architecture hardware includes gear motors with magnetic encoders, wheels, motor drivers, and microcontrollers (Arduino Nano), whereas a microprocessor (Raspberry Pi), keyboard/joystick terminal and remote station are the high-level architecture components. Dimensions of the platform were fixed after keeping a sufficient gap between both encoders so that the magnetic field of one encoder does not affect the output of the other encoder. Figure 1 shows the hardware setup indicating all the components mounted on the platform. The data is exchanged between low-level and high-level

hardware through I2C communication in which Raspberry Pi is acting as master and two Arduino Nano are the slaves. The logic level converter is used between master and slave as Raspberry signal works on 3.3 V, whereas Arduino works on the 5 V signal. Figure 2 shows the block diagram of the connections of the different components of the mobile robot.

Software algorithm for low-level architecture is written to receive the incoming data from magnetic encoder connected with motors. A dedicated microcontroller is used to acquire the pulses generated by the Hall effect-based magnetic disc encoder to calculate the revolution per second (RPS) of the motor. The microcontrollers are

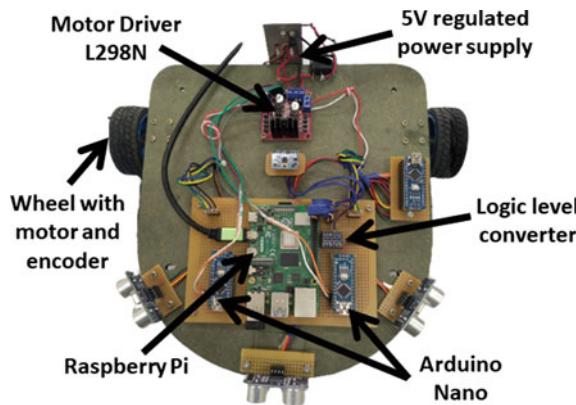


Fig. 1 Hardware setup of differential drive mobile robot

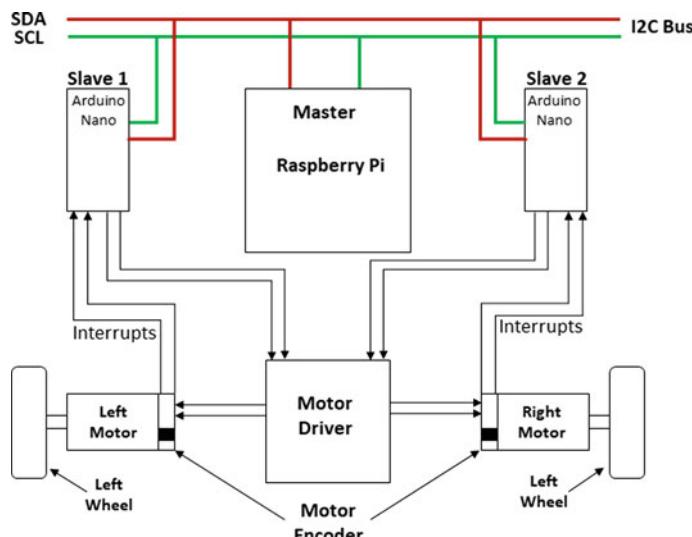


Fig. 2 Hardware block diagram

also responsible for transmitting the data to the microprocessor for further processing and giving the pulse wave modulation (PWM) signal to the motors.

The number of pulses per revolution of the wheel is acquired by the microcontroller, and RPS is calculated and fed to the microprocessor. After processing based on the inverse kinematics, the microprocessor sends the desired RPS value of each wheel to the corresponding microcontroller. The microcontroller sends the calculated PWM signal to the motor driver and the motor actuates accordingly.

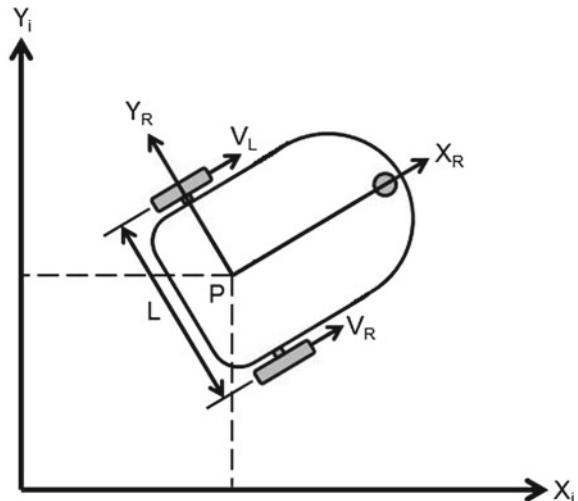
3 Kinematics of Differential Drive Robot

3.1 *Odometry and Kinematics*

Figure 3 illustrates the two-wheeled differential drive mobile robot in the Cartesian coordinates. Let $[X_i, Y_i]$ is the inertial reference frame and $\{X_R, Y_R\}$ is robot frame. The mobile robot under consideration is powered by two independent motors in which forward motion is achieved by rotating both motors in the forward direction at the same velocity. Turning of mobile robot in one direction is achieved by increasing the speed of the wheel on the other side. This robot can take a turn at the centre of the rear axis with zero turning radius by rotating the wheels in opposite directions at the same speed. A castor wheel at the front centre is provided for the stability of the mobile robot. A robot's movement is contributed by each wheel, which also places constraints on the robot's motion. The robot's wheels are assumed not to skid.

L is the distance between the wheels, and D denotes the diameter of each wheel. V_R and V_L represent the velocity of the right and left wheel, respectively. Figure 4

Fig. 3 Illustration of mobile robot in Cartesian coordinate system



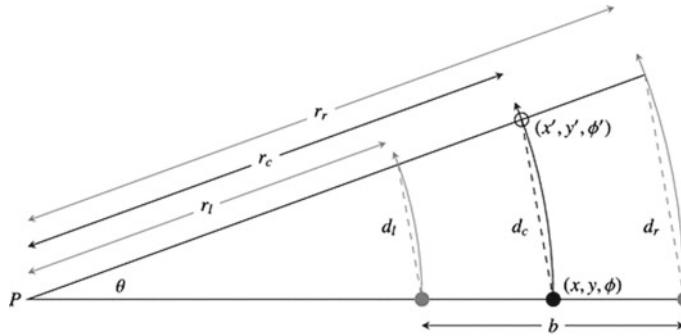


Fig. 4 Geometry of a left turn by a robot with two wheels [14]

shows the odometry geometry for a short period in which (x, y, ϕ) is the initial position and (x', y', ϕ') is the new position of the robot which is required to be estimated. The baseline is equal to the distance between the wheels, i.e. L . Distance travelled by the robot d_c can be calculated using the Eq. (1)

$$d_c = \frac{d_l + d_r}{2} \quad (1)$$

θ is the angle of rotation of the robot about the point P which is the centre of the arc and can be calculated using Eq. (2), and ϕ is the angle of rotation of the robot about its centre and can be computed using Eq. (3).

$$\theta = \frac{d_r - d_l}{2} \quad (2)$$

$$\phi' = \phi + \theta \quad (3)$$

The new position of the mobile robot can be estimated by using Eq. 4 and 5

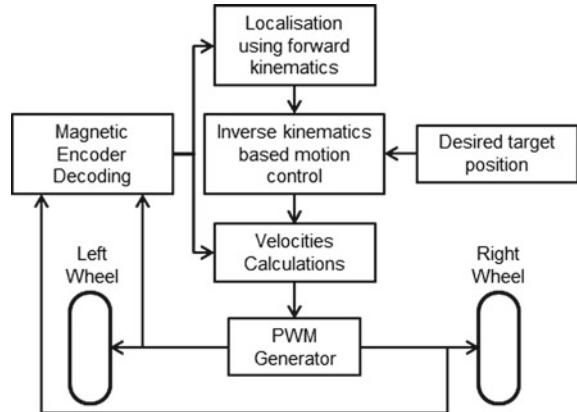
$$x' = x + d_c * \cos\phi \quad (4)$$

$$y' = y + d_c * \sin\phi \quad (5)$$

Therefore, the distance travelled by the left and right wheels is necessary to calculate the robot's present position. It is necessary to know the wheels' revolutions per minute (RPM) to compute the distance covered by both wheels. Distance travelled can be computed using Eq. (6)

$$\text{Distance travelled} = \frac{(3.14 * \text{wheel diameter}) * \text{RPM}}{60} \quad (6)$$

Fig. 5 Block diagram of proposed kinematics-based motion controller



3.2 Kinematic Equations

Figure 5 shows the block diagram of the strategy applied for the application of inverse kinematic on the differential drive mobile robot which is under consideration. It is required to calculate the velocities of each wheel to reach the set target position. Error is calculated using the difference between the set position and the current position, and the error is then corrected proportionally, as in Eq. (7). Required velocities of the right and left wheel are calculated using Eqs. (8) and (9), respectively.

$$W = K_p * \text{error} \quad (7)$$

$$V_R = \frac{(2*(d_c) + W*L)}{\text{tyre diameter}} \quad (8)$$

Angular velocity of the right and left wheel is computed using Eqs. (10) and (11), respectively, based on V_R and V_L . Equations (12) and (13) calculate the required travelling distance of each wheel. Desired RPM for both wheels is calculated by using Eqs. (14) and (15). Finally, the PWM signal is generated based on the required RPM, and the PWM signal is sent to the motors by the corresponding microcontroller. Both motors are also tested and calibrated for RPM and PWM values to investigate the RPM of the motor at different PWM values.

$$V_L = \frac{(2*(d_c) - W*L)}{\text{tyre diameter}} \quad (9)$$

$$\omega_R = \frac{2V_R}{\text{tyre diameter}} \quad (10)$$

$$\omega_L = \frac{2V_L}{\text{tyre diameter}} \quad (11)$$

$$D_{RR} = \frac{2\omega_R}{\text{tyre diameter}} \quad (12)$$

$$D_{RL} = \frac{2\omega_L}{\text{tyre diameter}} \quad (13)$$

$$\text{RPM}_R = \frac{D_{RR} * 60}{3.14 * \text{tyre diameter}} \quad (14)$$

$$\text{RPM}_L = \frac{D_{RL} * 60}{3.14 * \text{tyre diameter}} \quad (15)$$

4 Results

The most challenging but ultimately rewarding task was integrating the programme into the actual robot. Initially, the mobile robot was mounted on the stand and wheels were not touching the ground and Raspberry was connected to the screen. The simulation results were observed on screen in graph, and data is stored in.csv format. The simulation results are summarized in Fig. 6. After many trials, the robot was allowed to move on the ground and connected to the desktop through a VNC viewer. This enabled the simulation visualization on the desktop and at the same time robot following the desired trajectory on the ground. Many trials were made by giving the desired destination points to the mobile in Cartesian coordinate in all quadrants.

Initially, the robot is placed in positive x-direction and then the robot can navigate to the desired position. In the starting, the robot starts with an equal PWM signal for both motors and then depending on the required turn the RPM of both motors increases or decreases accordingly. Velocities of both wheels are controlled by a microcontroller depending upon the output of inverse kinematics to reach the target position. Simulation and experimental results were compared and found to be satisfactory with error less than 5%.

5 Conclusion and Future Scopes

In the present study, a platform for the mobile robot is designed to accommodate the two motors with encoders, castor wheel, and other electronics components, i.e. microprocessor, microcontroller, motor driver, logic level converter, and battery. The differential drive arrangement is selected for study as it does not require any special mechanism for steering purposes. The mobile robot can reach any target position in a

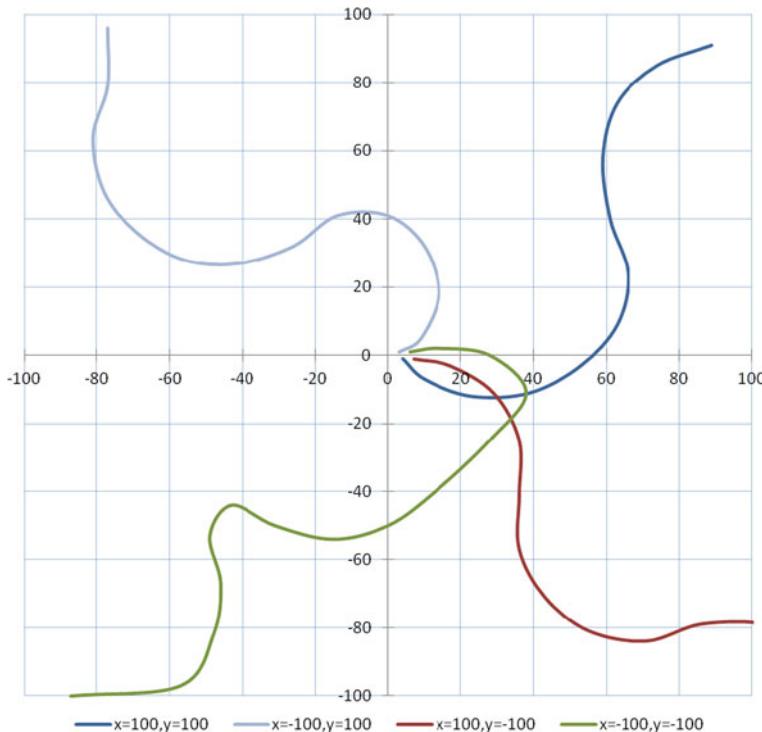


Fig. 6 Summary of simulation result observed on screen

two-dimensional Cartesian coordinate. The mobile robot is capable of localizing itself with the help of forward kinematics. The magnetic encoder provides the pulses to the microcontroller; microcontrollers calculate the RPS and send it to the microprocessor. The microprocessor calculates the current position and orientation of the mobile robot.

Inverse kinematic equations are used to calculate the desired velocities of each wheel at every instant to reach the destination. Then, the desired velocities are converted to RPS and RPS to PWM signal. The corresponding PWM signal is sent to the motors. It is observed from experimentation that the mobile robot can match the trajectory obtained in simulation.

In future study, more sensors like inertial measurement unit, and range sensors will be installed so that better state estimation can be achieved along with obstacle avoidance.

References

1. Hyla P, Szpytko J (2017) Automated guided vehicles—the survey. *J KONES Powertrain Transp* 24(3):101–110
2. Parekh D, Poddar N, Rajpurkar A, Chahal M, Kumar N, Joshi GP, Cho W (2022) A review on autonomous vehicles: progress methods challenges. *Electronics* 11(14):2162
3. Salem FA (2013) Dynamic and kinematic models and control for differential drive mobile robots. *Int J Curr Eng Technol* 3(2):253–263
4. Nitulescu M (2007) Solution for modeling and control in mobile robot. *CEAI* 9(3):43–50
5. Krichel SV, Agrawal SK, Sawodny O (2009) Modeling and control of two-wheeled vehicles using active caster wheel. In: IEEE/ASME international conference on advanced intelligent mechatronics, pp 1750–1756
6. Ailon A, Berman N, Arogeti S (2004) Robot controller design for achieving global asymptotic stability and local prescribed performance. *IEEE Trans Robot Autom* 20(5):790–795
7. Andrea B (1991) Modeling and control of non-holonomic wheeled mobile robots. In: Proceedings of IEEE international conference on robotics and automation, Sacramento, SUA, pp 1130–1135
8. Gupta V (2010) Working and analysis of the H—bridge motor driver circuit designed for wheeled mobile robots. *IEEE Int Conf Adv Comput Control (ICACC)* 3:441–444
9. Nitulescu M (2008) Theoretical aspects in wheeled mobile robot control. In: IEEE international conference on automation quality and testing, robotics proceedings vol 2, pp 331–336
10. Velazquez R, Ekuakille AL (2011) A review of models and structures for wheeled mobile robots: four case studies. In: IEEE the 15th international conference on advanced robotics, pp 524–529
11. Morales MCGC, Alexandrov VV (2012) Dynamic model of a mobile robot with two active wheels and the design an optimal control for stabilization, IEEE (CERMA), pp 219–224
12. Borenstein J (1995) Control and kinematic design of multi-degree-of-freedom mobile robots with compliant linkage. *IEEE Trans Robot Autom* 11(1):21–35
13. Singh R, Singh G, Kumar V (2020) Control of closed-loop differential drive mobile robot using forward and reverse kinematics. In: Third international conference on smart systems and inventive technology (ICSSIT), pp 430–433
14. Ben-Ari M, Mondada F (2018) Robotic motion and odometry. In: Elements of robotics. Springer, Cham

Analysis of Effectiveness of Online Classes During COVID



Disha Sriram, Lavanya Sanjay, Neha Nayak, Sathwik Sathish, Ashwini Kodipalli, and P. N. Anil

Abstract The motivation behind this paper is to provide knowledge with regard to the impact on students due to the outbreak of COVID-19. As rapid and as hopeful as the year 2020 started, we soon fell into lockdown to prevent further spread of the virus. COVID-19 is a respiratory illness in humans caused by SARS-CoV-2 virus. Education system became one among the affected sectors, majorly due to the transformation of the learning-teaching platform to online mode. Online mode put forth a number of disadvantages related to network issues, unavailability of resources, miscommunication, and a lot more. In addition to this, it took a toll on one's mental wellness. This study has complied the literature approach, using analytical and comparative techniques as basic research methods. The main objective of this research is to analyze the effectiveness of online classes and to find out which platform was preferred by majority of learners and the impact of online classes on the mental health. To analyze this, a survey through Google Forms which contained a set of precise questions was conducted. The responses put forth a night and day difference. For this research work, a total of 158 responses were collected, widely leaning toward offline mode. This signifies the importance of traditional education format. Since the questions were curated squarely targeting the course outcomes and the student expectations, the research work obtained crystallized results. Machine learning algorithms are applied to cross verify the data wherein, among all the classifiers, logistic regression is used and obtained an accuracy of 85.95%. To prevent overfitting of model, cross validated the data by splitting obtained dataset into 5-folds with k-fold accuracy of 71.678%

Keywords Online · Offline · Blended · COVID-19 · Technical · Computer-based network · Social media · Depression · Anxiety peer pressure

D. Sriram · L. Sanjay · N. Nayak · S. Sathish · A. Kodipalli (✉)

Department of Artificial Intelligence and Data Science, Global Academy of Technology, Bangalore, India

e-mail: dr.ashwinik@gat.ac.in

P. N. Anil

Department of Mathematics, Global Academy of Technology, Bangalore, India

1 Introduction

Education has an important role in molding a person. It makes a person self-reliant and knowledgeable and helps to contribute to the nation. Education is a life-long learning process where an individual becomes a better person every day. The COVID-19 pandemic which occurred in 2020 made a huge impact on education. This resulted in closing schools and colleges globally. As a result, education has changed drastically, with e-learning or the online learning.

Online learning deals with electronic applications and devices. It requires Internet connections, devices like smartphones, laptops, pc, etc. The learning takes place outside the classroom through digital medium [1]. The faculty delivers the lessons through animation, videos, and images which makes the class interesting.

The COVID-19 bought tremendous changes to the learning across India. All the classes and examinations were conducted online. The Government of Karnataka took the decision of cancelling the 10th exams during the 2020 pandemic. Reports tell that over 20,000 students skip SSLC exams on Day 1 [2]. Most of the students did not appear to the exam due to lack of preparation and due to family emergencies. Students might face a lot of troubles like depression, anxiety, peer pressure, and many more. This causes immense loss of concentration, focus, interest and leads in the downfall of the academic performance. Students do not get a chance to interact with their teachers and friends in person which causes lack of confidence and communication.

This refers to a traditional type of learning where the students learn face to face [3]. It is also called as classroom learning, where the students interact with teachers and students. Doubts regarding any topic can be cleared on spot. This learning enhances communication, discipline, and many more aspects of the students. Due to the pandemic, schools and colleges had to take a decision to conduct the classes online. Students and faculty did face a few difficulties in the beginning such as Internet issues, lack of devices, lack of resources, and many more.

Through offline learning, students got extra time to prepare and had access to immense resources over the Internet. This helped them gain more knowledge and helped them to work on the fields of their interest. This provided a completely different insight about learning and education.

The organization of the paper is as follows: Sect. 2 describes the detailed literature survey. Section 3 provides the methodology. Section 4 explains the results. Section 5 concludes the paper.

2 Literature Survey

According to the study by Mondol et al. [4], the quality of teaching imparted to the students depends on the teacher's hold on managing online classes efficiently. The teacher should be well versed to keep the class active, engaged, and motivated throughout. Although offline classes are preferred by most teachers and students, the

teachers should be trained to handle online classes, as a part of their job training to handle classes during such unprecedented situations like COVID as education and imparting knowledge cannot be hindered by external factors. [5, 6].

The research by Baun et al. [7] points that there are various challenges faced by the students during online classes like lack of strong network mostly in rural areas, lack of knowledge on how to handle software like Zoom or Teams, reduced drive to learn due to absence of human intervention to prevent distraction and a general sense of anxiety that thrives in the students' minds generally in online classes. From the teacher's perspective, although it is quite easy to handle theory classes using well-devised notes and pre-planned presentations or ppts, it is quite a hard task to handle practical or mathematical subjects as they seek more attention from teachers and are better learnt offline. [8, 9].

Szopinski et al. [10] show that the onset of the pandemic had institutions all across the world looking for the best and most effective mode of teaching. A number of students agree that online classes do have its own perks. As the teaching continued online, most teachers incorporated sources outside the given curriculum for better understanding by students. Teachers used various sources available online to make students analyze and understand the topic that resulted in better visualization and interpretation by students. Online education also has some major benefits such as saving time, energy and fuel, and investing these resources into much more productive work by both students and teachers. Also, clarifying doubts becomes easy by scheduling meetings with the teachers. Reviewing and revising topics become easier as students can simply rewatch the recorded lectures [11, 12].

According to Xing et al. [13], most students have adapted offline learning as they are much more effective in terms of building skills like communication, teamwork, ethics, and general behavior. It is quite important for a student to enjoy their campus life, build good social skills, and develop their personality. While online mode focuses on just learning the course, offline mode also deals with life practically and helps develop time management skills, friendships, and discipline. Also, the effectiveness of offline teaching is higher in terms of grades.

Chawkase et al. [14] point out that offline learning methods are usually based on syllabi laid out by an institution for the student, which may restrict the student's learning potential, whereas online learning methods offer a wider choice to the students wherein, enthusiastic students can choose the depths up to which they would like to explore and understand a topic they have chosen out of their own free will, which keeps them more driven and motivated to learn the topics they aspire to learn about, or are too passionate to work on. In most institutions of higher learning, students are in the quest to find their true calling, something they are truly passionate about which is catalyzed by e-learning [15, 16, 17].

Al-khresheh [18] states that both online and offline modes come with its own perks. So, by using a blended mode of offline and online techniques, when the curriculum is completed, it delivers more effective Program Outcomes, for the holistic growth and all-round development of a student. This can happen by devoting time, energy, resources toward building an effective e-learning classroom model making use of the latest available technology [19, 20, 21, 22].

3 Methodology

The objective of this survey is to know the opinion of the students as to how their experience was during online classes. To collect data, we asked a few questions using google forms as a review from the students. The questions included are follows:

1. Was the faculty's presentation of the topics clear during online classes as compared to traditional face to face classes? (Presentation, audibility, and vocal clarity)
2. How do you rate the quality and clarity of the contents delivered through online mode as compared to traditional offline classes?
3. Do you agree that online classes helped in developing critical thinking skills?
4. Rate the effectiveness of offline (Conventional) classes in terms of increasing technical skills
5. Do you feel Internet connectivity strength determines effectiveness online learning activities?
6. How do you rate interaction with teachers during online classes as compared to traditional offline classes?
7. How effective was the assessment in online mode?
8. Recommended mode of teaching:
 - OFFLINE • ONLINE • BLENDED

After collecting the data from students, the data was converted into a csv file and basic data cleaning operations like trim were performed on the dataset using Excel. Then analyzed the dataset by applying data science techniques and also applied certain machine learning algorithms. Then, generated graphs using Excel and Python libraries like matplotlib, seaborn, and interpreted the results. The detailed framework is shown in Fig. 1.

4 Results

The sudden decision to shift the mode of teaching to online after complete shutdown of all the educational institutions across the world have impacted the students understanding, effectivity, and reasonability. When a survey was conducted whether online class is a complete solution in the emergency arising from crisis like COVID-19, the responses were fairly divided. However, most of the respondents were positive in their take and a number of students were not clear about it.

1. For the questions posed about the clarity of presentations given by the teachers during online classes were clear in terms of presentation, audibility, vocal clarity when compared to traditional offline classes, it was found that most of the students preferred a blended mode (combination of both online and offline modes) of teaching. Out of the 158 responses received, 39% of students had moderately found it extremely clear. While 14% students did not agree upon the fact that it

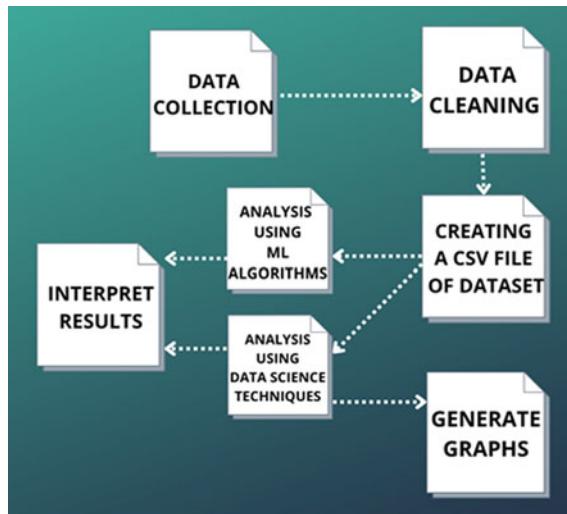


Fig. 1 Proposed framework

was clear, it was also found out that 7% students found the presentation totally unclear during the online classes as shown in Fig. 2.

2. For a query done about clarity and quality of the education through online classes during and post-pandemic, it was again found out that most of the student's preferred blended mode of education. Out of the total responses recorded, it was found out that 36% students found the classes to be moderately clear and

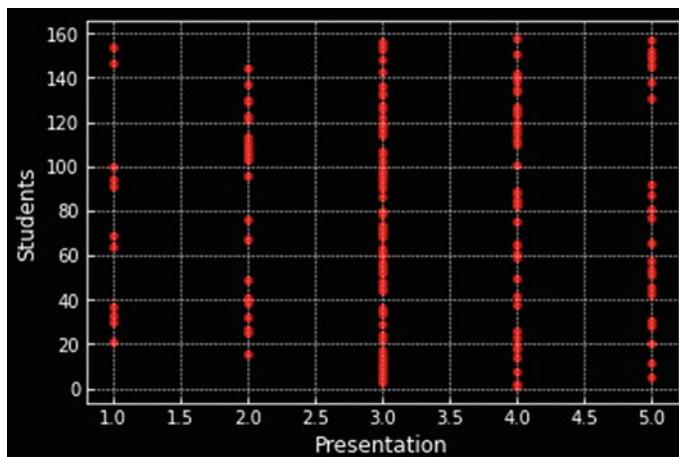


Fig. 2 Presentation of the concepts

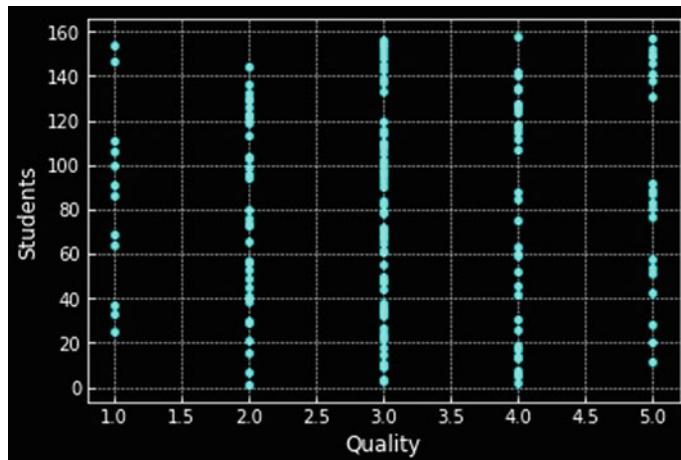


Fig. 3 Quality and clarity of content

preferable. Around 13% students found the contents of the online class to be clear while very few among responders found it to be unclear as shown in Fig. 3.

3. Out of the responses recorded for a question whether online classes helped the students elevate their critical thinking, it was found out that 1/4th of the students felt the online classes to be unhelpful in cultivating their critical thinking. This number summed up to around 25% of the students. While a moderate amount of people found the online classes helping them improve their critical thinking which was around 25%, handful of people found the online mode of teaching to be helpful as shown in Fig. 4.

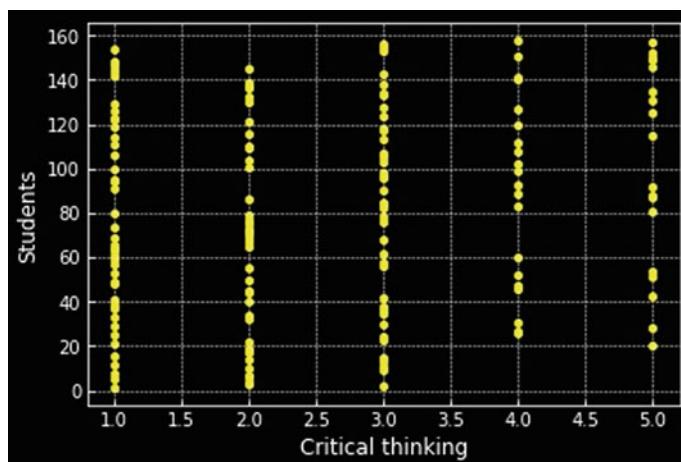


Fig. 4 Development of critical thinking skills

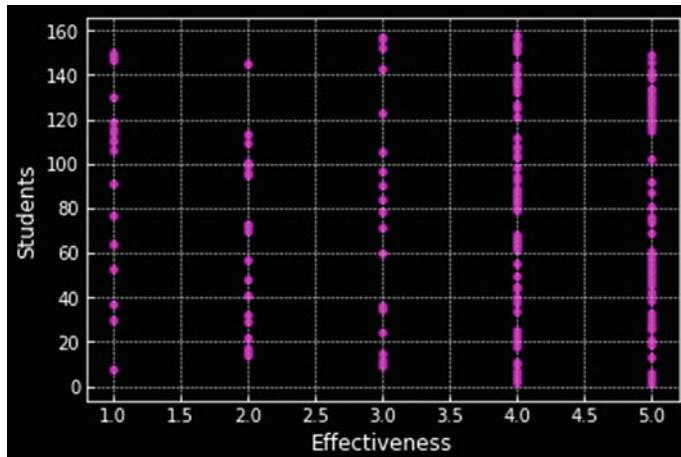


Fig. 5 Effectiveness of offline classes in developing technical skills

4. For the responses recorded whether offline (traditional classes) were more effective and impactful when compared to online classes, the majority of the students have been convinced that offline classes are the most effective. This sums up to around 31%. However, a number of 10% is found to disagree which shows the current understanding level of the students regarding the offline mode of classes while a minimal set of people around 11% moderately agreed on the fact that offline classes are more effective as shown in Fig. 5.
5. The Internet is very beneficial for students in all aspects. It serves as a teacher to students from where you can ask everything and you can fetch an answer for almost anything. So, when students were asked about how the Internet connectivity impacted their learning during the online classes, it was unfortunate that around 26% of people's poor Internet connectivity had a negative impact on their learning. While 24% among the lot felt the strength of Internet doesn't have much impact on learning, 25% of people almost disagreed to the fact that strength of Internet had an impact on their learning as shown in Fig. 6.
6. As the face-to-face interaction between students and teachers at an online learning academy is limited, it's important that there is an established relationship between teachers and students. Two-way communication is essential during online classes. On the survey conducted about how interactive classes were during the online phase, it was found out that around 1/3rd of the people found the classes to be moderately interactive. It was found out that very few among the lot agreed to the fact it was very much interactive and around 16% of the students found the online classes to be least interactive as shown in Fig. 7.
7. Testing a student through assessment is proved to be an effective way of testing one's learning. Due to restrictions during the pandemic, the ways of assessing a student diminished. Assessing the student online was the one way left. So, when asked about how many students found the online way of assessment effective,

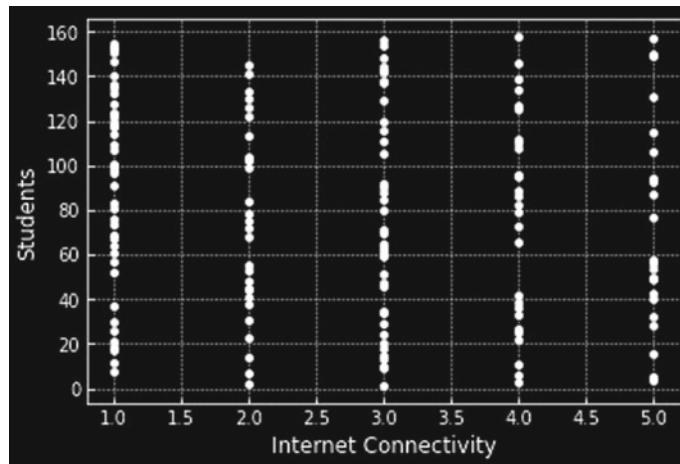


Fig. 6 Internet connectivity and its effects on online learning

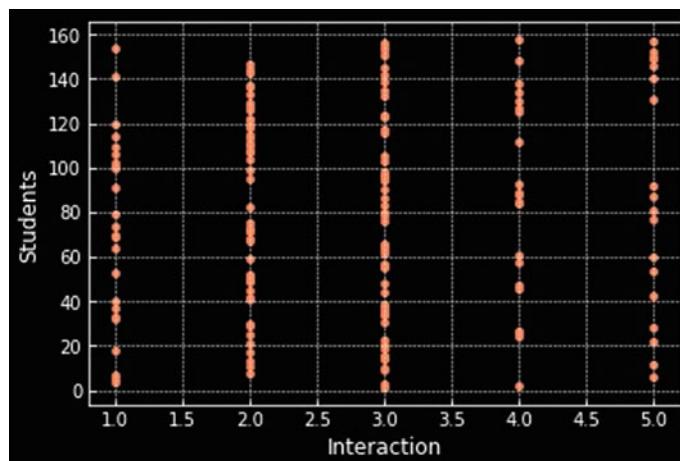


Fig. 7 Rate of interaction with teachers

the responses recorded portrayed that around 32% of students were moderately satisfied with this mode of assessment. While 16% of the students found the mode of assessment to be least affective, around 10% of people found the mode to be really effective as shown in Fig. 8.

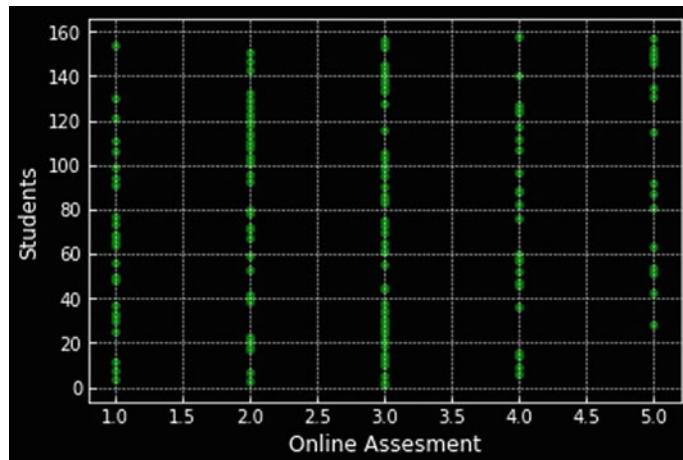


Fig. 8 Effectiveness of online assessment

Fig. 9 Favored learning platform

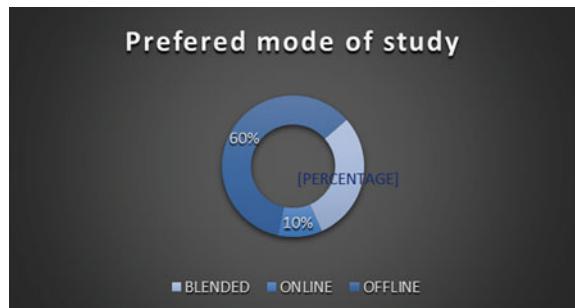
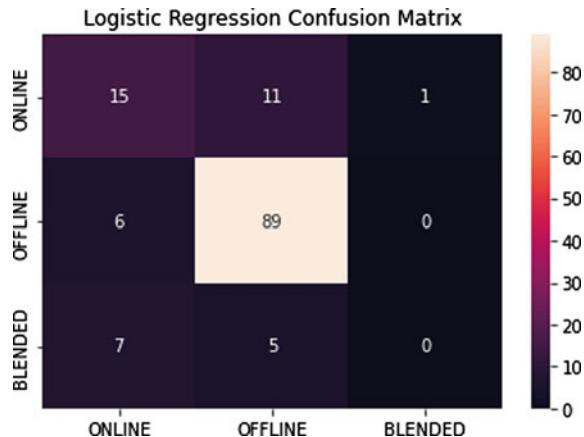


Fig. 10 Logistic regression confusion matrix



5 Conclusion

COVID-19 is a highly infectious disease which has caused shattering effects across a wide spectrum. As the result of its rapacious spree, it has brought forth several economical, physical, psychological, and social impacts. Educational field was also one among the affected. In order to examine the effectuality of online learning, then performed a survey posing clear-cut questions to reason out which among online, offline, and blended mode was favored by majority of the students, framed the questions targeting presentation, quality, critical thinking, effectiveness, Internet connectivity, interaction, and online assessment as our key points for analysis. As the research has demonstrated, offline mode was more popular and preferred by the students. It allows students to have live interaction with their peers and professor, aids in building their confidence, socialization skills, and discipline and allows students to comfortably understand the concepts. All-in all it ensures the overall development of an individual. To put it all together, offline learning is the best loved platform for various reasons as analyzed in our paper. Online learning also has its own set of advantages like time management, flexibility, and self-motivation and one can learn from the comfort of their homes. If such a calamity unfolds in the near future, can always resort to online learning by adapting an approach that reaps optimal results as there is always scope of improvement from the teaching as well as learning ends.

Acknowledgements The authors of this paper are grateful to our Principal Dr. RanaPratap Reddy and Management of Global Academy of Technology who gave us an opportunity to carry out this research work. The authors also thank fellow friends for giving their honest review through the google forms which helped us in collecting the data for our analysis.

References

1. <https://www.cbsedigitaleducation.com/essay-on-online-education/>
2. <https://timesofindia.indiatimes.com/city/bengaluru/over-20k-students-skip-sslc-exams-on-day-1-first-paper-easy-say-many/articleshow/90505398.cms>
3. <https://www.igi-global.com/dictionary/characterization-of-online-learners-or-students-in-namibia/97521>
4. Mondol MS, Mohiuddin MG (2020) Confronting Covid-19 with a paradigm shift in teaching and learning: A study on online classes. *Int J Soc Political Econ Res* 7(2):231–247
5. Subekti AS (2021) Covid-19-triggered online learning implementation: pre-service English teachers' beliefs. *Metathesis: J Eng Lang Lit Teach* 4(3):232–248
6. Quijano Abad DE, Gil Herrera R (2020) A forced migration due to covid19 from a face-to-face university education service to a total online learning management system support: UDELAS-Panamá study case. In: 13th conference of education, research and innovation
7. Baum SE (2020) Distance learning during coronavirus: how it works, benefits and challenges. *Teen Vogue*, 19th March
8. Vijayan R (2021) Teaching and learning during the COVID-19 pandemic: a topic modelling study. *Educ Sci* 11(7):347
9. Sirkema SJ (2014) Analysing e-learning and modern learning environments. *Int J Inf Educ Technol* 4(2):176

10. Szopiński T, Bachnik K (2022) Student evaluation of online learning during the COVID-19 pandemic. *Technol Forecast Soc Chang* 174:121203
11. Yamada Y, Furukawa K, Hazeyama A (2020) Conducting a fully online education of a software engineering course with a web application development component due to the COVID-19 pandemic and its evaluation. In: SEED/NLPaSE@ APSEC, pp 20–28
12. Lepičnik-Vodopivec J, Štemberger T, Retar I (2020) New challenges in education and schooling: an example of designing innovative motor learning environments. *Econ Res-Ekonomska istraživanja* 33(1):1214–1221
13. Xing X, Saghaian S (2022) Learning outcomes of a hybrid online virtual classroom and in-person traditional classroom during the COVID-19 pandemic. *Sustainability* 14(9):5263
14. Chowkase AA, Datar K, Deshpande A, Khasnis S, Keskar A, Godbole S (2022) Online learning, classroom quality, and student motivation: perspectives from students, teachers, parents, and program staff. *Gift Educ Int* 38(1):74–94
15. Wu J, Jen E, Gentry M (2018) Validating a classroom perception instrument for gifted students in a university-based residential program. *J Adv Acad* 29(3):195–215
16. Gururaj V, Shriya VR, Ashwini K (2019) Stock market prediction using linear regression and support vector machines. *Int J Appl Eng Res* 14(8):1931–1934
17. Sanjana S, Shriya VR, Vaishnavi G, Ashwini K (2021) A review on various methodologies used for vehicle classification, helmet detection and number plate recognition. *Evol Intel* 14(2):979–987
18. Kodipalli A, Devi S (2021) Prediction of PCOS and mental health using fuzzy inference and SVM. *Frontiers Publ Health* 9
19. Corwith S (2021) Programming for talent development outside of school. In: Talent development as a framework for gifted education, pp 63–93. Routledge
20. Sanderson E, Greenberger R (2011) Evaluating online programs through a gifted lens. *Gift Child Today* 34(3):42–55
21. Al-khresheh MH (2022) Revisiting the effectiveness of blackboard learning management system in teaching English in the era of COVID-19. *World* 12(1):1–14
22. Tucker S (2001) Distance education: better, worse, or as good as traditional education? *Online J Dist Learn Admin* 4(4)
23. Behzadi Z, Ghaffari A (2011) Characteristics of online education and traditional education. *Life Sci J* 8(3):54–58

Suspicious Crime Identification and Detection Based on Social Media Crime Analysis Using Machine Learning Algorithms



C. Jayapratha, H. Salome Hema Chitra, and R. Mahalakshmi Priya

Abstract Social networking sites, such as Facebook and Twitter, make it possible for people to communicate with each other. However, some people have exploited the influence of social networking sites for illegal activities, such as intimidating news stories and suspicious posts on local community websites. The main objective of this work is to identify suspicious posting activities on Facebook and predict crime rates among those posts. By using pre-processing steps, the dataset can be cleaned up for identification by eliminating missing values, avoiding duplicates, stemming, removing stop words from posts that may indicate a suspicious post, and encoding the dataset for common format conversion. In order to construct a suspected profile, posts from social media are clustered together. In the next step, relevant features related to crime are selected for prediction using a feature selection strategy. To predict the single crime that relates to the various suspicious activities, classification approaches will be applied.

Keywords Suspicious crime identification · Crime detection · Social media crime analysis · Feature selection · Machine learning

1 Introduction

Using crime databases for processing the information can also be nearly as accurate as using crime data mining. By using advanced approaches, data that was not evidently put in is brought to light. Criminal data exploration can only cover a small number of aspects without data mining. In order to ensure all potential associations

C. Jayapratha (✉)

Department of MCA, Karpaga Vinayaga College of Engg. Tech, Madurantakam, Tamil Nadu, India

e-mail: jayaprathaclement@gmail.com

H. S. H. Chitra

Department of CS, Sri Meenakshi Govt. Arts College for Women's, Madurai, Tamil Nadu, India

R. M. Priya

Department of CS, Mangayarkarasi College of Arts and Science, Madurai, Tamil Nadu, India

within the crime data, data mining is used. Artificial intelligence, mathematical and statistical methods, and visualization are used to uncover crime relations that would otherwise go unnoticed. Additionally, all crime dealings can be used to predict future events. The use of automatic crime prediction is one of the most difficult methods of implementing scarce resources effectively for averting crime. The conventional crime prediction prototypes that use Facebook data are limited to labeling the real-time reflection of criminal incidents. Crime models can achieve greater accuracy and predictive power by taking advantage of polarities of sentiment and possible factors.

The purpose of this model (crime prediction model) is to identify offenders in any investigating department. Along with the details of crimes, the technology also keeps track of criminal information in a database. The creation of automated criminal detection and prediction systems that can match and ultimately outperform human performance is the main objective of computer vision researchers. The functions and relationships between the nodes in a conceptual network are described through social network analysis. Using this crime prediction technique, investigators may create a network that explains the personalities of criminals, the movement of material and immaterial commodities and information, and connections between these entities. Additional investigation can highlight vulnerable groups, subgroups, and positions within the network. Investigators may be able to see criminal networks using this method, but their real leaders may still be difficult to find if they keep a low profile.

1.1 Objective and Motivation

- To detect suspicious activity postings, pre-processing and clustering algorithms were used to evaluate the linguistic content in the Facebook dataset.
- Obtaining important crime-related features by examining the pre-processed dataset aids in expanding the crime search using a feature selection technique.

1.2 Outline of the Work

The paper includes five main sections to describe the work flow. The introduction of this research is described in Sect. 1. Section 2 labeled the background study of the work contains various research ideas. The proposed model and its exploration are described in Sect. 3. Section 4 includes the performance evaluation, and finally, the conclusion of the work is concluded in Sect. 5.

2 Review of Literature

Bruin et al. [1] provided a framework for crime trends based on a novel distance metric for associating and classifying all persons based on their characteristics. The authors suggested a method for visualizing criminal activities that makes use of clustering and criminal class identification. To detect criminal trends and speed up the resolution of crimes, Nath [2] use a clustering algorithm in their data mining approach. The authors improved K -means clustering to help in the process of identifying criminal trends. Wang et al. [3] devised the Series Finder, a machine learning agent tasked with detecting patterns in crime perpetrated by the same perpetrator or groups of offenders. Clustering has also been used to examine criminal activity trends and regional criminal history. Phua [4] offer a multilayered recognition system with two extra layers: communal and spike detection. Communal detection uses actual social links to reduce suspicion and is resistant to fake social interactions. Spike detection detects spikes in duplicate copies to raise the suspicion score and is attribute probe robust. Bolla [5] provided a social media data-driven crime detection study. The mapping is accomplished by mining social media data from Twitter with a filtering mechanism that mines tweets and saves them in a database after pre-processing. Stanford's Recursive Deep model, an excellent sentiment analysis tool, is used on a set of crime data to determine the crime intensity of a certain place. Selvakuberan and Indradevi [6] combined feature selection methods with classification systems. Using feature selection methodologies, one may improve the precision, applicability, and understandability of the classification process. Yu et al. [7] concentrate on how to build a data mining algorithm-centered application system for group users. They offer a contextual analysis including the development of a fraudulent duty presentation acknowledgment framework using a decision tree classification algorithm. Huang et al. [8] focused on a novel approach for calculating criminal activity based on mining location-based social network interactions. They can obtain information by leveraging geographical linkages and data gathered from people by exploiting these relationships.

3 Proposed Scheme

This work aims to track the crimes that occur in a certain period over the certain timeframe. Facebook has become the most popular social networking site among online outlets. As a result, this research solely considers the second option, implying that the social media channels of the appropriate portal will be used as the data source. Because social media channels update their postings every minute, it will be extremely easy to receive continually updated posts, and crime data can be separated and tracked statistically in an organized way that is very easy to understandable. In addition, based on an online portal data source, this research would want to be expanded to determine the most crime-affected post (Fig. 1).

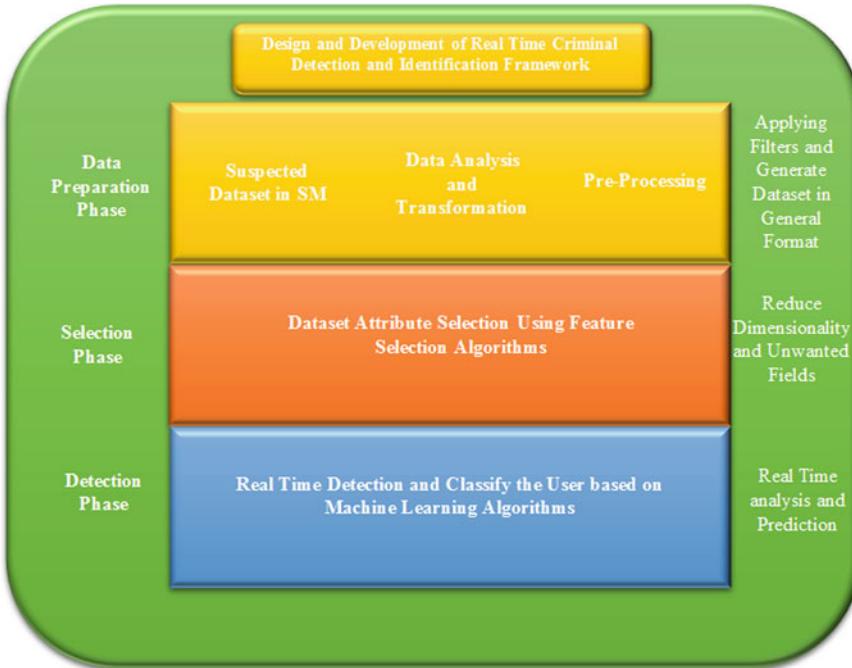


Fig. 1 Overall proposed architecture

3.1 Data Preparation Phase

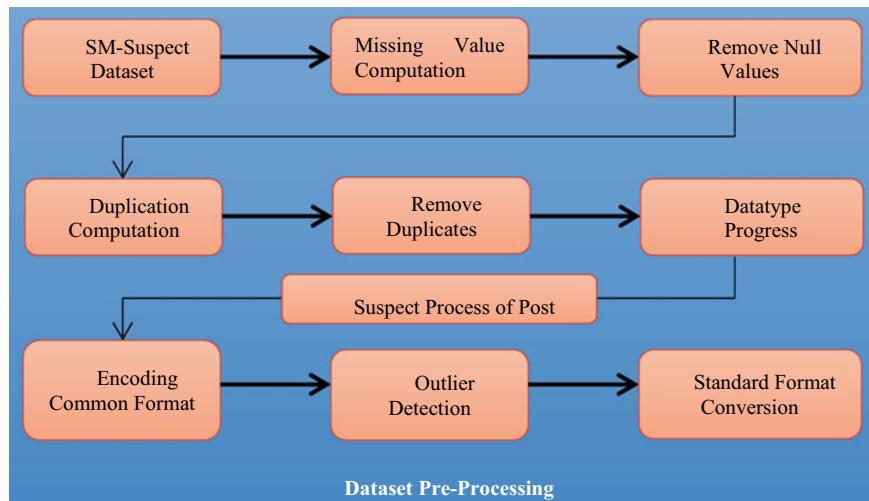
The plan for the criminal dataset collecting and pre-processing stages is given in this section of the work. Data from social media (Facebook) is combined with criminal data and pre-processed using stemming and stop word removal. At that time, the suspicious terms set by the cyber-crime department are investigated for the offenders' suspected list. With the pre-processed data, the suspected data is produced as the crime profile. The pre-processing step includes the methods for pre-processing, which is used to remove noisy data from the suspect dataset.

Table 1 includes the Facebook dataset, which will be used in subsequent operations. In the table, the dataset format is presented along with the fields within the dataset. The Facebook dataset analysis technique was used with the acquired dataset, which included posts, likes, shares, and posted media, among other things. Figure 2 depicts the pre-processing of the suggested models in conjunction with the following tasks.

This pre-processing stage (Algorithm 1) includes various stages such as missing value computation, feature encoding, dimensionality reduction, and so on. Feature encoding is the process that most closely corresponds to the previously discussed data categories. Because ML models cannot handle text input directly, feature encoding

Table 1 Dataset format

ID	Field(s)	Description
1	User ID	Unique ID of the user details (Using instead of Username) related to the suspect criminal name
2	Post ID	Unique ID of the post details that posted by the user
3	Published time	Post publishing time
4	Message	Posted content
5	Media type	Detail explain of the message media
6	Link name	Posted content-type such as video, text or link
7	Link	Post link URL
8	Target	Target type of the post
9	Num_React	Total reaction for the post
10	Num_Like	Total likes of the post
11	Num_Share	Total share of the post
12	Num_Love	Total love reaction of the post
13	Num_Wow	Total WOW reaction of the post
14	Num_HAHA	Total HAHA reaction of the post
15	Num_Sad	Total sad reaction of the post
16	Num_Angry	Total angry reaction of the post

**Fig. 2** Crime dataset pre-processing steps

converts categorical characteristics to numeric values. The performance of most ML algorithms changes depending on how categorical data is represented.

Algorithm 1: Pre-processing

```

Algorithm PreProcessing()
{
Input: DS- Data Set
Output: CM, DM and CR, DR- Missing Values and Redundant Information.
Read the suspected crime dataset DS, Import the DS to the pre-
process phase
For all data rows in the DS
Read_Rows (i) of DS
    CM= Count_NA(DS) & DS_FillNA (DS (i), 0) // impute 0 value to
                                                NA values
    DM= Drop_NA (DS (i)) from DS
    If (is_exists (DS (i))) then // check for duplicate rows in DS
        CR= Count_Redundant (DS) & DR= Remove_Rows (DS (i))
        Else (is_exists_Col (DS (i))) & DR= Remove_Cols (DS (i))
    End if
End for.
Return CM, DM, CR, DR
}.

```

The suspected word identification method from the suspected word database is depicted in Fig. 3. The conventional methodology for recognizing the suspicious post is to construct a list of keywords. The two major issues that arise when suspecting the post are: some of the static keywords are very sensible to receive which is present in the post that may be declared as suspicious; and secondly, the appropriately intellectual specific will not recognize that such assessment has been conducted and will use substitute words instead of known keywords. Post-extraction, stop words removal, stemming procedure, suspicious words matching, and ranking of the suspected post are all processes in suspicious identification.

It involves reducing curved or subordinate terms to their stems, such as the essential or root structure. Sometimes, it is referred to as conflation. The root words are stored in a tree structure, such as a parent-child structure limited; the child is any suspicious term that is planned into a root word. This suspicious wordlist is handled using techniques that are known to be string multi-sets. The suspicious words and



Fig. 3 Suspected words identification process

phrases that were developed as a vector for the identification of suspicious terms from the post are included in the wordlist. The number of times each word appears in the post is calculated using this vector. The procedure for doing common format encoding using the ordinal format is provided. This encoded information comprises the dataset's common format for further grouping of the suspected crime information.

3.2 Selection Phase

The required features are extracted from the crime database at this step. Initially, features were gathered, and some of them were physically removed since they were considered irrelevant to the selection. Finally, only conditional and one class properties have been assessed. The acquired data was structured in tables in a format suitable for the data mining system employed. The data is cleaned by removing the inconsistent values and using the same standard value for all of the data. Cleaning also entails applying the majority data strategy to fill in the missing items. Because some immaterial qualities of the obtained characteristics may reduce the presentation of the classification model, a feature selection strategy is used to select the most acceptable feature subset. If a difference in local search is replaced in a deterministic manner, the *Adaptable Assessment of Local Optimum based Attribute Selection* (AAALOAS) approach is gained. Accept that a fundamental solution 'x' is offered in the interpretations of its calculations. Most local search algorithms at their drop stage use a limited number of regions. The final answer should be a local minimum including all ' K_{\max} ' neighborhood; hence, the chances of arriving at a global one are higher when using this method than when using a single local structure. Algorithm 2 depicts the algorithmic representation of the AAALOAS method.

Algorithm 2: AAALOAS

```

Algorithm AAALOAS ()
{
  Input:  $D_{ST}$  - Data Set
  Output:  $V_{SF}$  - Optimized Feature Subset
  Read the standardized dataset  $D_{ST}$  from the pre-processing phase
  For each feature in the set  $D_{ST}$ 
    Initialize the set of local optimum variables for dataset  $D_{ST}$  and the solution  $S$ 
    Select the random position among the features and define the higher value of local optima
    Apply local optimum assessment to the selected solution
    If (obj_value is better) then
      Replace the value by local optimum variable
      Reconstruct the current solution by randomly selected local optimum
       $V_{SF} = \text{Optimal\_Features from } D_{ST}$ 
      Repeat the process
    End if
  End for
}
```

```

Return VSF
}

```

3.3 Detection Phase

After the pre-processing and feature selection steps, the number of attributes is drastically reduced and is now more exact for creating data mining models. Crime status can be forecasted using several data mining methods. A classification work is used for prediction in this work. Classification is a well-known data mining supervised learning approach that is used to extract useful information from big datasets and may be used to predict unknown classes. As presented model, this work employs the multi-class SVM (MSVM) for crime prediction. The criminal characteristics obtained during the attribute selection step are used as the training set for the classifier for crime prediction. The MSVSM use the K -fold validation approach to build a function with the decision construction using the provided ' N ' training samples from the vector length $\text{Len}(n)$ and $y_i \in \{1, \dots, M\}$ signifies the sample class. The aspects of crime that create ' M ' classifiers for each class. The hyperplane's leftover class for the classifier generates the training set. A new testing set is assigned to the class with the greatest reserve from the boundary in a positive direction. A collection of binary classifiers is trained in this method to be ready to separate each class from all others [9].

Algorithm 3: MSVM classification

```

Algorithm MSVM ()
{
Input: T- Training set
Output: Cout - Multiclass Classifier
Initialize the target class 'C' & Obtain initial feature set
FSet and T
Set Multi-SVM type and parameter
Train 1..., N classifier
For j = 1 to N do
    For each test samples C1 to Cj classes
        Train jth binary SVM & Classify the training samples
        If (j > 1)
            Compute fuzzy scores for all training samples
            Thresholds by splitting the curve of sorted relevance scores
            Output 1 ...N SVM classifier
        End for
    Classify the model
    Compute weight wi,j in class
    TestClassifier(Ti ∈ T)
End For
If (Class (Ti) in Class (Ci)
    COut = Predict (Ti)
End If

```

```

    Return COut
}

```

Then, each data object is categorized according to its most significant decision value. With this system, ' N ' SVMs are trained (the number of classes), and there are ' N ' decision functions. Due to moderately uneven training sets, it suffers from mistakes despite its speed. This method applies successive classifiers to each pair of classes, preserving the first class computed for each object. Then, a max-win operator is used to determine which category the objects are eventually allocated to. The use of this approach necessitates the use of $N*(N-1)/2$ machines. Despite being more computationally expensive than the 'one against all' approach, it has been proved that it may be more suited for multi-class classification challenges; hence, it was chosen for SVM-based classification.

4 Performance Evaluation

This section describes and illustrates the performance evaluation for the pre-processing, feature selection, and classification phases, as well as the technique used in this study. All work is done on a Rack Server with an Intel(R) Xenon Processor and 64 GB RAM running Ubuntu. Python 3 and its related libraries are used to implement the pre-processing, feature selection, and prediction algorithms.

Table 2 shows the details of the dataset and missing data for the dataset. The table illustrates the size of the dataset after missing and duplicate values are removed, and encoding is specified. And it shows the average evaluation of the missing values and duplicated values deleted rows based on the dataset.

(a) Prediction Rate (PR)

Prediction rate is based on number of positive identified from the true positive data. It is used to measure the accuracy of the all classification model. Error rate is derived from the prediction rate. Error rate is opposite to the prediction rate. It shows the how much errors occurs in the classification model [10].

$$PR = \frac{\text{No. of True Positive Identified}}{\text{Total no. of True Positive}} * 100 \quad (1)$$

Table. 2 Duplication computed dataset

Detail of dataset	Size (MB)	Missing data	Values
Total size	21.67	Missing cells count and (%)	2789 and 0.2%
After missing and duplication removal	18.32	Duplicate rows	0
After encoding	11.45		

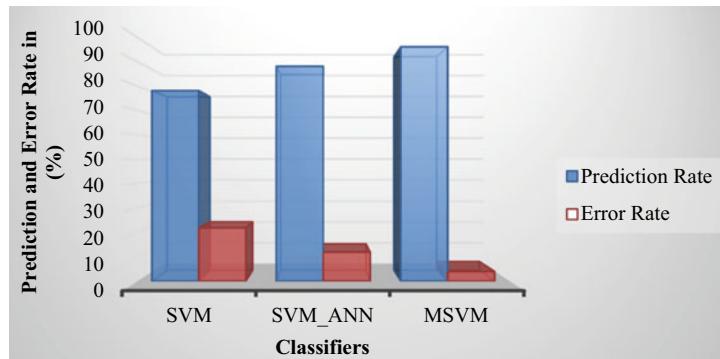


Fig. 4 Prediction rate for classification methodologies

$$ER = 1 - \text{PredictionRate} * 100 \quad (2)$$

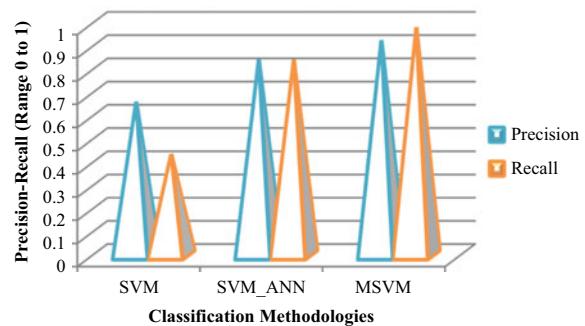
Figure 4 depicts the prediction rate evaluation for the classification algorithms. The graph shows that the suggested approach MSVM has a higher prediction rate than others.

Precision is the fraction of true positives among all persons predicted by the model to be criminally active. This denotes the precision with which a good outcome was predicted. The TPR or Recall is the number of true positives person divided by the total number of true positives and false negatives persons. The evaluation of precision–recall for the classification methods is shown in Fig. 5.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

Fig. 5 Precision–recall of classification methodologies



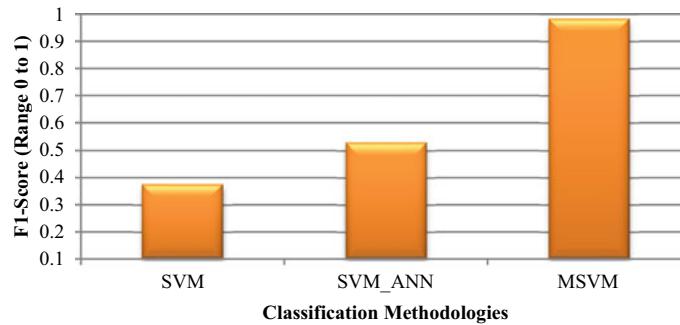


Fig. 6 *F1*-score for classification methodologies

The *F*-score, also known as the *F1*-score, is a model's accuracy on a criminal suspect dataset. It is used to estimate classification systems that categorize instances as 'positive' or 'negative.' Figure 6 shows the *F1*-score for the proposed classifiers.

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

5 Conclusion

This work is intended to discriminate the criminals in any investigation department. The implement charges the images of criminals within the database alongside the criminal information. The key goal of computer vision researchers is to source automated criminal identification and prediction systems which can identify in social media. Social network analysis defines the roles of and collaborations among nodes in a conceptual network. This proposed work achieved the key objectives by examine the group of textual content in social media posts and completed the initial pre-processing phase with Facebook Crime dataset. The work expanded to select the best features in the Processed Dataset using the feature selection approach and achieves the best accuracy related to a single crime detection with the classification methods. Future work will include more case analyses by keep searching and investigation, to have more trained data for machine learning.

References

1. Bruin JS, Cox TK, Kosters WA, Laros J, Kok JN (2006) Data mining approaches to criminal career analysis. In: Proceedings of the sixth international conference on data mining, pp 171–177
2. Nath SV (2010) Crime pattern detection using data mining. *IEEE Trans Knowl Data Eng* 18(009):41–44
3. Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from twitter posts. In: Social computing, international conference on behavioral-cultural modeling and prediction, pp 231–238
4. Phua C (2012) Resilient identity crime detection. *Trans Knowl Data Eng (IEEE)* 24(3):533–546
5. Bolla RA (2014) Crime pattern detection using online social media. Masters Theses
6. Selvakuberan K, Indradevi (2008) Combined feature selection and classification—A novel approach for categorization of web pages. *J Inf Comput Sci* 3(1):10–18
7. Yu F, Qin Z, Jia X (2003) data mining issues in fraudulent tax declaration detection. In: Proceedings of the second international conference on machine learning and cybernetics, pp 2202–2206
8. Huang YY, Li CT, Jeng SK (2015) Mining location-based social networks for criminal activity prediction. In: Proceedings of 24th international conference on wireless and optical communication (IEEE), pp 185–190
9. Gupta A, Mohammad A, Syed A, Halgamuge MN (2016) A comparative study of classification algorithms using data mining: crime and accidents in Denver city the USA. *Int J Adv Comput Sci Appl* 7(7):374–381
10. Hu R, Zhu X, He W, Yan Y, Song J, Zhang S (2017) Graph self-representation method for unsupervised feature selection. *Neuro-computing* 220(1):130–137



C. Jayapratha MCA., M.E., M.Phil., Ph.D., Professor, Department of MCA, Karpaga Vinayaga College of Engineering and Technology, Madurantakam Tamil Nadu., She has 16 year experience in teaching field. Completed her Ph. d in Bharthiyar University at 2021, M.E in G. K. M College Of Engineering, Anna University at 2011. Published 4 Journals, 3 Conferences and Conduct 2 seminars. Interested on Data structure and Algorithms, Data Mining and Machine Learning.



H. Salome Hema Chitra M.Sc., M.Phil., Ph.D, Lecturer in Computer Science, Sri Meenakshi Government Arts College for Women, Madurai. She completed her UG Degree (B.Sc (Physics)) LDC, Madurai in 1999, and PG Degree (M.Sc (Computer Science)) in MKU College, Madurai in 2001. She has been awarded her Ph.D Degree in Computer Science, 2018 and 15+ years of experience in teaching field. She has published 13 + research papers in journals and conferences. She has published her patent work at 2022. She has interests in domains like, Image Processing and Data science.



R. Mahalakshmi Priya M.Sc.,M.Phil, B.Ed., Assistant Professor, Mangayarkarasi College of arts and science, Madurai. She has two year's teaching experience years in LMIT and 4+ year's experience in Software Development. Completed her M.Sc. (CS&IT) at Mannar College in 2009–11 and finished M.Phil in Mother Teresa University at 2018. Published 2 Journals, 4 Conferences and Filed 1 Patent. Interested on Image Processing, Data Mining and Machine Learning.

Deep Learning-Based Similar Languages’ POS Tagging: Experiments on Bhojpuri, Maithili, and Magahi



Rajesh Kumar Mundotiya, Praveen Gatla, Nikita Kanwar, and Anil Kumar Singh

Abstract Monolingual corpora and similar language resources are vastly available for a few languages. These resources stimulate the exploration and building of potential NLP tools for new languages or dialects. This paper deals with the part-of-speech (POS) tagging for the Indo-Aryan languages, i.e., Magahi, Maithili, and Bhojpuri, a dialect of Hindi. The POS model is trained by BiLSTM-CRF and explores the effectiveness of Word2Vec, GloVe as word and FastText, and BPE as subword-level embeddings, trained on the raw corpus of these languages. All these languages are dialects of Hindi; hence, multilingual embedding at the BPE level has been evaluated. Better results are obtained than with monolingual BPE embedding. However, the best results have been obtained from word embeddings, i.e., GloVe on Maithili and Magahi, with 81.23% and 82.24%, respectively.

Keywords POS tagging · Low-resource language · Word embedding

1 Introduction

Parts-of-speech (POS) tagging is a process of marking a word in a corpus corresponding to a known tagset and based on its morphological features. It is a primary step of text processing in a typical natural language processing (NLP) pipeline. A word has a different part of a speech tag based on different contexts in different

R. K. Mundotiya (✉)
University of Petroleum and Energy Studies, Dehradun, India
e-mail: rajeshkm.mundotiya@gmail.com

P. Gatla
Department of Linguistics, Faculty of Arts, BHU, Varanasi, India

N. Kanwar
Computer Science and Engineering Department, PRATAP Institute of Technology and Science, Sikar, India

A. K. Singh
Department of Computer Science and Engineering, IIT (BHU), Varanasi, India
e-mail: aksingh.cse@iitbhu.ac.in

sentences, making it challenging to have a generic mapping for POS tags. POS tags are building blocks for many other tasks involving preprocessing steps to syntactic parsing, lemmatizers, and corpus linguistics.

Bhojpuri, Magahi, and Maithili belong to the Indo-Aryan language family. They share almost the same typological features. Among these three languages, Maithili is a scheduled language. These languages are spoken majorly in Bihar, Uttar Pradesh, and Jharkhand. In addition to India, Maithili, Magahi, and Bhojpuri are also spoken in Fiji, Nepal, Surinam, and Mauritius, hence it is reasonable to classify them as living languages with considerable user populations. These languages are characterized as **low-resource languages** because they lack an abundance of annotated data. This limitation degrades the performance of the NLP tools, including POS tagging. Multilingual POS tagging is one of the premier solutions to enhance performance within such limitations.

After gaining deep learning in NLP, multilingual POS tagging has been performed by either tag projection or multilingual word embedding. The tag projection method has been popular with a conventional learning algorithm, which requires massive parallel corpora. Nowadays, a plausible method in multilingual word embedding uses minimal parallel corpora for supervised settings or monolingual corpora for unsupervised settings. Ideally, the attainment of parallel corpora requirements in both methods is not feasible for low-resource languages. Magahi, Maithili, and Bhojpuri are all low-resource and similar languages. Hence, parallel corpora among these languages were not available until now.

In this work, we exploited a raw corpus and syntactically annotated corpora for performing monolingual and multilingual experiments on POS tagging and tried to figure out two problems, which are

- Is the available raw corpus helpful for improving the performance of POS tagging on Magahi, Maithili, and Bhojpuri?
- For related languages, embedding at the subword level affects the efficiency of neural machine translation (NMT) systems [4]. Does it work for the neural POS tagger for these three languages in a multilingual approach as similar languages?

To accommodate these questions, we have performed experiments by using a BiLSTM-CRF model with multilingual and monolingual word embedding generated from a monolingual corpus of these languages. The following are the paper's contributions:

1. BiLSTM-CRF, a neural-based approach for these languages was explored for designing a more accurate POS tagging system.
2. These languages are similar and also follow the same writing script, i.e., the Devanagari script. Subword-level BPE embedding provides a better embedding space instead of the monolingual BPE that has been explored empirically for POS tagging.

2 Sequential Labeling Model

Sequence labeling is a core NLP task that ascribes a label/tag to each word of the input sequence using a pre-defined label set. The popular token labeling tasks are word chunking, named entity recognition (NER), and POS tagging in NLP. POS tagging is a major area that is explored in this paper.

2.1 *BiLSTM-CRF Model*

Word embedding represents a single word well, but these words' ordering in a sentence is not well-treated. The occurrences of unique words are not only transformed into high-dimensional vectors, but the ordering is kept by introducing neural network layers. The character and word embedding identify a pair of homonyms in the same vector representation, though these are not the same words. It reduces the accuracy of the model. The long-term dependencies are captured by LSTM, which is a variant of a recurrent neural networks [7, 10]. A LSTM cell consists of a collection of memory blocks that are interconnected recurrently. The block consists of one or more recurrently connected memory cells and three multiplicative units—the input, output, and forget gates—that enable continuous equivalents of write, read, and reset operations. A contextual embedding mechanism helps to understand the context of a word. LSTM sequences belong to the contextual embedding layer. Bidirectional-LSTM (BiLSTM) composed of both backward as well as forward-LSTM sequences [6]. The forward-LSTM layer captures past dependencies, and the backward LSTM layer captures future dependencies. The merged output embeddings encode information from both past (backward) and future (forward) states from the backward-and forward-LSTM simultaneously. Back-propagation through time [1] is a training algorithm that is used to update weights in bidirectional LSTM networks. The strong interdependence between POS tags captured by CRF [12] in the form of linear-chain CRF can be employed with BiLSTM. Predicting current tags using neighbor tag information considering sentence-level information focused by the CRF model. BiLSTM, in tandem with linear-chain CRF, can help the model generate a valid dependent series of POS tags rather than an independent label.

2.2 *Word Embedding*

Word embedding is a deep learning-based feature extraction method of text that permits a similar representation of words with similar meanings in a pre-defined shared space. Word embeddings are mappings from words into n -dimensional vectors. There are lots of ways to build word embedding. (a) One of the most popular is Google's Word2Vec [16] which is a statistical method with two different learning

models, used to learn the word embedding. The continuous bag-of-words (CBOW) model learns the embedding by predicting the current word based on its context. The continuous SkipGram model learns by predicting the surrounding words of a given word. (b) Stanford University’s Global Vectors (GloVe) [20] is an unsupervised learning algorithm for obtaining distributed word representation by mapping words into a meaningful space where the distance between words is related to semantic similarity. (c) Facebook’s AI Research FastText is a library for efficient learning of word representations and sentence classification [2]. It is based on the SkipGram model, where each word is represented as a bag of character n -grams. Character n -grams are shared across words. This model does better than word embedding models for out-of-vocabulary words—it can generate an OOV word embedding. In FastText, each center word is represented as a set of subwords. Instead of representing a sentence using words, we need to go with its subword level. Specifically, in a language with rich morphology, it would be nice to use this morphological information to better represent a sentence.

The implementation of Word2Vec is fast, efficient, and widely used. However, it can not easily represent a sentence using Word2Vec, and it does not exploit morphology. The GloVe and FastText systems are essentially extensions of the Word2Vec model. Only the vectors for entire words that are obtained in the training corpus are learned by Word2Vec. FastText is able to learn vectors for the n -grams that are contained within each individual word and is useful for languages with lots of internal structure and formation methods. Due to character n -grams, FastText encodes OOV and rare words more effectively.

The SkipGram model is trained with the morphological information by exploiting regularities present in the word representation [26]. The Word2Vec model and GloVe embedding are trained on a large corpus containing billions of tokens, which may not hold for low-resource languages. Training word vectors in such a setting is a challenging problem.

2.3 *Byte-Pair Embedding*

In data compression techniques, particularly Byte-Pair Encoding, the most common pair of bytes is replaced with a new pair of bytes that have never existed in the data before. The BPE algorithm represents each token in the corpus as a group of characters and a specified end-of-word token. Iteratively, it counts character pairs with all tokens of the vocabulary and adds the new character n -gram to the vocabulary after merging each occurrence of the most frequent pairs. Repeatedly, each frequent pair is to be merged until the desired vocabulary size or the defined number of merge operations is reached [24].

2.4 Multilingual Word Embedding

Traditionally, word embeddings are language-specific and separately trained in each language that exists in totally different vector spaces. Multilingual word embeddings (MWEs) share the vector space of words from more than one language. It can leverage the dependencies among participating languages, which generates a multilingual embedding space for low-resource languages. Aligning the embeddings of various languages is a series of geometrical transformations. In multilingual word embeddings, each of the languages in the same vector space and words with similar meanings (regardless of language) are close together. To generate this vector space, the raw corpora of all languages are concatenated before generating the embeddings. Here, we have used the subword-level (BPE) embedding for this experiment. These concatenated segmented subwords are used to train the Word2Vec embeddings. This multilingual training setup is inspired by MultiBPEmb [9].

3 Experimental Details

3.1 Dataset

For compiling our methodology and experimental setups, we have used an in-house (annotated) corpora for Magahi, Maithili, and Bhojpuri. The annotation of these languages follows the Bureau of Indian Standard (BIS) tagset. This paper [18] described similarities (semantic and syntactic), number of tokens, number of annotated sentences, and frequency of words for each tag, for each language. A brief description of the annotated corpus is summarized in Table 1. These datasets have been split into 80-10-10 ratios to train, test and validate our model. During the training, test data was not used.

3.2 Parameter Settings

For performing the BiLSTM-CRF model training, we have used four kinds of embedding which belong to word and subword levels, i.e., Word2Vec, GloVe (word level),

Table 1 Statistics of POS tagging datasets

Language	# Token	# Types	# Sentence
Bhojpuri	245,482	26,202	16,067
Maithili	208,640	21,410	12,310
Magahi	171,509	14,077	14,669

FastText, and Byte-Pair Encoding (BPE) (subword). The raw corpus of Magahi, Maithili and Bhojpuri [17, 18] has been exploited for training these embeddings by using standard libraries of each embedding technique that are GloVe-Python¹ for GloVe, Gensim² for Word2Vec, fastText³ for FastText and SentencePiece⁴ for BPE. However, all these three languages follow the same writing script of Hindi as well as similar languages; hence, we have followed the Hindi tokenizer from the Indic NLP Library⁵ which prevents from tokenization relevant errors (such as an unknown word) during training.

We have assumed that 100-dimensions can significantly capture linguistically-based relevant information. As these languages are low-resource languages, we consider each token (including rare types) during embedding training. We kept the adjacency of words up to 5 and trained over 100 epochs using the CBOW training algorithm for training the word-level embeddings. The FastText embedding used the SkipGram training algorithm at a learning rate of 0.01 over 100 epochs. For generating the embeddings by BPE, we have used character coverage of 1 (due to less raw corpus) and 1000 for vocabulary size as a parameter. The rest of the parameters have default values for all of these embedding training.

4 Result and Analysis

From the set of experiments, we tried to improve the POS tagging performance on the annotated dataset of these languages. The BiLSTM-CRF model is trained up to a maximum of 100 epochs with early stoppage criteria and a learning rate reducer to prevent overfitting during training. We have used the available raw corpus of these languages to improve the model performance. We evaluated BPE on different vocabulary sizes, such as 1000, 2000, and 3000, out of which we got the best score of 1000. Hence, the reported results for each language in monolingual and multilingual settings are performed on this. The model performance has been evaluated in terms of accuracy, F-score, precision, and recall. In this direction, FastText achieved the best F-score for Bhojpuri. However, the best score of Bhojpuri is very low compared to the baseline score, as mentioned in Table 2.

Similarly, the best F-score on Magahi and Maithili obtained from GloVe embedding is better than the baseline score, reported in the same table.

A closer analysis of these results has shown that word-level embedding provides better results compared to the subword level for all these languages due to dealing with word representations directly, whereas subword level decreases the performance of Maithili and Magahi, since we have used pretrained embeddings of FastText and

¹ <https://github.com/maciejkula/glove-python>.

² <https://radimrehurek.com/gensim/>.

³ <https://github.com/facebookresearch/fastText>.

⁴ <https://github.com/google/sentencepiece>.

⁵ https://github.com/anoopkunchukuttan/indic_nlp_library.

Table 2 POS tagging results (in %) obtained on different word embedding

	Embedding	Precision	Recall	F-score	Accuracy
Bhojpuri	GloVe	82.13	82.26	82.19	82.26
	Word2Vec	81.38	81.32	81.17	81.32
	FastText	82.50	82.03	81.79	82.50
	BPE	79.14	79.27	80.75	80.89
	Baseline [18]	89.00	89.00	88.00	89.00
Maithili	GloVe	81.84	81.74	81.23	81.74
	Word2Vec	81.01	81.16	80.70	81.16
	FastText	73.61	74.58	73.11	74.58
	BPE	76.43	76.38	76.22	76.74
Magahi	Baseline [18]	77.00	77.00	76.00	77.00
	GloVe	83.19	82.57	82.24	82.57
	Word2Vec	80.29	79.36	78.71	79.36
	FastText	77.96	77.51	76.21	77.51
	BPE	75.93	75.33	74.69	75.62
	Baseline [18]	80.00	78.00	78.00	78.00

Table 3 Multilingual BPE embedding result

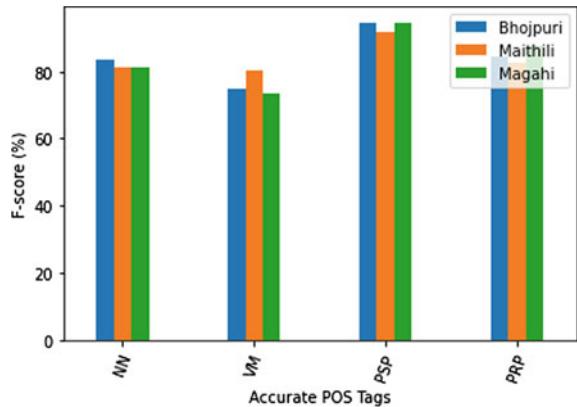
Language	F-score	Accuracy
Bhojpuri	80.75	80.93
Maithili	76.38	76.50
Magahi	75.21	75.29

BPE. Using Maithili pretrained BPE [9] embedding of the 100 dimensions provides 80.80, 80.97, 80.23, and 80.97 as precision, recall, F-score, and accuracy, respectively. Similarly, pretrained FastText [5] embedding of the 300 dimensions for Maithili produces 74.00, 73.86, 72.71, 73.86 as precision, recall, F-score, and accuracy, respectively.

The BiLSTM-CRF model is language agnostic; hence, all the languages' raw corpora are concatenated before BPE. These concatenated segmented subwords are used to train the Word2Vec embeddings. This multilingual training setup is inspired by MultiBPEmb [9]. The obtained F-score and accuracy for each language have been mentioned in Table 3. The result indicates that multilingual BPE improves the POS system's performance as compared to monolingual BPE embedding.

We found that our model performs well for the following tags for these three languages: we noticed for Bhojpuri, our model is performing well for pronoun (PRP), noun (NN), postposition (PSP), and verb (VM). Also similar to Maithili, model performance is quite impressive for NN, VM, and PRP. For Magahi, our model's performance is quite similar to that of Bhojpuri and Maithili. We also observed that

Fig. 1 Most accurate POS tags for Magahi, Maithili, and Bhojpuri



the model performs pretty well for closed-class words like postpositions and particles. The training data contains the high frequency of these tokens in comparison with close class tokens.

Additionally, we found almost similar accuracy for all languages. Because Bhojpuri, Magahi, and Maithili are close-cognate languages and share common morphological and typological features. Maithili is known as one of the most polite languages, and it is reflected in our results as well. The model predicts the PRP as the most frequent token in the entire Maithili data. The summarized result is mentioned in Fig. 1.

5 Related Work

POS tagging is an essential NLP task that creates a base for several tasks such as syntactic parsing and furthermore. Many experiments have been done on the POS tagger which are [13, 14] on PTB-WSJ, but there remain issues. The accuracy measure token-by-token is simple because a high degree of ambiguous words such as punctuation is assigned by proper POS tags readily through the tagger, as stated by Manning [15], whereas accuracy at sentence-level maps to a more realistic criteria for it.

Ambiguous tokens along with low-resource languages present a potential problem for POS taggers, as their availability and accuracy are substantially reduced [22]. Moreover, much more work has already improved the POS tagger's accuracy, like Yasunaga et al. [29] worked on the adversarial training POS (AT) tagger. A lot of work on the parallel sentences POS tagger data has been done [21, 27, 28]. Multilingual transfer is addressed by Kim et al. [11] through hidden word classification. Hana et al. [8] used POS tagging transfer without parallel data. Using topologically similar languages and the morphological properties of their data for training, models were able to learn probabilities of transitions that reflect the morphological properties.

Other methods used the robust projection method on the aligned data to guess the POS information and noun phrase bracketing. Gaddy et al. [3] used the coarse mapping word embedding to predict the POS tagging information for ten different languages.

POS annotation tool for a language with limited resources, Bhojpuri [25], achieved 88.6% accuracy with 33 tagsets trained using a support vector machine (SVM) classification model. Employing SVM and CRF for Indo-Aryan languages such as Bhojpuri, [19] experiment with 90K training tokens where CRF (82–89%, based on the type of test data) performs qualitatively better than the SVM model. The Maithili POS tagger developed [23] using CRF with a tagset comprising 27 tags, and 2460 sentences reports the accuracy of CRF with orthography features is 82.67%, and neural word embedding is 85.88%. On the same dataset, the CRF technique has been evaluated, which produced 89%, 77%, and 78% as baseline accuracy for Magahi, Maithili, and Bhojpuri [18], respectively.

6 Conclusion

Part-of-speech tagging is the preliminary module for various NLP and natural language understanding-based tasks like named entity recognition, automatic question answering bot, and machine translation system. Here, the in-house standard corpus of Magahi, Maithili, and Bhojpuri is used to train word embeddings such as Word2Vec, GloVe, BPE, and FastText for each language. These trained embeddings have been used in the BiLSTM-CRF-based POS model. The BPE-based embedding is further used to generate multilingual embeddings due to the capability of sharing subword information. The obtained F-score results exhibit a better POS model by using GloVe embedding than the baseline model for Maithili and Magahi by +5.23% and +4.24%, respectively, whereas it lacks the Bhojpuri language.

References

1. Boden M (2002) A guide to recurrent neural networks and backpropagation. The Dallas project
2. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 5:135–146
3. Gaddy DM, Zhang Y, Barzilay R, Jaakkola TS (2016) Ten pairs to tag-multilingual POS tagging via coarse mapping between embeddings. Association for Computational Linguistics
4. Goyal V, Kumar S, Sharma DM (2020) Efficient neural machine translation for low-resource languages via exploiting related languages. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics: student research workshop. Association for Computational Linguistics, pp 162–168
5. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. In: Proceedings of the international conference on language resources and evaluation (LREC 2018)

6. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 6645–6649
7. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 18(5–6):602–610
8. Hana J, Feldman A, Brew C (2004) A resource-light approach to Russian morphology: tagging Russian using Czech resources. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 222–229
9. Heinzerling B, Strube M (2018) BPEmb: tokenization-free pre-trained subword embeddings in 275 languages. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA)
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
11. Kim Y-B, Snyder B, Sarikaya R (2015) Part-of-speech taggers for low-resource languages using CCA features. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1292–1302
12. Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, pp 282–289
13. Ling W, Luís T, Marujo L, Astudillo RF, Amir S, Dyer C, Black AW, Trancoso I (2015) Finding function in form: compositional character models for open vocabulary word representation. arXiv preprint [arXiv:1508.02096](https://arxiv.org/abs/1508.02096)
14. Ma X, Hovy E (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint [arXiv:1603.01354](https://arxiv.org/abs/1603.01354)
15. Manning CD (2011) Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, pp 171–189
16. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
17. Mundotiya RK, Kumar S, Kumar A, Chaudhary UC, Chauhan S, Mishra S, Gatla P, Singh AK. Development of a dataset and a deep learning baseline named entity recognizer for three low resource languages: Bhojpuri, Maithili and Magahi. ACM Trans Asian Low Resour Lang Inf Process (Just accepted)
18. Mundotiya RK, Singh MK, Kapur R, Mishra S, Singh AK (2021) Linguistic resources for Bhojpuri, Magahi, and Maithili: statistics about them, their similarity estimates, and baselines for three applications. ACM Trans Asian Low Resour Lang Inf Process 20(6)
19. Ojha AK, Behera P, Singh S, Jha GN (2015) Training and evaluation of POS taggers in Indo-Aryan languages: a case of Hindi, Odia and Bhojpuri. In: The proceedings of 7th language and technology conference: human language technologies as a challenge for computer science and linguistics, pp 524–529
20. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
21. Petrov S, Das D, McDonald R (2011) A universal part-of-speech tagset. arXiv preprint [arXiv:1104.2086](https://arxiv.org/abs/1104.2086)
22. Plank B, Søgaard A, Goldberg Y (2016) Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics, vol 2 (short papers), pp 412–418
23. Priyadarshi A, Saha SK (2020) Towards the first Maithili part of speech tagger: resource creation and system development. *Comput Speech Lang* 62:101054
24. Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics, vol 1 (long papers), pp 1715–1725

25. Singh S, Jha GN (2018) Part-of-speech tagger for Bhojpuri. In: WILDRE4—4th workshop on Indian language data: resources and evaluation, p 36
26. Sorice R, Och FJ (2015) Unsupervised morphology induction using word embeddings. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, pp 1627–1637
27. Täckström O, McDonald R, Nivre J (2013) Target language adaptation of discriminative transfer parsers
28. Wisniewski G, Pécheux N, Gahbiche-Braham S, Yvon F (2014) Cross-lingual part-of-speech tagging through ambiguous learning. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1779–1785
29. Yasunaga M, Kasai J, Radev D (2018) Robust multilingual part-of-speech tagging via adversarial training. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, vol 1 (Long papers), New Orleans, Louisiana, June 2018. Association for Computational Linguistics, pp 976–986

CS-Jaya: Hybridization of Cuckoo and Jaya Algorithm



Megha Varshney, Pravesh Kumar, and Tarun Kumar Sharma

Abstract Jaya algorithm and Cuckoo Search Algorithm are newly developed meta-heuristic algorithms for solving optimization algorithms. In this paper, we have proposed a new algorithm, hybridizing the concept of the Jaya algorithm with the Cuckoo Search Algorithm, named CS-Jaya. CS-Jaya algorithm utilizes the advantage of both algorithms. Five test functions are selected from the literature for testing of CS-Jaya performance. Then, lastly, numerical results reveal the efficiency of the presented algorithm.

Keywords Optimization · Cuckoo search algorithm · Jaya algorithm

1 Introduction

Input or specification setting for a device, a mathematical function, or an experiment that aims to maximize or reduce the output (result) is known as optimization. In optimization problems, the process or function is known as the cost function or the fitness function, and the output is known as the cost or fitness. The input of these issues often has a fundamental answer known as a global optimal solution. In some challenging issue optimization problems, obtaining a global optimal solution takes a lot of time and calculation. So population-based optimization algorithms have the benefit of offering appropriate solutions with less calculation and computation times [1–3]. Numerous strategies in the field of optimization that were inspired by nature have been created to address this. Ant Colony Optimization (ACO), Cuckoo Search (CS), Artificial Bee Colony (ABC), Differential Evolution (DE), Genetic Algorithm (GA),

M. Varshney (✉) · P. Kumar

Rajkiya Engineering College Bijnor (AKTU Lucknow), Lucknow, India

e-mail: varshney.megha31@gmail.com

P. Kumar

e-mail: pkumarrecb@gmail.com

T. K. Sharma

Shobhit Institute of Engineering and Technology (Deemed to be) University, Meerut, India

e-mail: taruniitr1@gmail.com

Jaya algorithm (JA), Particle Swarm Optimization (PSO), and others are examples of well-known techniques.

The cuckoo search technique was first introduced in 2009 by Yang and Deb [4]. This technique, which is used to solve continuous and NP-hard (Non-deterministic polynomial time) problems, was inspired by the natural behavior of cuckoo birds. Numerous academics have evaluated it using different benchmark functions. The ABC Method [5], PSO Algorithm [6], and GA Algorithm [7] were found to produce much worse results than the Cuckoo Search Algorithm. You may find a thorough analysis of the Cuckoo Search Algorithm in the [8–13].

Rao originally presented the Jaya algorithm in 2015 [14]. It is a comprehensive algorithm for resolving numerous real-world issues. This algorithm's simple implementation and condensed design are its key benefits. This algorithm has many variations, and its applications are discussed in [15–24].

Recent research has resulted in the development of the hybrid algorithm CS-Jaya, which combines the benefits of the algorithms Cuckoo Search and Jaya into a single variation. Later in the study, the significance of a planned algorithm is covered.

The study is organized as follows: Sect. 2 describes how the Cuckoo Search and Jaya algorithm work. Section 3 describes the Contemplated CS-Jaya. Section 4 discusses benchmark functions, parameter settings, performance standards, outcomes, and comparisons of CS-Jaya with other variations. Section 5 has the conclusion.

2 Working of Basic Algorithms

This section explains details functioning of Jaya algorithm and Cuckoo Search.

2.1 Cuckoo Search Algorithm

CS is an evolutionary metaheuristic algorithm based on population size. The CS algorithm is straightforward, easy to use, and has a small number of control parameters. It is based on the Levy's flight-based obligatory parasitism of some cuckoo species in flocks.

The Cuckoo Search procedure is described below [25].

1. Each time a cuckoo lays an egg, it places it in a different nest
2. The next generation inherits the best nest that produces an egg
3. There are fixed number of host nests available.

Cuckoo eggs are more likely to be found by host birds with probability P_a (0, 1). In this situation, the host bird has two options: either toss the eggs or to leave the nest and make a brand-new one. Each egg in this case can be thought of as a potential solution.

Each solution is created at random during the first stage of the CS process while creating the i th solution in the $(k+1)$ th generation. The Equation (1) updates the nest's or egg's location (1).

$$x_i(k+1) = x_i(k) + \alpha \otimes L(l) \quad (1)$$

where $x_i(k+1)$ is the nest's i th position of the $(k+1)$ generation in the population, α is a real number, it marks the step size, \otimes represents entry wise multiplications, and $L(\lambda)$ is the random search vector produced by Levy distribution.

Levy flight is a crucial component of the CS [26] for local and global searching:

$$L(l) \sim u = t^{-\lambda}, (1 < \lambda < 3) \quad (2)$$

Here, the cuckoo's subsequent steps, which adhere to a step length distribution, constitute a random walk. Levy walks are used to produce some of the new solutions in order to speed up local searches for the optimal answer.

2.2 Jaya Algorithm

For the purpose of resolving both limited and unconstrained optimization issues, Rao has created a revolutionary metaheuristic algorithm [14]. The Sanskrit word “Jaya” means “Victory.” In this method, the best solution is kept while the poorest solution is ignored in each generation, while updating the location. Every suggestion made here is referred to as a particle.

The position of the particle is updated mathematically by Equation (3)

$$x_i(k+1) = x_i(k) + r_1 \times [x_{\text{best}} - |x_i|] - r_2 \times [x_{\text{worst}} - |x_i|] \quad (3)$$

where r_1 and r_2 are uniform random numbers with a value in range $[0, 1]$. The best and worst candidate values are x_{best} and x_{worst} respectively.

3 Proposed CS-Jaya Algorithm

The CS algorithm is incredibly effective and only uses a few regulating parameters. The Levy flight's great randomness is used to speed up the search process. As a result, the algorithm's ability to search globally is highly strong. However, because of the algorithm's high level of randomness, a blind search process is started, which slows convergence and makes it much easier to find the best answer. Therefore, Jaya is included to the random initialization of the abandoned nests procedure of CS to enhance its performance. As a result, CS-Jaya is better able to manage the location of the abandoned nests.

Working of CS-Jaya

```

Begin_Cuckoo_Search_Jaya_Replacement
    Objective function  $f(\bar{x})$ ,  $\bar{x} = [x_1, x_2, \dots, x_d]$ 
    Randomly initialize the population with  $n$  host
    nests  $\bar{x} = (x_1, x_2, \dots, x_n)$ 
    While (t < Max Generation or (stop criterion))
        Randomly select a cuckoo based on Levy's flight
        and rate its fitness  $f_i$ .
        Randomly select one of the  $n$  nests a  $j^{th}$  nest and
        evaluate its fitness  $f_j$ .
        If ( $f_i > f_j$ )
            Replace  $j^{th}$  nest with New Solution
        End If
        Some of the bad nests are discarded and new nests
        are built in new locations via the Jaya algorithm.
        Organize solutions that hold the best solution (or
        nest of quality solutions) and find the current best.
    End While
    Terminate

```

4 Simulation Results and Discussion

This section contains details of evaluation criterion, parameter settings, test functions, and discussion on numerical results and comparison.

4.1 Evaluation Criterion and Parameter Settings

- **Evaluation:** Performance of algorithms is evaluated in term of *Average Error*, *Average NFE*, and *Acceleration rate*
- **System Settings:** The proposed algorithm is implemented in MATLAB. Besides, the system supports an Intel Core i5—3470 processor running at 3.20 GHz with a RAM chip of 4 GB, Windows 7, and 64-bit OS
- **Population Size:** 100
- **Dimension:** 30 for each test function
- **Probability Pa:** 0.25

Table 1 Test Functions

Function	Name	Property	Search space
F1	Sphere function	Unimodal, separable, scalable	[- 100, 100]
F2	Schwefel's problem2.22	Unimodal, separable, scalable	[- 10, 10]
F3	Rosenbrock's function	Multimodal, non-separable, scalable	[- 30, 30]
F4	Griewank function	Multimodal, non-separable, scalable	[- 600, 600]
F5	Ackley function	Shifted, multimodal, separable, scalable	[- 32, 32]

- $\sigma_u : 0.6969$
- **Maximum NFE:** 150,000
- **Test Run:** 30
- **Acceleration Rate:**

$$AR = \left(1 - \frac{NFE_{CS-Jaya}}{NFE_{\text{other}}} \right)$$

4.2 Test Functions

Following five test functions have chosen from literature [27, 28] for testing of CS-Jaya and comparison with other algorithms (Table 1).

4.3 Results and Comparisons of the CS-Jaya Search Algorithm with Jaya and Cuckoo

This section discusses numerical results and contrasts the proposed CS-Jaya algorithm with the Cuckoo Search and the Jaya algorithm.

In Table 2, numerical results are given in terms of mean error and standard deviation. Here, the comparison is also taken with CS and Jaya algorithm. We can easily see that our proposed CS-Jaya gives better accuracy in comparison with CS and Jaya algorithms.

Table 3 presents numerical results in term of mean NFE and AR. Here, we can see that the NFE of CS-Jaya is very less to reach fixed VTR for each test function. Average NFE counts for Cuckoo search and Jaya algorithms are 76,600 and 352,025, respectively, while average NFE taken by CS-Jaya are only 67,200. We can also see the acceleration speed of CS-Jaya over Cuckoo search and Jaya algorithm for each

Table 2 Results and comparisons in terms of mean error and standard deviation

Function	Max NFE (K)	Mean error (Standard deviation)		
		Cuckoo search	Jaya	CS-Jaya
F1	150	1.72E-21 1.03E-21	2.06E-16 7.87E-16	1.44E-27 6.91E-27
F2	150	4.96E-21 3.10E-21	5.23E-08 5.37E-08	7.98E-23 2.15E-23
F3	150	5.84E-06 2.82E-06	6.05E+00 1.16E-01	4.66E-08 1.31E-07
F4	50	7.11E-06 2.08E-06	5.96E-09 2.07E-09	5.23E-07 2.08E-06
F5	50	7.26E-06 1.94E-06	4.04E-02 2.17E-02	4.21E-07 1.22E-06

Table 3 Results in terms of mean NFE and AR

Function	VTR	Mean NFE			AR	
		Cuckoo search	Jaya	CS-Jaya	CS-Jaya versus Cuckoo	CS-Jaya versus Jaya
F1	10^{-08}	45,000	982,100	40,000	11.11	95.92
F2	10^{-08}	64,000	182,000	58,000	9.37	68.13
F3	10^{-08}	158,000	NA	140,000	11.39	NA
F4	10^{-08}	49,000	100,000	35,000	28.57	65.00
F5	10^{-08}	72,000	144,000	63,000	12.50	56.25
Average AR		76,600	352,025	67,200	14.58	71.32

function. The average acceleration rate of CS-Jaya is 14.58 and 71.32 with respect to Cuckoo search and Jaya algorithm, respectively.

4.4 Results and Comparisons of the CS-Jaya Search Algorithm with GWO, PSO-GWO, and jDE

This section compares the proposed CS-Jaya to other popular population-based search algorithms, GWO, PSO-GWO, and one enhanced DE variants “jDE” [29]. The outcomes are expressed in Table 4 as the average NFE of 100 runs with errors fixed 10^{-08} for all test functions. For GWO, PSO-GWO, and jDE, parameter sets and numerical outcomes are taken from [29–31]. From Table 4, it is clear that CS-Jaya produces the best results the fastest when compared to other algorithms. Table 4 also includes each algorithm’s rank-wise performance. The average rank of GWO, PSO-GWO, and jDE is 3.6, 2.2 and 3 while the average rank of CS-Jaya is 1.2 which

Table 4 Comparison of CS-Jaya with GWO, PSO-GWO, and jDE in term of mean NFE

Function	NFE				Rank			
	CS-Jaya	GWO	PSO-GWO	jDE	CS-Jaya	GWO	PSO-GWO	jDE
F1	40,000	42,000	41,500	60,000	1	3	2	4
F2	58,000	85,000	60,000	83,000	1	4	2	3
F3	140,000	121,000	150,000	100,000	2	4	3	1
F4	29,000	66,000	32,000	63,000	1	4	2	3
F5	63,000	72,100	70,000	91,000	1	3	2	4
Average	66,000	77,220	70,700	79,400	1.2	3.6	2.2	3

demonstrating the superiority and robustness of the suggested CS-Jaya performance over the competition.

5 Conclusion

In this study, a new algorithm called “CS-Jaya” that combines the Cuckoo and Jaya algorithms is put up as a solution to problems involving global optimization. Two algorithms are combined in a methodical manner so that the benefits of both can be taken advantage of to increase efficiency in terms of convergence speed and result correctness. The average error and average NFE on 5 test functions for the proposed CS-Jaya. The outcomes of CS-Jaya are contrasted with those of the three well-known algorithms GWO, PSO-GWO, and jDE. The comparisons and results have demonstrated CS-Jaya’s superiority over the other algorithms.

References

1. Dehghani M, Trojovsky P (2021) Teamwork optimization algorithm: a new optimization approach for function minimization/maximization. *Sensor* 21:4567. [https://doi.org/10.3390/s21134567\(2021\)](https://doi.org/10.3390/s21134567(2021))
2. Sharma SK, Kumar V, Katal N, Singh P (2021) Multiarea economic dispatch using evolutionary algorithms. *Hindawi mathematical problems in engineering*
3. Kumar V, Sharma V (2021) GAMS environment based solution methodologies for ramp rate constrained profit based unit commitment problem. *Iranian J Sci Technol Transac Electr Eng* pp 1325–1342
4. Yang XS, Deb S (2009) ‘Cuckoo search via Lévy flights. In: *Proceedings world congress on nature and biologically inspired computing—NaBIC*, Coimbatore, India, pp 210–214
5. Single S, Jarial P, Mittal G (2015) Hybridisation of cuckoo search & artificial bee colony optimization for satellite image classification. *Int J Adv Res Comput Commun Eng* 4(6)
6. Kanagaraj G, Ponnambalam SG, Jawahar N (2013) A hybrid cuckoo search and genetic algorithm for reliability-redundancy allocation problems. Accepted

7. Jianwen G, Zhenzhong S, Hong T, Xuejun J, Song W, Xiaohui Y, Guoliang Y, Guohong W (2015) Hybrid optimization algorithm of particle swarm optimization and cuckoo search for preventive maintenance period optimization. Hindawi Publishing Corporation, Accepted
8. Victoria YM, Rodrigo MS, Carolina M (2018) Cuckoo Search approach enhanced with genetic replacement of abandoned nests applied to optimal allocation of distributed generation units. IET J
9. Kamoona M, Patra J, Stojcevski A (2018) An enhanced cuckoo search algorithm for solving optimization problems. Conference paper
10. Al-Abaji MA (2020) A literature review of cuckoo search algorithm. J Educ Pract 11(8)
11. Fister JRI, Yang X, Fister D, Fister I (2020) Cuckoo search: a brief literature review. Part of studies in computational intelligence book series. SCI 11:49–62
12. Yang XS, Deb S (2013) Cuckoo search recent advances and applications. Neural Comput Appl 24(March):169–174
13. Shehab M, Khader AT, Al-Betar MA (2017) A survey on applications and variants of the cuckoo search algorithm
14. Rao RV (2016) Jaya: 'A simple and new optimization algorithm for solving constrained and unconstrained optimization problems.' Int J Ind Eng Comput 7:19–34
15. Mishra S, Ray PK (2016) Power quality improvement using photovoltaic fed DSTATCOM based on Jaya optimization. IEEE Trans Sust Energ 99:1–9
16. Gong C (2017) An enhanced Jaya algorithm with a two group adaption. Int J Comput Intell Syst 10:1102–1115
17. Yu K, Liang J, Qu B, Chen X, Wang H (2017) Parameters identification of photovoltaic models using an improved JAYA optimization algorithm. Energ Convers Manage 150, 742–753 (2017)
18. Gao K, Zhang Y, Sadollah A, Lentzakis A, Su R (2017) Jaya harmony search and water cycle algorithms for solving large-scale real-life urban traffic light scheduling problem. Swarm Evol Comput 37:58–72
19. Rao RV, More KC (2017) Design optimization and analysis of selected thermal devices using self-adaptive Jaya algorithm. Energ Convers Manage 140:24–35
20. Singh SP, Prakash T, Singh V, Babu MG (2017) Analytic hierarchy process based automatic generation control of multi-area interconnected power system using Jaya algorithm. Eng Appl Artif Intell 60:35–44
21. Rao RV, Saroj A (2017) Economic optimization of shell-and-tube heat exchanger using Jaya algorithm with maintenance consideration. Swarm Evol Comput 116:473–487
22. Rao RV, Saroj A (2017) A self-adaptive multi-population based Jaya algorithm for engineering optimization. Swarm Evol Comput 37:1–37
23. Rao RV, Saroj A (2018) Multi-objective design optimization of heat exchangers using elitist-Jaya algorithm. Energ Syst 9:305–341
24. Yu J-T, Kim C-H, Wadood A, Khurshaid T, Rhee S-B (2019) Jaya algorithm with selfadaptive multi-population and lévy flights for solving economic load dispatch problems. IEEE Access 7:21372–21384
25. Li X-T, Yin M-H (2013) A hybrid cuckoo search via Levy flights for the permutation flow shop scheduling problem. Int J Prod Res 51(16):4732–4754
26. Zumofen G, Klafter J, Shlesinger M-F (1999) Levy flights and Levy walks revisited. In: Anomalous diffusion from basics to applications: proceedings of the XIth Max Born Symposium Held at Łądek Zdroj, Poland, 20–27 May 1998, vol. 519 of *Lecture Notes in Physics*, pp. 15–34, Springer, Berlin, Germany, (1999)
27. Basak J, Roy S, Chaudhuri SS (2015) Benchmark function analysis of cuckoo search algorithm. In: Information systems design and intelligent applications, advances in intelligent systems and computing 339
28. Civicioglu P, Besdok E (2011) A conceptual comparison of the cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. Springer
29. Kumar P, Sharma A (2022) MRL-JAYA: a fusion of MRLDE and JayaAlgorithm. Palestine J Math 11(Special Issue I), pp 65–74

30. Marichelvam MK, Prabaharan T, Yang XS (2014) Improved cuckoo search algorithm for hybrid flow shop scheduling problems to minimize makespan. *Appl Soft Comput J* 19:93–101
31. Mirjalili S, Mirjalili S, Lewis A (2013) Grey wolf optimizer. *Advances in engineering software*, Elsevier

Plant Leaf Disease Detection Using ResNet



Amit Kumar, Manish Kumar Priyanshu, Rani Singh, and Snigdha Sen

Abstract The agriculture sector can be considered as the backbone for any developing economy. To obtain the maximum yield from the crops, it is required that farmers should be provided with the best technologies and methodologies. Artificial intelligence is having its vast applications in various sectors. Due to its ability to perceive the problems, developing the appropriate reasons for that and to establish optimal solutions for it, artificial intelligence can act as a great aid in addressing the diseases of crops. Throughout this paper, we used the ResNet9 architecture, which is a neural network concept CNN (Conv) in collaboration with this approach, which seems to be effective in classification tasks. This paper is also able to tell the difference between of healthy and the sick. As a result, recognizing leaf plants as damaged or healthy can help all end users, as they will be able to identify the best solution for afflicted plants, leading to higher major crops and economic benefits.

Keywords Deep learning · ResNet9 · CNN · Gradio · Plant leaf disease

1 Introduction

One of the main economic sectors in India is agriculture. For thousands of years, it has existed throughout the nation. As it evolved over time, nearly all of the conventional farming techniques were replaced by the employment of modern machinery and technology. In addition, some small farmers in India continue to employ antiquated conventional farming techniques due to a lack of funding. Furthermore, this is the only industry that has helped both its own and the rest of the economy of the nation thrive. We have been practicing agriculture for countless years, although it has long been in its infancy. As a result of our independence, we also used to import food grains from other nations to meet our needs. However, the majority of people

A. Kumar · M. K. Priyanshu (✉) · R. Singh · S. Sen

Department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India

e-mail: rmanishsingh111@gmail.com

S. Sen

e-mail: snigdha.sen@gat.ac.in

are unaware of the crop loss we are experiencing as a result of the disease and poorly affected leaf that is causing crop loss. Depending on the traits and properties that each leaf possesses, it might be of numerous varieties. As a result, spotting illnesses and harmed plants is now the most major factor to achieve in order to avoid catastrophic crop loss. Manually, distinguishing the type of disease a plant leaf has will be quite tough and getting the exact disease would take a long time. As a result, it should be easy to classify these disorders at any time. It is both simpler and less expensive to automatically detect illnesses by looking at their symptoms on plant leaves. Given that it makes use of statistical machine learning and image processing algorithms, the recommended method for plant disease diagnosis is computationally less expensive and necessitates less time for prediction than previous deep learning-based systems. We employed a hybrid approach in this study that combines ResNet9 architecture and CNN, both of which are neural network concepts in Deep Learning. Gradio application enables the serving of the sample using a basic web server, which can be applied to produce a sharing link. To identify the plant illness, we created a web application based on Gradio and deployed it both on Heroku and on a local host (free cloud hosting server). Our system's predictions are stated. It indicates that the disease was successfully discovered by the system. The fundamental objective of this architecture is to extract features and categorize healthy and diseased plant leaves based on those features. It has a more efficient algorithm and improved accuracy. Considering the huge applications and potential of machine learning and deep learning in crucial domains [1–7] we experimented with ResNet architecture for our leaf disease detection task.

The following is a list of the sections that make up the manuscript. The literature reviewed survey is covered in Sect. 2. Section 3 covers the proposed Methodology Sect. 4 contains the experimental setup and results discussion, as well as graphs and types. Finally, we will make some closing remarks.

2 Literature Survey

This part contains the studied results of the impacted plant leaf, which has an impact on crop production, as well as descriptions of other researchers' work and the outcomes of various algorithms. Pesticides are advised based on [8] crop condition identification. The model was trained on both healthy and afflicted photos, and the dataset was kept. Researchers in [9] employed deep learning and image processing to classify plant leaf disease. The dataset is examined again, this time with both impacted and unaffected plant leaves, and they are then classified into acceptable classifications. Khirade et al. attempted to detect plant diseases using back propagation neural network (BPNN) and digital image processing methods in 2015 [10]. Different methods for detecting plant disease have been developed by authors employing the leafy pictures. Algorithms to divide up the affected leaf area, border, and spot detection are employed.

When they research, Medha Wyawahare and Shiroop Madiwalar conducted an examination various image-processing techniques for identifying plant diseases [11]. The ability to detect plant illness using color and texture traits was examined by authors. On 110 RGB pictures in a dataset, they tested their algorithms. The characteristics gleaned for classification were the image's a measure and its standard deviation, the RGB and YCbCr channels' means and standard deviations, and the properties of the grey level cooccurrence matrix (GLCM) filtered with a Gabor convolution. The classification process employed a classifier that uses support vector machines. The authors concluded that the GLCM properties can accurately identify normal leaves. Gabor filter and color characteristics, nonetheless, are regarded as the most effective for diagnosing leaf spots and leaves with anthracnose, respectively. They've accomplished utilizing every feature that was retrieved, the maximum precision was 83.34%.

Imaging with hyperspectral technology in the task of detecting plant diseases was demonstrated by Peyman Moghadam et al. [12]. The visible, near-infrared (VNIR), and short-wave infrared (SWIR) spectrums were used in this investigation. K-means has been employed by authors spectral clustering technique for leaf segmentation. They have to take the grid out of the hyperspectral photos, an unique grid removal algorithm has been presented. Incorporating vegetation indices with VNIR spectral data, authors have attained an accuracy of 83% range and 93% accuracy across the whole spectrum.

Convolutional neural networks were utilized by Garima Shrestha and colleagues to determine the plant disease [13]. With 88.80% accuracy, the authors could effectively list and describe 12 phytopathogens. RGB images with 3000 dpi photos were utilized as the dataset for experimenting. The network consists of three layers for convolution and pooling. The network now has powerful processing costs.

Various test was carried out to assess the effectiveness of the newly constructed model. The writers and researchers of [14] advocated the use of an artificial neural network (ANN) technique for categorization. They were mostly interested in the color shifts in the photographs. For better and efficient results, the authors of [15] investigated the histogram of directed gradient operation and extracted the characteristics that are used for classification model using SVM and ANN algorithms. Therefore, our main objective is to employ a hybrid approach (ResNet9 and CNN) that outperforms all other papers currently being examined. With just one click after uploading an image, Gradio serves as the user interface for websites that identify the type of diseased plant. Additionally, it can be applied to mobile applications.

3 The Proposed Methodology

The proposed system consists of two phases, first is training phase which contains Data Collection and Data preprocessing and feature extraction. Second is Implementation phase which includes System Architecture and Classification Algorithms.

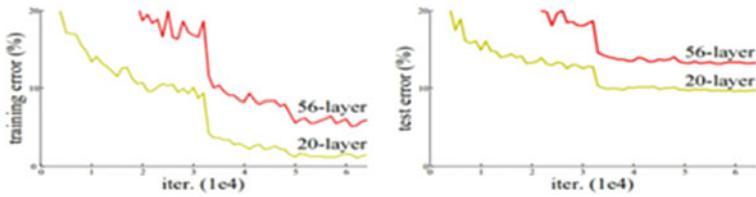


Fig. 1 Efficiently trained networks with 100 layers and 1000 layers also



Fig. 2 Some crop leaf images from dataset

The workflow of system architecture as shown in Fig. 3, in which first we collect a suitable plant leaf dataset (Fig. 2) for the training, then data pre-processing will be done on that dataset to make that data fit for the machine learning model which is a really necessary step before training and then visualize the dataset on graphs and plot the relationship between various dataset attributes. Then, we will train the machine learning model with the pre-processed dataset for predicting the correct plant leaf diseases with suitable machine learning model which gives the highest accuracy on prediction. The error vs iteration graph has been shown in (Fig. 1).

3.1 *How the Proposed Method is Different from the Existing Methods*

- Method used in this paper is ResNet 9 which gives an accuracy of 99.3%, compared to other techniques used earlier ResNet9, is quite high.
- Won first place in the ILSVRC 2015 classification competition with a top-5 error rate of 3.57%
- Won the first place in ILSVRC and COCO 2015 competition in ImageNet Detection, ImageNet localization, Coco detection, and Coco segmentation.

- Replacing VGG-16 layers in Faster R-CNN with ResNet-101. They observed relative improvements of 28%
- Widely known and used architectures are Alexnet, VGG-16, VGG19 and others. The model will learn more complex features from images by increasing the number of layers, but they got to know that a 56 layer network is performing very bad than 20 layer network even on the training data, as you can see from below image.

3.1.1 Training

Data collection

The data related are collected form the Kaggle website which provides different kinds of plant leaf images with healthy and unhealthy leaf images. The data collected basically to train the model. This dataset consists of about 87 K rgb images of healthy and diseased crop leaves which is categorized into 38 different classes. There are an equal amount of healthy and diseased photos in the study. The total dataset is divided into 80/20 ratio of training and validation set preserving the directory structure.

Data Pre-processing and Feature extraction

A crucial duty in any machine character system is data gathering. Prior to feature extraction, certain noise level should be eliminated to obtain accurate findings. Pre-processing an image removes unwanted distortions and increases key aspects that are crucial for further processing and analysis tasks. It covers picture enhancement, color space conversion, and image segmentation.

Step in image analysis is a technique used to make an image's representation more concise, meaningful, and understandable. After segmentation, the sick component, or the area of interest, was excised. The meaning of a particular sample can be ascertained using the relevant features that are retrieved in the following phase. Actually, color, form, and texture aspects are typically included in picture attributes. The majority of academics are currently focusing on plant leaf texture as the key characteristic for classifying plants. Plant diseases are divided into distinct categories using textural properties.

3.2 *Implementation*

System Architecture

Design and architecture demonstrate how the system interacts and how control proceeds from one cycle point to another. A system's structure, behavior, and other aspects are all defined by its conceptual model, or system architecture as shown in Fig. 3. The Basic Methodology describes the control flow of the whole model.

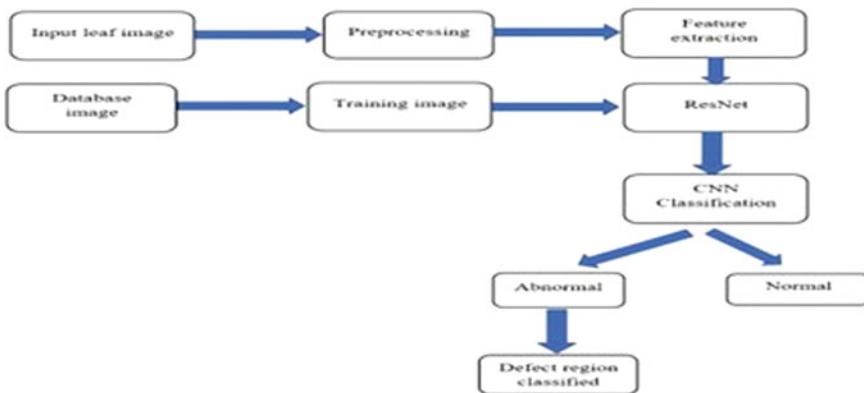


Fig. 3 Basic methodology

Classification Algorithms

A software procedure was created that uses various neural network classifiers for training and testing. These are the methods for classifying texture features.

ResNet9

Artificial neural networks include residual neural networks (ResNet). ResNet central concept is the introduction of a “identify quick access link” that omits one or more levels as shown in Fig. 4. A residual block is a tower of layers configured so that each layer’s output gets added to a layer further down the stack. After combining it with the output of the equivalent layer in the main path, the nonlinearity is then applied.

Convolutional Neural Network Model

Artificial neural networks with a unique architecture generally known as convolutional neural networks (CNN). CNN makes use of some visual cortex properties. Image categorization is one of the most well-liked applications of this architecture. Convolutional layers and pooling layers, which make up convolutional neural networks, are two extremely basic building blocks. For a specific computer vision issue, there are practically limitless ways to assemble these layers, despite their simplicity.

To test the results in this project, we are going to implement UI interface (Gradio) in which the image has to be uploaded, and usage of the algorithm is implemented; images are used to predict the plant diseases.

4 Experimental Setup and Result

We used a dataset collected called Plant Village dataset compiled by Sharada P Mohanty at AI to create the database containing healthy and afflicted plant leaves. We utilized Anaconda, TensorFlow for classification, which automatically discovers and can train using many designs of neural networks available within its zoo. We employed the ResNet9 architecture model in this work, which is more accurate and identifies faster than other models. The first stage is ResNet9, and the results are passed to the next stage, CNN. The settings and parameters of TensorFlow are kept in protobuf files, which were also installed. Training data, which contains both images and annotations, is created in the TF Records file format, and a collection of records is formed for training and testing. The content will be portrayed as affected or fine in images and graphs. This experiment's specificity and learning rate and the style feature it. Finally, the middle aged a grade to the disease detection and type.

We've created a method that uses computer vision and has an average accuracy of 99.2% for detecting plant diseases. Statistical processing of thumbnails with learning algorithms are also used to make the suggested method computationally efficient model of education.

5 Results

Figure 5 describes the number of images per each class of plant diseases. As it clearly visible, there are number of plants with different disease (Fig. 5).

Figure 6 describes healthy and affected plant leaf images. As it is clearly visible, the pores of leaf and the part which is affected and also the color of the leaf says it all, whereas the healthy leaf has a clear picture containing one color of whole leaf and doesn't have any affected.

Figure 7 describes graphs 1. Accuracy and No. of epochs 2. Learning rate and Batch No. The performance of the model is shown, and the variation it's achieving 3. validation loss.

In Figure 8, a plant leaf is uploaded and based on the model, it predicts the respective plant disease. The GUI has been developed using Gradio, a package to

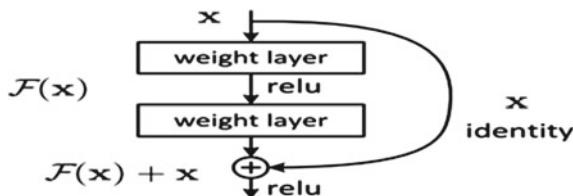


Fig. 4 Residual block

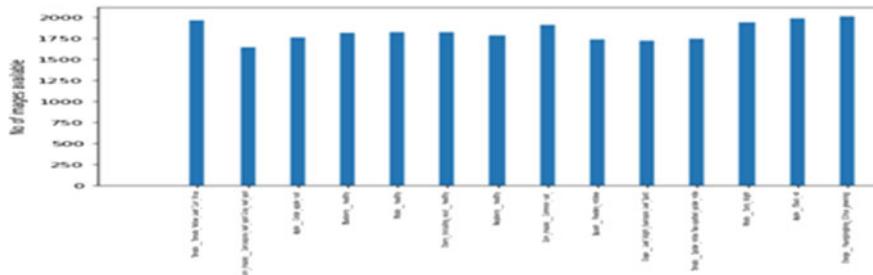


Fig. 5 Histogram

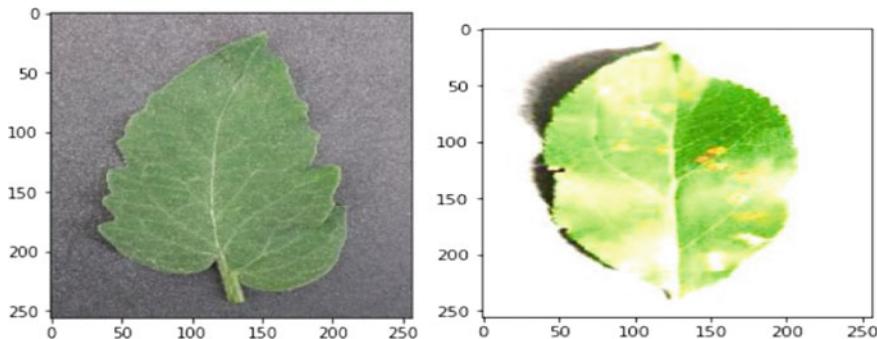


Fig. 6 **a** Healthy leaf, **b** affected leaf

develop web interface. The performance of the proposed method has been compared with other existing techniques and shown in (Table 1).

6 Conclusion

The main motive of this study was to brief the applications and available techniques of artificial intelligence to solve the problems of farmers in getting the required yield. The basic functions and understanding of ResNet was to give better results in terms of accuracy and faster process with neural net. Outputs show the flexibility of ResNet in detecting plant leaf disease and the type. In order to identify plant leaf disease, every end user will find the Gradio user interface to be very helpful and simple. Future researchers should organize a proper dataset covering all arena of agriculture and enhance the available technologies to increase the productivity of primary sectors [17, 18].

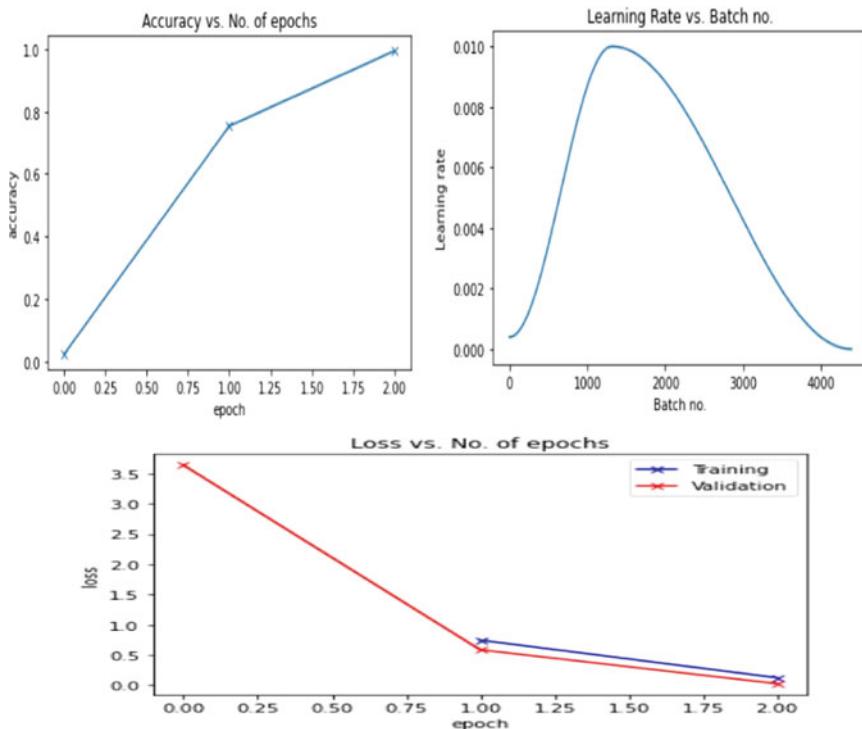


Fig. 7 Graphs

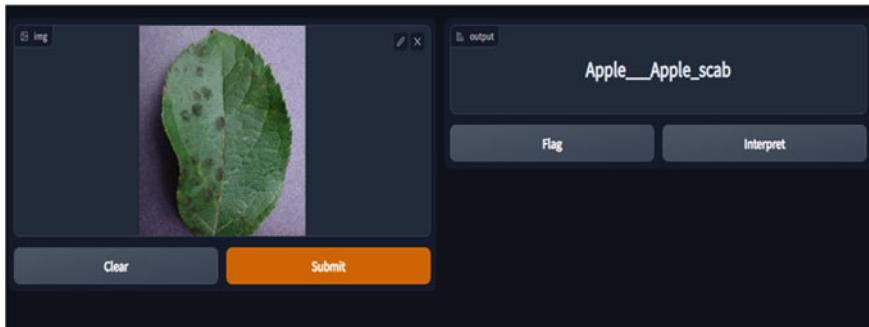


Fig. 8 Predicted plant disease

Table 1 Comparison of proposed system with other existing system

Author	Khirade et al. [10]	Madiwala r et al. [11]	Moghadam et al. [12]	Sharath DM et al. [16]	Garima Shrestha et al. [13]	Proposed method
Algorithm	Digital image processing and BPNN	Digital image processing and SVM	Hyperspectral imaging and SVM	Digital image processing	CNN	Resnet9 architecture with CNN
Accuracy (%)	–	83.34	93	–	88.80	99.3

References

1. Sen S, Agarwal S, Chakraborty P et al (2022) Astronomical big data processing using machine learning: a comprehensive review. *Exp Astron.* <https://doi.org/10.1007/s10686-021-09827-4>
2. Sen S et al (2021) Analysis, visualization and prediction of COVID-19 pandemic spread using machine learning. In: innovations in computer science and engineering. pp 597–603. Springer Singapore
3. Sen S, Singh KP, Chakraborty P (2023). Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network. *New Astron* 99:101959
4. Sen S, Amrita I (2022). A transfer learning based approach for lung inflammation detection. *Advanced Techniques for IoT Applications*. In: Proceedings of EAIT 2020. Springer Singapore
5. Monisha R et al (2022). An approach toward design and implementation of distributed framework for astronomical big data processing. *Intelligent Systems*. In: Proceedings of ICMIB 2021. pp 267–275. Singapore: Springer Nature Singapore
6. Mayank K, Sen S, Chakraborty P (2022). Implementation of cascade learning using apache spark. In: 2022 IEEE international conference on electronics, computing and communication technologies (CONECCT). IEEE
7. Khasnis NS, Sen S, Khasnis SS (2021). A machine learning approach for sentiment analysis to nurture mental health amidst COVID-19. In: Proceedings of the international conference on data science, machine learning and artificial intelligence
8. Poonguzhal R, Vijayabhanu A (2019) Crop condition assessment using machine learning. *Int J Recent Technol Eng (IJRTE)* 7
9. Sladojevic S et al (2016) Deep neural networks-based recognition of plant diseases by leaf image classification. *Comput Intell Neurosci* 2016
10. Khirade SD, Patil AB (2015) Plant disease detection using image processing. In: 2015 international conference on computing communication control and automation. IEEE
11. Madiwalar SC, Wyawahare MV (2017) Plant disease identification: a comparative study. In: 2017 international conference on data management, analytics and innovation (ICDMAI). IEEE
12. Moghadam P et al. (2017) Plant disease detection using hyperspectral imaging. In: 2017 international conference on digital image computing: techniques and applications (DICTA). IEEE
13. Shrestha G, Das M, Dey N (2020) Plant disease detection using CNN. In: 2020 IEEE applied signal processing conference (ASPCON). IEEE
14. Shah N, Jain S (2019) Detection of disease in cotton leaf using artificial neural network. In: 2019 amity international conference on artificial intelligence (AICAI). IEEE
15. Ramesh S et al (2018) Plant disease detection using machine learning. In: 2018 international conference on design innovations for 3Cs compute communicate control (ICDI3C). IEEE
16. Sharath DM, Kumar SA, Rohan MG, Prathap C (2019). Image based plant disease detection in pomegranate plant for bacterial blight. In: 2019 international conference on communication and signal processing (ICCP) pp 0645–0649. IEEE

17. Al Bashish D, Braik M, Bani-Ahmad S (2011) Detection and classification of leaf diseases using K-means-based segmentation and. *Inf Technol J* 10(2):267–275
18. Ferentinos KP (2018) Deep learning models for plant disease detection and diagnosis. *Comput Electron Agric* 145:311–318

Automatic Infographic Builder Using Natural Language Statements



Chetali Neema and Anuradha Purohit

Abstract “A picture is worth a thousand words,” visual memories are the strongest memories. Combining graphics with information, Infographics are an effective and eye-catching way to deliver information in a memorable manner. Several kinds of authoring tools have been developed for creating infographics, but casual users either find them difficult to use or do not invest their time in learning to understand the working of these tools. In this paper, an alternative approach for the creation of Infographics from user friendly natural language statements has been proposed. The approach incorporates dynamic layouts, designs, and graphics. The approach takes a natural, proportion-related statement as an input and analyze it using natural language processing algorithms. It dynamically selects graphics and images for infographic creation using a Deep Neural Network model. Taking into consideration, the various visual aspects of Infographic creation, the approach generates a set of Infographics from which the user can select and refine according to their use. The training of the system performed on the MS COCO dataset and the parameter results obtained after testing demonstrates the effectiveness of the system. With an increasing use of visual Infographics in various domains, this system proves to be efficient and time-saving as illustrated by the generated samples and results.

Keywords Automatic visualization · Natural language processing · Machine learning

1 Introduction

Information Graphics also known as Infographics is one of the important techniques for data visualization. Infographics are the combination of elements like graphics, icons, and images, with various data visualization methods. Infographics are effective and engaging as it combines visuals with data which is one of the most memorable ways of message delivery. Infographics find applications in a wide variety of domains

C. Neema · A. Purohit (✉)

Department of Computer Engineering, Shri G. S. Institute of Technology and Science, 23 Sir M. Visvesvaraya Marg, Indore, Madhya Pradesh 452003, India
e-mail: anuradhapurohit78@gmail.com

including education, business, mass media, health care, and entertainment industry. However, it is intricate to build a professional infographic. A professional infographic requires designing and analytical skills for its creation and is often a time-consuming process.

Various authoring tools [16] for infographic creation has been proposed. These authoring tools are easy to use, fast in creating infographics, but targets mainly on advanced users like professional editors, data scientists, graphic designers, etc. Besides these advanced users, there is also a vast majority of casual users [17] including students, business employees, etc., which finds this tools difficult to understand and hence less useful. Based on various surveys and samples, the most widely expressed information is based on proportion facts (e.g., “More than **13%** of the dog bites cause rabies disease.”). Thus, the paper aims to focus on a sub-field of infographics, i.e., proportion-related infographics and to build a system that automatically generate Infographics from proportion-related statements.

The proposed approach takes a natural language statement as input analyze it using various natural language processing algorithms and uses [13] real time object detection algorithm for the generation of dynamic graphics and images. Based on the results obtained, the proposed system provides an efficient way for converting simple user-friendly statements to engaging Infographics in an easy and speedy way along with increased accuracy.

Section 1 gives a brief introduction of the subject and describes the proposed approach. Section 2 describes Infographics and their classification along with summarizing the various works and studies in Visualization and Infographics domain. Section 3 describes the proposed approach for Automatic Generation of Infographics using Natural Language Statements. Section 4 consists of the sample results and discussions. Finally, Sect. 5 discusses the future aspects and scope of the proposed approach.

2 Related Work

2.1 *Information Graphics*

Infographics also known as Information Graphics are one of the most effective techniques of data visualization which can deliver very complex data content in an attractive and appealing manner. Due to the huge demand and use, Infographics find applications in a wide variety of domains including education, business, mass media, health care, and entertainment industry.

Infographics are capable of representing a broad range of information. However, a valid infographic should be able to deliver completely at least one message along with including one or more graphical elements. Also, it should not be divisible into smaller units conveying the same meaning.

2.2 *Literature Survey*

Infographics are one of the most effective techniques of data visualization which can deliver very complex data content in an attractive and appealing manner. Research in Infographic generation started from programming which uses languages and libraries for data visualization. Later developments focused on semi-automatic way to generate infographics. The semi-automatic way seemed more user-friendly. Later research in Infographic generation focuses on automatic infographic generation very useful to casual users.

Key et al. [1] proposed “Vizdeck” which was the first visualization recommendation system. They proposed a technique which was based on disorganized relational dataset. The visualizations were recommended by training a model that can learn correlations among properties of visualization and data records to predict user-preferred charts. Luo et al. [2] presented a novel system for automatic data visualization tasks called as “DeepEye.” They worked toward classifying a visualization as good or bad by training a binary classifier. Also, their system used a supervised machine learning model for determining the quality of visualization. Later, Srinivasan et al. [3] introduced “Voder,” which used a series of statistical functions to combine natural language generation techniques into their system.

Battle et al. [4] proposed an approach that can automatically extract visualizations from the Internet and can describe visualizations with annotation. They concluded that charts like bar charts, line charts, scatter charts, and geographic maps are the most frequently used visualizations. Hu et al. [5] proposed “VizML,” a model trained via 1,000,000 dataset-visualization pairs that learns visualization design choices using neural networks with increased accuracy. Cui et al. [6] proposed “Text-to-Viz,” a model for the automatic generation of infographics using natural language statements. Their system takes a proportion related statement as an input and converts it into a set of professional infographics using a combination of Convolutional Neural Network and Conditional Random Field (CNN and CRF).

Cao et al. [7] proposed “VisGuide,” an assistive data exploration system that helps users to create context-based sequence trees for visualization. Their system was also capable of recommending meaningful charts according to user’s choice. Lai et al. [8] proposed “Vis-Annotator” that uses [18] a Mask R-CNN model to identify and extract visual elements in the target visualizations, along with their visual properties. Liu et al. [9] proposed “AutoCaption,” which aims to identify the correlations among visual elements using a multi-layer perceptron classifier and a 1D convolutional residual network. The system can analyze and extract essential features from charts and the relationship between various visual elements. Qian et al. [10] presented a system called “Retrieve-Then-Adapt,” an example-based automatic generation for proportion-related infographics that aimed to generate infographics by automatically imitating online examples. Suitable examples are fetched, and an initial draft is then produced which can be updated with user information later. The approach used recursive neural networks for obtaining better results.

3 Proportion-Related Information

Due to the dominance of proportion-related information used in day-to-day life, the proposed system focuses on Infographics based on proportion related facts. This domain comprises two significant areas, i.e., the Text Space and the Visual Space of the infographic.

3.1 *Text Space*

One significant role of the proposed system is to allow its users to visualize and provide information using their natural daily language rather than any formal high level language like XML, JSON, etc. This eases out the process of creating infographics by hiding away the technical details of the system.

From the general trends observed, most statements are of the form $x\%$, x out of y , x in y . All such statements depicts proportion related information. Thus, the Text Space of the proposed system involves statements of the above form which are used as an input to the proposed system. The input statements then undergo text processing to extract important features from the statements.

3.2 *Visual Space*

The four important elements of the visual space are the design aspects of Infographic creation [6], i.e., Layout, Description, Graphic and Color, which are summarized as follows.

Layout. Layout, or the blueprint, comprise two significant visual elements mainly the graphic and the description. The layout for infographics can be arranged in horizontal manner, vertical manner, or in a tiled manner. Also, description can be overlaid on top of the graphic if required.

Description. Description is an important element in data-driven infographics. While the initial description is a normal user-friendly statement, the system aims to extract important information from the input statement including the number (proportion) and subject of the statement to build the final infographic and deliver the complete message.

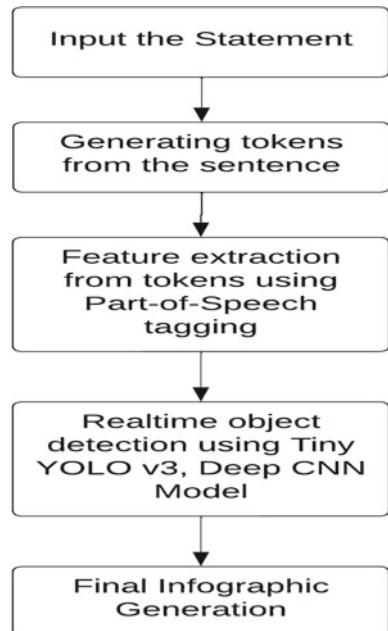
Graphic. This element of the visual space involves the selection and designing of the visual elements called graphics. Since, different graphics conveys different message to the Infographic and hence to the user, the graphic which is semantically most accurate to the content and description of the original statement needs to be selected

Color. Infographics are attractive and visually appealing due to the different usage of colors. Our system aims to assign various colors to the Infographic including the background area. Also, since our system emphasize on proportion-related Infographics, the number part of the description is highlighted in a suitable harmonic color to the description.

4 Proposed System

The paper proposes a system that comprises two main modules, namely the text analyzer and the visual generator. First, a proportion-based textual statement, such as “More than 80% of students got placements.” is provided by the user. Then, the text analyzer performs tokenization by converting the words to the corresponding tokens. Then, it identifies and tags the essential parts in the statement, like proper nouns, pronouns, etc. Then, the tagged words are searched on the web for images. Related images from the web get downloaded on the system. The images are then fed to real-time object detection, Tiny YOLO v3 algorithm for real-time detection of objects. The model classifies and detects the objects present in the images with their accuracy. The detected image along with the original statement is then fed into the visual generator for infographic generation. Figure 1 shows the block diagram of the proposed approach. The users can directly export the infographic into their reports and presentations and can refine further as per their need.

Fig. 1 Block diagram of infographic builder



4.1 Text Analyzer

Given a statement based on proportion facts, the text analyzer undergoes the below described phases for extracting the useful information from the input statement.

Tokenization. First, the input statement is converted into a sequence of tokens. For example, given the statement, “Less than 40% of students got placements.” The sequence of tokens will be “less,” “than,” “40,” “%,” “of,” “students,” “got,” “placements,” “.”. Next, all the stop-words including “,” “.” etc., are removed. Then, the extracted tokens are tagged using Natural Language Processing Part of Speech Tagging. [11] The tag describes whether the word is a noun, adjective, verb, and so on. For example, for the statement “More than 80% of students got placement.” output would be [(“students,” “NNS”), (“placement,” “NN”)], where NNS stands Plural Noun and NN stands for Singular Noun.

Image Processing. The extracted words are then searched on the web to get images for the tagged words. For achieving this purpose, our system first requests for an HTTP connection with Google’s API and search for the related images. The response is a set of related images for the extracted words fetched from the huge corpus of images from the web which gets downloaded in the current working directory of the client’s system.

Real-time Object Detection. The set of images are now fed to real-time object detection model, Tiny YOLO V3. This algorithm [12] applies a single neural network to the full image by dividing the image into regions and predicts the bounding boxes and probabilities for each region.

The model takes the set of images downloaded on the system as an input and evaluates and detects objects present in it. Accordingly, the model classifies the objects and predicts the accuracy of the various objects present in the image. The obtained image and the original statement are now fed to the visual generator for Infographic building process.

4.2 Visual Generator

The output provided by the text analyzer module is fed as an input to the visual generator that builds multiple infographic candidate designs based on pre-designed templates and blueprints.

Layout. The proposed system creates layouts and templates using the Pillow(PIL) library of Python. These layouts and templates are then used for further infographic generation.

Description. The descriptive elements present in the statement are identified by the text analyzer. The proposed system involves descriptions to be broken down to different lengths and different components according to the layout.

Graphic. The graphic module of the proposed system aims to provide the various graphical elements, including images, icons, charts, etc., for the infographic

Color. The color module of the proposed system aims to assign a set of harmonic colors for the various elements of the infographic including various icons, figures, and descriptions.

5 Experimentation and Results

The Infographic Builder system takes a natural, proportion-related statement as the input. It imports Python's Natural Language Toolkit (NLTK) library to convert words of the statements to corresponding tokens using `nltk.sent_tokenize()` method. It then tags the tokens with appropriate tags to extract features like noun, pronoun, etc. Using `nltk.pos_tag()` method, the system then establishes a HTTP connection using Google API to search for the image of the tagged word on the web. The image is then fed to the Tiny YOLO v3 object detection model which detects the objects in the image using `detector.detectObjectsFromImage()` method. The detection model detects and returns the object along with its percentage probability. Finally, the methods of Python's Pillow (PIL) library are used to combine graphics with the original statement to generate the final infographic.

The opportunities of the proposed approach are vast. The proposed approach facilitates the easy building of Infographics which is beneficial for casual users to adapt in their daily life as illustrated by the samples results shown in Fig. 2. The real-time object detection model, Tiny YOLO V3 used can process images at 30 frames per second with a Mean Average Precision (mAP) value of 57.9% on COCO dataset. Results of higher version like YOLO v6 algorithm reveals it better both in terms of speed and accuracy with almost a 3.9% increased accuracy and 29.4% increased speed as compared to YOLO v5 algorithm. However, the system's approach through Tiny YOLO v3 algorithm also proves itself as demonstrated by the obtained parameter results. Although, Tiny YOLO v3 algorithm lacks somewhere in terms of speed and accuracy, it is quite well capable and stable in performing real-time detection tasks. Tiny YOLO v3 also capable on running small processors proves itself in sustaining detection accuracy along with a lightweight model as compared to its higher versions. Thus, our system employed Tiny YOLO v3 as the object detection algorithm and aims to upgrade and expand it according to the requirements in the future.

Table 1 shows the various parameter results obtained by the proposed system. The precision of 0.591 indicates that our deep learning model is returning quite accurate and precise results in terms of quality of objects detected, whereas a higher recall measure of 0.812 is indicating that the rate of false negative class prediction out of all classes is low. Since, F1 Score is a harmonic mean value of recall and precision, a F1 Score of 0.685 is indicating that our model is able to balance between the true positive and false negative cases quite well.

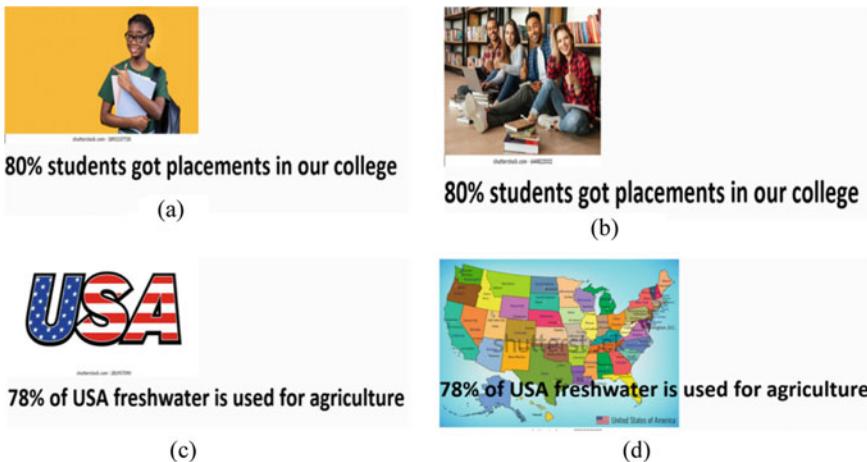


Fig. 2 Examples produced as output by our system “Automatic Infographic Builder.” **a, b** are generated from the statement: “80% students got placements in our college.” **c and d** are generated from the statement: “78% of US freshwater is used for agriculture

Table 1 Results displaying the various performance parameters

Mean average precision (mAP)	Recall	F1 score
0.591	0.812	0.685

6 Conclusion

This paper aims to build a system for the automatic generation of infographics from natural language statements. The proposed system takes a natural, proportion-based statement as input and converts it to a set of easy to use, professional infographics. The infographics generated can be selected and refined according to user’s choice and requirements. The approach used to build the Infographic Builder is easy, less time-consuming, and does not involve any complex authoring process. The sample results and the values of various testing parameters generated by the system proves the usability and the ease of adopting it in everyday working life.

In the future, the system may be expanding toward incorporating more variety of information, more layout designs, and blueprints. Also, the goal of our system will be toward implementing Infographic Builder concerned to some particular domains like education and advertisements.

References

1. Key A, Howe B, Perry D, Aragon C (2012) Vizdeck: self-organizing dashboards for visual analytics. In: Proceedings of ACM SIGMOD international conference on management of data, pp 681–684
2. Luo Y, Qin X, Tang N, Li G (2018) DeepEye: towards automatic data visualization. In: Proceedings of IEEE international conference on data engineering (ICDE), pp 101–112
3. Srinivasan A, Drucker SM, Endert A, Stasko J (2019) Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Trans Visual Comput Graph* 25(1):672–681
4. Battle L, Duan P, Miranda Z, Mukusheva D, Chang R, Stonebraker M (2018) Beagle: automated extraction and interpretation of visualizations from the web. In: Proceedings of CHI conference on human factors in computing systems, pp 1–8
5. Hu K, Bakker MA, Li S, Kraska T, Hidalgo C (2019) Vizml: a machine learning approach to visualization recommendation. In: Proceedings of CHI conference on human factors in computing systems, pp 1–12
6. Cui W et al (2020) Text-to-Viz: automatic generation of infographics from proportion-related natural language statements. *IEEE Trans Visual Comput Graph* 26(1):906–916
7. Cao YR, Pan JY, Lin WC (2020) User-oriented generation of contextual visualization sequences. In: Extended abstracts of the 2020 CHI conference on human factors in computing systems (CHI EA '20). Association for computing machinery, New York, USA, pp 1–8
8. Lai C, Lin Z, Jiang R, Han Y, Liu C, Yuan X (2020) Automatic annotation synchronizing with textual description for visualization. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for computing machinery, New York, USA, pp 1–13
9. Liu C, Xie L, Han Y, Wei D, Yuan X (2020) AutoCaption: an approach to generate natural language description from visualization automatically. In: IEEE pacific visualization symposium, pp 91–195
10. Qian C, Sun S, Cui W, Lou J-G, Zhang H, Zhang D (2021) Retrieve-Then-Adapt: example-based automatic generation for proportion-related infographics. *IEEE Trans Visual Comput Graph* 27(2):443–452
11. Github (2021) <https://github.com/IshtyM/Parts-of-Speech-Tagging>. Last accessed 30 July 2021
12. Github (2021) <https://github.com/DhamuSniper/YOLO-Object-Detection---Andrew-Ng-s-course>. Last accessed 30 July 2021
13. Kumar S, Yadav D, Gupta H, Verma OP, Ansari IA, Ahn CW (2020) A novel yolov3 algorithm-based deep learning approach for waste segregation: towards smart waste management. *Electronics* 10(1):14
14. Gupta H, Verma OP (2022) Monitoring and surveillance of urban road traffic using low altitude drone images: a deep learning approach. *Multimedia Tools Appl* 81(14):19683–19703
15. Kumar S, Gupta H, Yadav D, Ansari IA, Verma OP (2022) YOLOv4 algorithm for the real-time detection of fire and personal protective equipments at construction sites. *Multimedia Tools Appl* 81(16):22163–22183
16. Wang Y, Zhang H, Huang H, Chen X, Yin Q, Hou Z, Zhang D, Luo Q, Qu H (2018) InfoNice: easy creation of information graphics. In: Proceedings of the 2018 CHI conference on human factors in computing systems (CHI '18). Association for computing machinery, New York, USA, Paper 335, pp 1–12
17. Liu Z, Thompson J, Wilson A, Dontcheva M, Delorey J, Grigg S, Kerr B, Stasko J (2018) Data illustrator: augmenting vector design tools with lazy data binding for expressive visualization authoring. In: Proceedings of the 2018 CHI conference on human factors in computing systems (CHI '18). Association for computing machinery, New York, USA, Paper 123, pp 1–13
18. Zhu S, Sun G, Jiang Q, Zha M, Liang R (2020) A survey on automatic infographics and visualization recommendations. *Visual Inf* 4(3), pp 24–40, ISSN 2468–502X, <https://doi.org/10.1016/j.visinf.2020.07.002>

A Novel Type-2 Fuzzy Programming Approach for Solving Multiobjective Programming Problems



Animesh Biswas, Debjani Chakraborty, Bappaditya Ghosh, and Arnab Kumar De

Abstract This article introduces a novel type-2 fuzzy programming approach for solving multiobjective programming problems. The linguistic decision variables of the model are considered as type-2 fuzzy numbers, each of which is represented by four ordinary fuzzy numbers. In the solution process a Takagi–Sugeno type-2 fuzzy inference system has been developed. The outputs achieved through the inference system are then aggregated by developing an equivalent fuzzy multiobjective programming model. Finally, a weighted fuzzy goal programming is derived to find the compromise solution of the objectives in deterministic environment. A numerical example is solved to demonstrate the application potentiality of the developed method. The achieved solutions are then compared with the existing method to establish superiority of the proposed method.

Keywords Type-2 fuzzy number · Takagi–Sugeno fuzzy inference system · Linguistic variable · Multiobjective decision making · Defuzzification

1 Introduction

Involvement of fuzziness in multiobjective programming problems is inevitable in modern decision making contexts due to increasing complexities in making decisions. To tackle uncertainties associated with those problems, various multiobjective decision making methods are being developed. Most of the decision making methods developed so far are based on ordinary fuzzy sets [1, 2]. Afzal and Ramis [3] developed a novel technique for solving multiobjective optimization (MOO) problem in thermal management of battery system using fuzzy logic and particle swarm

A. Biswas (✉) · B. Ghosh
University of Kalyani, Kalyani 741235, India
e-mail: abiswaskln@rediffmail.com

D. Chakraborty
Indian Institute of Technology Kharagpur, Kharagpur 721302, India

A. K. De
Government College of Engineering and Textile Technology, Serampore 712201, India

algorithm. To minimize the emission of polluting gases, a fuzzy MOO model was developed by Hashemi [4]. The notion of adaptive fuzzy logic in the field of MOO processes was successfully introduced by Brindha and Amali [5]. Guo [6] used fuzzy self-defence algorithm to solve MOO problems in cloud computing. A computational model for selecting best strategy in building energy retrofit was proposed by Pazouki et al. [7] using fuzzy robust MOO approach.

With the advancement of decision making models, it is frequently observed that uncertainties cannot always be captured by ordinary fuzzy sets. To overcome such drawback the concept of interval valued fuzzy sets [8] were introduced. An interval fuzzy optimization technique is developed by Derghal et al. [9] to solve environmental/ economic dispatch problem. In compare to ordinary fuzzy sets, the use of interval valued fuzzy sets is very limited. For capturing uncertainties in a better way Liang and Mendel [10] developed the concept of interval type-2 fuzzy (IT2F) system. There after a large number of research works have been performed using that concept. For the evaluation of failures of ship diesel generator, an IT2F model is developed by Yucesan et al. [11] using best-worst method. Gomes and Serra [12] used IT2F concept for forecasting and real time filtering of novel Coronavirus. An innovative concept of IT2F fractional inference system has been recently developed by Mazandarani and Xiu [13]. IT2F numbers are special type of type-2 fuzzy numbers (T2FNs) [8] which cannot completely perform the role of T2FNs. As a consequence, it might have a chance of losing some information to interpret fuzzy phenomenon. It is the fact that T2FNs possess more complex behaviour than other variants of fuzzy sets. Lv et al. [14] introduced the concept of triangular type-2 fuzzy sets (T2FSs) to maintain a reasonable balance between information protection and computational complexity reduction. A two phase defuzzification process for T2FNs was introduced by Biswas and De [15]. For indoor air quality assessment Ghorbani and Zamanifar [16] used semantic knowledge concept of type-2 fuzzy ontology. In compare to IT2F sets, the development of T2FSs are limited. A review work on T2FSs has recently presented by De et al. [17]. To develop a multiobjective model, linguistic variables are represented using type-2 triangular fuzzy numbers (T2TFNs). The developed model is then solved using type-2 Takagi–Sugeno fuzzy inference system (T2TSFIS). A numerical example has been provided to illustrate the proposed method and compared with existing methods.

2 Preliminaries

2.1 *Definition [8]*

An ordinary triangular fuzzy numbers, \tilde{A} is denoted by $\tilde{A} = (a^L, a, a^R)$, where $a^L, a, a^R \in \mathbb{R}$ and is defined by the membership function as follows:

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{x-a^L}{a^R-a^L}, & a^L \leq x \leq a^R \\ \frac{a^R-x}{a^R-a}, & a < x \leq a^R \\ 0, & \text{elsewhere} \end{cases}$$

2.2 Definition [15]

It is well-known that T2FS is the generalization of interval valued fuzzy set. Thus, in T2FSs a tolerance on lower and upper boundary of the interval valued fuzzy sets are allowed. Therefore, a T2FS can be represented in terms of four ordinary fuzzy sets within its foot print of uncertainty. Hence, a T2TFN is represented in the form $\tilde{A} = (\tilde{A}^1, \tilde{A}^2, \tilde{A}^3, \tilde{A}^4)$ where $\tilde{A}^1 = (a_1^L, a, a_1^R)$, $\tilde{A}^2 = (a_2^L, a, a_2^R)$, $\tilde{A}^3 = (a_3^L, a, a_3^R)$, $\tilde{A}^4 = (a_4^L, a, a_4^R)$ are all ordinary triangular fuzzy numbers which satisfies $a_4^L \leq a_3^L \leq a_2^L \leq a_1^L \leq a \leq a_1^R \leq a_2^R \leq a_3^R \leq a_4^R$ and is shown in Fig. 1.

The primary membership function $\mu_{\tilde{A}}$ of a T2TFN can be written in the following form:

$$\mu_{\tilde{A}}(x) = \begin{cases} (\mu_{1j}^L(x), \mu_{2j}^L(x), \mu_{3j}^L(x), \mu_{4j}^L(x)) & \text{if } a_{j+1}^L \leq x \leq a_j^L \\ (\hat{\mu}_{1j}^R(x), \hat{\mu}_{2j}^R(x), \hat{\mu}_{3j}^R(x), \hat{\mu}_{4j}^R(x)) & \text{if } a_j^R \leq x \leq a_{j+1}^R, \\ 0 & \text{otherwise} \end{cases}$$

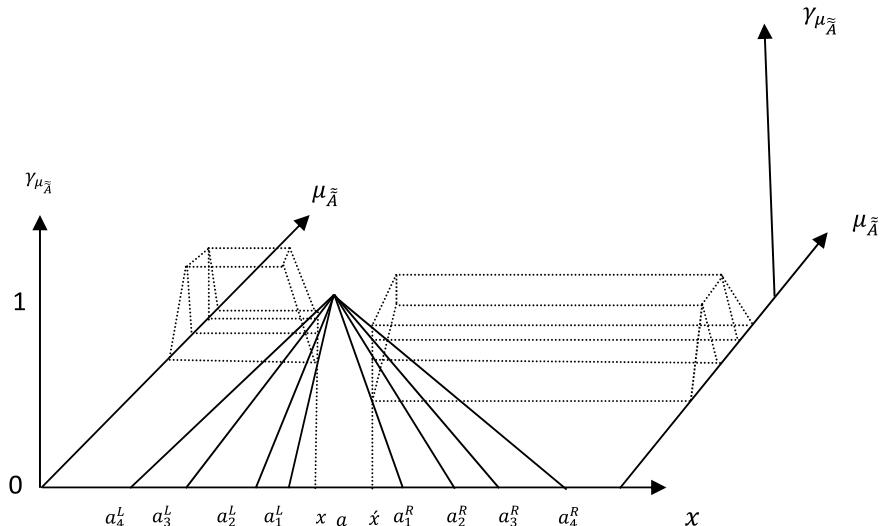


Fig. 1 T2TFN

$$\text{where } \mu_{ij}^L(x) = \begin{cases} \frac{x-a_i^L}{a-a_i^L} & i > j \\ 0 & i \leq j \end{cases}; (j = 0, 1, 2, 3; i = 1, 2, 3, 4) \text{ and } \hat{\mu}_{ij}^R(x) = \begin{cases} \frac{a_i^R-x}{a_i^R-a} & i > j \\ 0 & i \leq j \end{cases}; (j = 0, 1, 2, 3; i = 1, 2, 3, 4).$$

2.3 Definition [2]

It is assumed that all the decision variables of the proposed fuzzy multiobjective model are treated as linguistic variables. For convenience, here, a linguistic variable is considered as a variable whose state is a T2TFN. Thus, a linguistic variable is characterized by $(x, J(x), X, G, M)$, where x is the name of the linguistic variable; $J(x)$ is the set of names of linguistic values of x with each value being a T2TFN defined on X ; G is a syntactic rule for generating the names of values of x ; and M is a semantic rule for associating with each value its meaning.

It is supposed that the values of each of the linguistic variables x_1, x_2, \dots, x_n are defined in the interval $[a, b] \subset \mathbb{R}$, i.e. $X = [a, b]$ and let $J(x)$ consists of $(K + 1)$ terms as:

$$J = \{\text{low}, \text{around}(a + \beta), \text{around}(a + 2\beta), \dots, \text{around}(a + (K - 1)\beta), \text{high}\},$$

where $\beta = (b - a)/K$, $K \geq 2$ and each term \tilde{A}_i , $i = 1, 2, \dots, K + 1$ is a T2TFN, i.e. $\tilde{A}_i = (\tilde{A}_i^1, \tilde{A}_i^2, \tilde{A}_i^3, \tilde{A}_i^4)$ which may be represented using its membership functions $\{\mu_{\tilde{A}_1}^{\tilde{A}_1}, \mu_{\tilde{A}_2}^{\tilde{A}_2}, \dots, \mu_{\tilde{A}_K}^{\tilde{A}_K}, \mu_{\tilde{A}_{K+1}}^{\tilde{A}_{K+1}}\}$ of the following form:

$$\mu_{\tilde{A}_1}^{\tilde{A}_1}(x) = \mu_{\text{low}}(x) = (\hat{\mu}_{11}^R(x), \hat{\mu}_{21}^R(x), \hat{\mu}_{31}^R(x), \hat{\mu}_{41}^R(x)) \text{ if } a \leq x \leq b,$$

$$\text{where } \hat{\mu}_{11}^R(x) = \begin{cases} \frac{a_1^R-x}{a_1^R-a} & a \leq x \leq a_1^R \\ 0 & a_1^R \leq x \leq b \end{cases}; \hat{\mu}_{21}^R(x) = \begin{cases} \frac{a_2^R-x}{a_2^R-a} & a \leq x \leq a_2^R \\ 0 & a_2^R \leq x \leq b \end{cases}, \hat{\mu}_{31}^R(x) = \begin{cases} \frac{a_3^R-x}{a_3^R-a} & a \leq x \leq a_3^R \\ 0 & a_3^R \leq x \leq b \end{cases}; \hat{\mu}_{41}^R(x) = \frac{b-x}{b-a} a \leq x \leq b.$$

$$\mu_{\tilde{A}_K}^{\tilde{A}_K}(x) = \mu_{\text{around}(a+(K-1)\beta)}(x) = \begin{cases} (\hat{\mu}_{1K}^L(x), \hat{\mu}_{2K}^L(x), \hat{\mu}_{3K}^L(x), \hat{\mu}_{4K}^L(x)) & \text{if } a \leq x \leq a + (K - 1)\beta \\ (\hat{\mu}_{1K}^R(x), \hat{\mu}_{2K}^R(x), \hat{\mu}_{3K}^R(x), \hat{\mu}_{4K}^R(x)) & \text{if } a + (K - 1)\beta \leq x \leq b \end{cases}$$

$$\text{where } \hat{\mu}_{1K}^L(x) = \begin{cases} \frac{x-(a+(K-1)\beta)_1^L}{(a+(K-1)\beta)-(a+(K-1)\beta)_1^L} & (a + (K - 1)\beta)_1^L \leq x \leq (a + (K - 1)\beta) \\ 0 & a \leq x \leq (a + (K - 1)\beta)_1^L \end{cases},$$

and similarly, $\hat{\mu}_{2K}^L(x), \hat{\mu}_{3K}^L(x), \hat{\mu}_{4K}^L(x)$ can also be defined.

$$\hat{\mu}_{1K}^R(x) = \begin{cases} \frac{(a+(K-1)\beta)_1^R - x}{(a+(K-1)\beta)_1^R - (a+(K-1)\beta)} & (a+(K-1)\beta) \leq x \leq (a+(K-1)\beta)_1^R, \\ 0 & (a+(K-1)\beta)_1^R \leq x \leq b \end{cases}$$

and similarly, $\hat{\mu}_{2K}^R(x)$, $\hat{\mu}_{3K}^R(x)$, $\hat{\mu}_{4K}^R(x)$ are defined.

$$\mu_{\tilde{A}_{K+1}}^{\tilde{L}}(x) = \mu_{\text{high}}(x) = (\hat{\mu}_{1(K+1)}^L(x), \hat{\mu}_{2(K+1)}^L(x), \hat{\mu}_{3(K+1)}^L(x), \hat{\mu}_{4(K+1)}^L(x))$$

if $a \leq x \leq b$

$$\text{where } \hat{\mu}_{1(K+1)}^L(x) = \begin{cases} \frac{x-b_1^L}{b-b_1^L} & b_1^L \leq x \leq b \\ 0 & a \leq x \leq b_1^L \end{cases} \text{ and others can be defined similarly.}$$

3 T2TSFIS

Takagi and Sugeno fuzzy inference system which was introduced by Takagi and Sugeno [18] is a systemic approach which generates fuzzy rules from a given input-output data set and provides outputs as crisp numbers. In this study, a novel approach of T2TSFIS is introduced which is as follows:

Step 1: Selection of input and output parameters

Let $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ be n number of input parameters defined in the universe of discourse X_1, X_2, \dots, X_n , respectively, where each \tilde{A}_j is a type-2 fuzzy number of the form $\tilde{A}_j = (\tilde{A}_j^1, \tilde{A}_j^2, \tilde{A}_j^3, \tilde{A}_j^4)$, for $j = 1, 2, \dots, n$. Also, let $z = f(x_1, x_2, \dots, x_n)$ be the desired output in the form of a linear function, where $x_j \in X_j (j = 1, 2, \dots, n)$.

Step 2: Formation of rule base

A rule of a T2TSFIS is of the form:

R_k : If x_1 is \tilde{A}_{k1} and x_2 is \tilde{A}_{k2} and ... and x_n is \tilde{A}_{kn} then $z_k = p_{k1}x_1 + p_{k2}x_2 + \dots + p_{kn}x_n + q_k$, where p_{kj} and q_k are real numbers for $k = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$.

Step 3: Assessment of firing strength of each rule

The firing strength of a rule signifies the degree to which the input matches the antecedent part of the rule. The firing strength of the k th ($k = 1, 2, \dots, m$) rule for the input vector $y = (y_1, y_2, \dots, y_n)$ is represented by the four tuple vector $(l_k^1, l_k^2, l_k^3, l_k^4)$, where $l_k^r = I(\mu_{\tilde{A}_{k1}}^r(y_1), \mu_{\tilde{A}_{k2}}^r(y_2), \dots, \mu_{\tilde{A}_{kn}}^r(y_n))$, $r = 1, 2, 3, 4$ and I represents a suitable t -norm.

Step 4: Evaluation of crisp output of each rule

The output of k th ($k = 1, 2, \dots, m$) rule, z_k , is evaluated as:

$$z_k(y) = \sum_{s=1}^n p_{ks} y_s + q_k$$

Step 5: Evaluation of final output

The final output of the proposed inference system is evaluated as:

$$\begin{aligned} z_{\text{T2TSFIS}} = & \lambda_1 \left(\frac{\sum_{k=1}^m l_k^1 z_k}{\sum_{k=1}^m l_k^1} \right) + \lambda_2 \left(\frac{\sum_{k=1}^m l_k^2 z_k}{\sum_{k=1}^m l_k^2} \right) + \lambda_3 \left(\frac{\sum_{k=1}^m l_k^3 z_k}{\sum_{k=1}^m l_k^3} \right) \\ & + \lambda_4 \left(\frac{\sum_{k=1}^m l_k^4 z_k}{\sum_{k=1}^m l_k^4} \right), \end{aligned}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

4 Multiobjective Programming Problem Under Type-2 Fuzzy Environment

4.1 MOO Model Using Type-2 Fuzzy Rule Base

A MOO problem is considered as follows:

$$\begin{aligned} \text{Max } f(x) = & (f_1(x), f_2(x), \dots, f_t(x)); \\ \text{subject to } & \{R_1(x), R_2(x), \dots, R_m(x) | x \in Y \subset \mathbb{R}^n\} \end{aligned}$$

where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is the i th objective, x_j ($j = 1, 2, \dots, n$) is a linguistic variable and $Y \subset \mathbb{R}^n$ is the set of constraints on X_j ($j = 1, 2, \dots, n$) and the k th rule, R_k is presented as:

$R_k(x)$: If x_1 is \tilde{A}_{k1} and x_2 is \tilde{A}_{k2} and ... and x_n is \tilde{A}_{kn} then $f_i(x) = z_{ik} = \sum_{s=1}^n p_{kis} x_s + q_{ki}$, ($i = 1, 2, \dots, t$).

4.2 Construction of Equivalent Deterministic Type-2 MOO Model

The proposed model can be converted into equivalent deterministic multiobjective optimization model as follows:

Step 1: Normalization of the values of linguistic variables

The value of the n th linguistic term, $\tilde{A}_{kn} = (\tilde{A}_{kn}^1, \tilde{A}_{kn}^2, \tilde{A}_{kn}^3, \tilde{A}_{kn}^4)$ is represented by:

$$\tilde{A}_{kn}^r = (a_{rkn}^L, a_{kn}, a_{rkn}^R), \quad r = 1, 2, 3, 4.$$

Therefore the normalized value of \tilde{A}_{kn} is represented by $\tilde{N}_{kn} = (\tilde{N}_{kn}^1, \tilde{N}_{kn}^2, \tilde{N}_{kn}^3, \tilde{N}_{kn}^4)$ and is evaluated as follows:

$$\tilde{N}_{kn}^r = \left(\frac{a_{rkn}^L}{a_{4kn}^R}, \frac{a_{kn}}{a_{4kn}^R}, \frac{a_{rkn}^R}{a_{4kn}^R} \right), \quad r = 1, 2, 3, 4.$$

Therefore, the k th rule of the proposed model can be converted in the following form:

$R_k(x)$: If x_1 is \tilde{N}_{k1} and x_2 is \tilde{N}_{k2} and ... and x_n is \tilde{N}_{kn} then $f_1(x)$ is z_{1k} , $f_2(x)$ is z_{2k} , ..., $f_t(x)$ is z_{tk} .

Step 2: Formation of equivalent type-2 fuzzy rule base

The compositional equivalent type-2 fuzzy rule base contains $m.t$ number of rules which are of the following form:

$R_{1k}(x)$: If x_1 is \tilde{N}_{k1} and x_2 is \tilde{N}_{k2} and ... and x_n is \tilde{N}_{kn} then $f_1(x)$ is z_{1k} ,
 $R_{2k}(x)$: If x_1 is \tilde{N}_{k1} and x_2 is \tilde{N}_{k2} and ... and x_n is \tilde{N}_{kn} then $f_2(x)$ is z_{2k} ,

...

$R_{tk}(x)$: If x_1 is \tilde{N}_{k1} and x_2 is \tilde{N}_{k2} and ... and x_n is \tilde{N}_{kn} then $f_t(x)$ is z_{tk} ,
 $k = 1, 2, \dots, m$

Step 3: Aggregation of output of each rule

The i th objective function f_i at $y \in Y \subset \mathbb{R}^n$ is obtained by:

$$f_i(y) = \lambda_1 \left(\frac{l_1^1 z_{i1} + l_2^1 z_{i2} + \dots + l_m^1 z_{im}}{l_1^1 + l_2^1 + \dots + l_m^1} \right) + \lambda_2 \left(\frac{l_1^2 z_{i1} + l_2^2 z_{i2} + \dots + l_m^2 z_{im}}{l_1^2 + l_2^2 + \dots + l_m^2} \right) + \lambda_3 \left(\frac{l_1^3 z_{i1} + l_2^3 z_{i2} + \dots + l_m^3 z_{im}}{l_1^3 + l_2^3 + \dots + l_m^3} \right) + \lambda_4 \left(\frac{l_1^4 z_{i1} + l_2^4 z_{i2} + \dots + l_m^4 z_{im}}{l_1^4 + l_2^4 + \dots + l_m^4} \right),$$

where $l_k^r = I(\mu_{\tilde{a}_{k1}^r}^r(y_1), \mu_{\tilde{a}_{k2}^r}^r(y_2), \dots, \mu_{\tilde{a}_{kn}^r}^r(y_n))$, $r = 1, 2, 3, 4$, I is a suitable T -norm and $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

Step 4: Construction of equivalent crisp multiobjective programming problem

Finally, the optimization problem transforms to an equivalent crisp multiobjective programming problem as follows:

$$\begin{aligned} \text{Max } f(x) &= (f_1(x), f_2(x), \dots, f_t(x)); \\ \text{subject to } y &\in Y. \end{aligned}$$

Here, in most of the cases, f_i 's are non-linear functions in nature.

Each objective is now solved independently under the set of constraints to find the upper bound and lower bound of each objective. Let f_i^u ($i = 1, 2, \dots, t$) and f_i^l ($i = 1, 2, \dots, t$) be the upper bound and lower bound of the i -th objective, respectively. Based on the upper and lower value of the objectives the membership functions are constructed as:

$$\mu_{f_i} = \frac{f_i - f_i^l}{f_i^u - f_i^l}, \quad (i = 1, 2, \dots, t).$$

Now, after converting the above membership functions into flexible fuzzy membership goals, a weighted fuzzy goal programming (FGP) model is developed for finding the compromise solution of all the objectives.

Thus, the weighted FGP model is formulated as:

$$\begin{aligned} \text{Minimize } D &= \sum_{i=1}^t w_i d_i^- \\ \text{subject to } &\frac{f_i - f_i^l}{f_i^u - f_i^l} + d_i^- - d_i^+ = 1, \quad (i = 1, 2, \dots, t). \\ &\text{and } y \in Y. \end{aligned}$$

where $w_i = \frac{1}{f_i^u - f_i^l} \geq 0$, signifies the weight of the goals.

The developed multiobjective model in crisp environment is then solved to find the most satisfactory result in decision making environment.

5 Numerical Example

In this section, a modified version of the numerical example considered earlier by Chakraborty et al. [2] is illustrated to establish the application potentiality of the proposed model which is presented as follows:

$$\begin{aligned} &\text{Max } (f_1(x), f_2(x)) \\ &\text{subject to } \{0 \leq x_1 \leq 10, 10 \leq x_2 \leq 70, 6x_1 + x_2 \leq 90\}, \end{aligned}$$

where

$R_1(x)$: If x_1 is high and x_2 is very high then $f_1(x) = -x_1 + x_2$ and $f_2(x) = x_1 + x_2/2$.

$R_2(x)$: If x_1 is low and x_2 is high then $f_1(x) = x_1 + x_2$ and $f_2(x) = -x_1 + x_2$.

5.1 Construction of Equivalent Deterministic MOO Model

The proposed model can be converted into equivalent deterministic MOO model as follows:

Step 1: Normalization of the values of linguistic variables

In this step, the linguistic scales along with their corresponding T2FNs are demonstrated and are normalized using the process as described in Sect. 4.2. The normalized scales for different values of x_1 and x_2 are presented in Tables 1 and 2, respectively.

Step 2: Formation of equivalent type-2 fuzzy rule base

The firing strengths of the rules R_1 and R_2 are computed using product t -norm as the four tuple vector $l_1 = (l_1^1, l_1^2, l_1^3, l_1^4)$ and $l_2 = (l_2^1, l_2^2, l_2^3, l_2^4)$, respectively.

Now, the compositional equivalent type-2 fuzzy rules for the proposed example are of the following form:

$R_{11}(x)$: If x_1 is high and x_2 is very high then $f_1(x) = -x_1 + x_2$.

$R_{21}(x)$: If x_1 is high and x_2 is very high then $f_2(x) = x_1 + x_2/2$.

$R_{12}(x)$: If x_1 is low and x_2 is high then $f_1(x) = x_1 + x_2$.

$R_{22}(x)$: If x_1 is low and x_2 is high then $f_2(x) = -x_1 + x_2$.

The respective output of rules R_{11} , R_{12} , R_{21} and R_{22} are $z_{11} = -y_1 + y_2$, $z_{12} = y_1 + y_2$, $z_{21} = y_1 + y_2/2$ and $z_{22} = -y_1 + y_2$.

Step 3: Aggregation of output of each rule

Now, the objective function $f_1(y_1, y_2)$ is computed as:

$$f_1(y_1, y_2) = \lambda_1 f_1^1(y_1, y_2) + \lambda_2 f_1^2(y_1, y_2) + \lambda_3 f_1^3(y_1, y_2) + \lambda_4 f_1^4(y_1, y_2),$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$, $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ and

Table 1 Linguistic value for x_1

Linguistic term	T2TFN	Normalized T2TFN
Very low	((0,0.8), (0,0.8.5), (0,0,9.5), (0,0,10))	((0,0,0.8), (0,0,0.85), (0,0,0.95), (0,0,1))
Low	((2,2.5,8), (1.5,2.5,8.5), (0.5,2.5,9.5), (0.2,5,10))	((0.2,0.25,0.8), (0.15,0.25,0.85), (0.05,0.25,0.95), (0,0.25,1))
Medium	((2,5,8), (1.5,5,8.5), (0.5,5,9.5), (0.5,10))	((0.2,0.5,0.8), (0.15,0.5,0.85), (0.05,0.5,0.95), (0,0.5,1))
High	((2,7.5,8), (1.5,7.5,8.5), (0.5,7.5,9.5), (0,7.5,10))	((0.2,0.75,0.8), (0.15,0.75,0.85), (0.05,0.75,0.95), (0,0.75,1))
Very high	((2,10,10), (1.5,10,10), (0.5,10,10), (0,10,10))	((0.2,1,1), (0.15,1,1), (0.05,1,1), (0,1,1))

Table 2 Linguistic value for x_2

Linguistic term	T2TFN	Normalized T2TFN
Very low	$((10,10,62), (10,10,64), (10,10,68), (10,10,70))$	$((0.143,0.143,0.886), (0.143,0.143,0.914), (0.143,0.143,0.971), (0.143,0.143,1))$
Low	$((18,20,62), (16,20,64), (12,20,68), (10,20,70))$	$((0.257,0.286,0.886), (0.229,0.286,0.914), (0.171,0.286,0.971), (0.143,0.286,1))$
Medium low	$((18,30,62), (16,30,64), (12,30,68), (10,30,70))$	$((0.257,0.429,0.886), (0.229,0.429,0.914), (0.171,0.429,0.971), (0.143,0.429,1))$
Medium	$((18,40,62), (16,40,64), (12,40,68), (10,40,70))$	$((0.257,0.571,0.886), (0.229,0.571,0.914), (0.171,0.571,0.971), (0.143,0.571,1))$
Medium high	$((18,50,62), (16,50,64), (12,50,68), (10,50,70))$	$((0.257,0.714,0.886), (0.229,0.714,0.914), (0.171,0.714,0.971), (0.143,0.714,1))$
High	$((18,60,62), (16,60,64), (12,60,68), (10,60,70))$	$((0.257,0.857,0.886), (0.229,0.857,0.914), (0.171,0.857,0.971), (0.143,0.857,1))$
Very high	$((18,70,70), (16,70,70), (12,70,70), (10,70,70))$	$((0.257,1,1), (0.229,1,1), (0.171,1,1), (0.143,1,1))$

$$f_1^1(y_1, y_2) = \frac{l_1^1 z_{11} + l_2^1 z_{12}}{l_1^1 + l_2^1}$$

$$= \begin{cases} \frac{\frac{y_1-0.2}{0.55} \cdot \frac{y_2-0.257}{0.743} (y_2-y_1) + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6} (y_1+y_2)}{\frac{y_1-0.2}{0.55} \cdot \frac{y_2-0.257}{0.743} + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6}} & \text{if } 0.257 \leq y_1, y_2 \leq 0.75 \\ \frac{\frac{0.8-y_1}{0.05} \cdot \frac{y_2-0.257}{0.743} (y_2-y_1) + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6} (y_1+y_2)}{\frac{0.8-y_1}{0.05} \cdot \frac{y_2-0.257}{0.743} + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6}} & \text{if } 0.75 \leq y_1, y_2 \leq 0.8 \\ 0 & \text{elsewhere} \end{cases}$$

Similarly, f_1^2 , f_1^3 and f_1^4 can be found.

Next, the objective function $f_2(y_1, y_2)$ is computed as:

$$f_2(y_1, y_2) = \lambda_1 f_1^1(y_1, y_2) + \lambda_2 f_1^2(y_1, y_2) + \lambda_3 f_1^3(y_1, y_2) + \lambda_4 f_1^4(y_1, y_2),$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$, $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

$$f_2^1(y_1, y_2) = \frac{l_1^1 z_{21} + l_2^1 z_{22}}{l_1^1 + l_2^1}$$

$$= \begin{cases} \frac{\frac{y_1-0.2}{0.55} \cdot \frac{y_2-0.257}{0.743} (y_1+y_2/2) + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6} (y_2-y_1)}{\frac{y_1-0.2}{0.55} \cdot \frac{y_2-0.257}{0.743} + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6}} & \text{if } 0.257 \leq y_1, y_2 \leq 0.75 \\ \frac{\frac{0.8-y_1}{0.05} \cdot \frac{y_2-0.257}{0.743} (y_1+y_2/2) + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6} (y_2-y_1)}{\frac{0.8-y_1}{0.05} \cdot \frac{y_2-0.257}{0.743} + \frac{0.8-y_1}{0.55} \cdot \frac{y_2-0.257}{0.6}} & \text{if } 0.75 \leq y_1, y_2 \leq 0.8 \\ 0 & \text{elsewhere} \end{cases}$$

Table 3 Upper bound and lower bound of f_1 and f_2

Objective functions	Upper bound	Lower bound
f_1	0.968	-0.345
f_2	1.045	0.044

Table 4 Compromise solution of the objectives

Solution point		Objective 1	Objective 2
y_1	y_2	f_1	f_2
0.55	0.75	0.7145	0.6919

In a similar manner f_2^2 , f_2^3 and f_2^4 can be derived.

Step 4: Construction of equivalent crisp multiobjective programming problem

Finally, the optimization problem transforms to an equivalent crisp multiobjective programming problem as follows:

$$\begin{aligned} \text{Max } f(y) &= (f_1(y), f_2(y)) \\ \text{subject to } 0.67y_1 + 0.78y_2 &\leq 1 \\ y_1 \in [0, 1], y_2 \in [0.143, 1]. \end{aligned}$$

Each objective is now solved individually under the system constraints defined above to find the upper bound and lower bound of the objectives.

The upper bound and lower bound of f_1 and f_2 are shown in Table 3.

On the basis of upper and lower bound of the objectives, the membership functions μ_{f_1} and μ_{f_2} of the corresponding objectives are formed.

Now, the compromise solution of the objectives are evaluated by solving the weighted FGP model after converting the above defined membership functions into membership goals. The *software LINGO* (ver. 19) is used to find solution of the problem. The compromise solution is shown in Table 4.

6 Comparative Analysis

In this section the solution obtained by proposed methodology are compared to the solution obtained by the methodology described in the article Chakraborty et al. [2]. This comparison is shown in Table 5.

The above table clearly suggests the superiority of the proposed methodology over the existing methodology in terms of attaining the objective values.

Table 5 Comparison of solutions

Method	Solution point	Objective value
Proposed methodology	$y_1 = 0.55$ $y_2 = 0.75$	$f_1 = 0.7145$ $f_2 = 0.6919$
Method described by Chakraborty et al. [2]		
GA	$y_1 = 0.625$ $y_2 = 0.746$	$f_1 = 0.6441$ $f_2 = 0.6313$
PSO	$y_1 = 0.619$ $y_2 = 0.75$	$f_1 = 0.6564$ $f_2 = 0.6277$

7 Conclusions

This article proposes a new methodology for solving type-2 fuzzy multiobjective programming problems. Instead of considering the linguistic variables as ordinary fuzzy numbers, T2TFNs are considered by keeping in mind that T2FS captures uncertainties more efficiently than ordinary fuzzy sets. Although, some amounts of computational complexities may increase due to the inclusion of type-2 fuzzy notions, but it is witnessed that, capturing uncertainties more powerfully, a highly acceptable and reasonable decisions can be assessed from the proposed methodology. Thus, the methodology presented in this article can be applied to different real life applications for obtaining most satisfactory solution in the type-2 fuzzy decision making environments.

Acknowledgements The authors remain grateful to the anonymous reviewers for their comments and suggestions. The authors are thankful to the Science and Engineering Research Board, Department of Science and Technology, Government of India for providing financial assistance through the project under Teachers Associateship for Research Excellence (TARE) scheme vide Reference No. TAR/2019/000272.

References

1. Khorram E, Ezzati R, Valizadeh Z (2014) Linear fractional multi-objective optimization problems subject to fuzzy relational equations with a continuous Archimedean triangular norm. *Inf Sci* 267:225–239
2. Chakraborty D, Guha D, Dutta B (2016) Multi-objective optimization problem under fuzzy rule constraints using particle swarm optimization. *Soft Comput* 20:2245–2259
3. Afzal A, Ramis MK (2020) Multi-objective optimization of thermal performance in battery system using genetic and particle swarm algorithm combined with fuzzy logics. *J Energy Storage* 32:101815
4. Hashemi SE (2021) A fuzzy multi-objective optimization model for a sustainable reverse logistics network design of municipal waste-collecting considering the reduction of emissions. *J Clean Prod* 318:128577
5. Brindha S, Amali SMJ (2021) A robust and adaptive fuzzy logic based differential evolution algorithm using population diversity tuning for multi-objective optimization. *Eng Appl Artif Intell* 102:104240

6. Guo X (2021) Multi-objective task scheduling optimization in cloud computing based on fuzzy self-defense algorithm. *Alex Eng J* 60:5603–5609
7. Pazouki M, Rezaie K, Bozorgi-Amiri A (2021) A fuzzy robust multi-objective optimization model for building energy retrofit considering utility function: a university building case study. *Energy Build* 241:110933
8. Zadeh LA (1965) Fuzzy sets. *Inf Control* 8(3):338–353
9. Dergah A, Golea N, Essounbouli N (2016) An interval fuzzy optimization-based technique to optimal generation scheduling with load uncertainty. *IFAC-PapersOnLine* 49(12):1122–1127
10. Liang Q, Mendel JM (2000) Interval type-2 fuzzy logic systems: theory and design. *IEEE Trans Fuzzy Syst* 8:535–550
11. Yucesan M, Başhan V, Demirel H, Gul H (2022) An interval type-2 fuzzy enhanced best–worst method for the evaluation of ship diesel generator failures. *Eng Fail Anal* 138:106428
12. Gomes DCS, Serra GLO (2022) Interval type-2 fuzzy computational model for real time Kalman filtering and forecasting of the dynamic spreading behavior of novel coronavirus 2019. *ISA Trans* 124:57–68
13. Mazandarani M, Xiu L (2022) Interval type-2 fractional fuzzy inference systems: towards an evolution in fuzzy inference systems. *Expert Syst Appl* 189:115947
14. Lv Z, Jin H, Yuan P (2009) The theory of triangle type-2 fuzzy sets. In: Proceedings of the 2009 IEEE international conference on computer and information technology. IEEE, Piscataway, pp 57–62
15. Biswas A, De AK (2018) A unified method of defuzzification for type-2 fuzzy numbers with its application to multiobjective decision making. *Granular Comput* 3:301–318
16. Ghorbani A, Zamanifar K (2022) Type-2 fuzzy ontology-based semantic knowledge for indoor air quality assessment. *Appl Soft Comput* 121:108658
17. De AK, Chakraborty D, Biswas A. Literature review on type-2 fuzzy set theory, soft computing (accepted)
18. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modelling and control. *IEEE Trans Syst Man Cybern* 15(1):116–132

Person Detection Using YOLOv3



Bhawana Tyagi, Swati Nigam, and Rajiv Singh

Abstract In today's world, person detection in video surveillance is very important. It has many applications like crowd counting, single and multiple object tracking, crowd behavior analysis, anomaly detection, etc. There are different models to detect a person in an image and video. But, the majority of the models focused on many object classes which sometimes lead to poor performance for detecting specific objects. In this paper, a single class of object is considered, i.e., person. Here, we have used transfer learning for generating the person detection system by using YOLOv3. We have generated the content specific customized dataset, and annotated the dataset manually by using Label Tool. The result shows that the proposed model detects and classifies the person with higher accuracy.

Keywords Object detection · Deep learning · YOLOv3 · Convolution neural network

1 Introduction

Object detection is an area of computer vision, which includes applications like driverless cars [1, 2], video surveillance [3], crowd behavior analysis [4] and traffic monitoring systems [5, 6]. Previously, saturated accuracy was the major problem with the computer vision. After the emergence of deep learning methods [7], there is drastic improvement in terms of accuracy and results. The convolution neural network (CNN) [8] and You Look Only Once (YOLO) are common methods used for object detection. Challenges associated with object detection are multiple objects in the single frame, occlusion, size of an object, position of an object, angle of the camera from which image is taken, etc. In general, we detect a person in the images from video surveillance which can further be used for crowd behavior analysis in order to provide the safety to individuals and it may help to reduce the occurrence of crime [9]. Human detection system is very useful in terms of security. It can

B. Tyagi · S. Nigam · R. Singh (✉)

Department of Computer Science, Banasthali Vidyapith, Banasthali, Rajasthan, India

e-mail: jkrajivsingh@gmail.com

help us to count the number of persons in an image which can identify how many persons can be easily accommodated in the sports arena, live concerts, political rallies, etc. Traditionally, there was a security officer for surveillance who closely monitors the closed-circuit television (CCTV), but this approach was vulnerable to many human errors like negligence, boredom, fatigue, etc. The conventional methods also suffered from poor detection when the number of people is more in the frame due to occlusion. To tackle these issues, here we generate human detection system by using YOLOv3 [10] network. There is many customized YOLO-based object detection system available. Lemon YOLO model is proposed in [11] to detect lemon in real time natural environment by using YOLOv3. To detect and classify leukocytes in leukemia, YOLO model is used in [12]. To detect small objects, [13] used YOLO. Sonavane et al. [14] detects the dental cavity by using YOLO. The YOLO-based pedestrian detection is performed by Lan et al. [15], Hsu et al. [16], Shao et al. [17]. Saurav et al. [18] used YOLOv3 to detect whether a person wearing a mask or not, to avoid the outbreak of COVID-19. Siddhi et al. [19] proposed model based on YOLOv3 tiny and YOLOv4 tiny to detect the diseases in the rice plant at early stage. YOLOv4 was used by Saurav et al. [20] to detect fire at construction sites in real time.

Object detection is two-step process. In the first step, image classification is done in order to recognize the class of object present in an image and name it. The second step is object localization, which recognizes the location of every object in the image and generates the bounding box around those objects. Here, we have used transfer learning [21] to train the model. Transfer learning is used to provide learning parameters of large dataset to alike smaller dataset. The model is trained for particular task, and it is used as an initial point for model on the second task. The block diagram is depicted in Fig. 1. We have created our own context specific and custom dataset, then annotated it. Then transfer learning is applied by considering YOLOv3 as a base model. In this paper, we have used transfer learning for generating the person detection system by using YOLOv3 [10]. The experimental results demonstrated on customized dataset gathered from open image dataset (OID v4) [22] and person-COCO dataset [23].

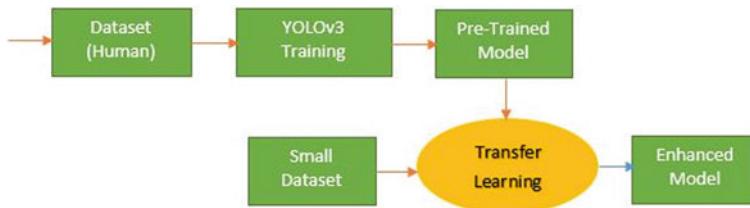


Fig. 1 Transfer learning

2 Related Works

The object detection methodologies can be categorized into two parts: two stage detection and one stage detection as depicted in Fig. 2. In two stage detection, the region of interest can be found in first step and then, classification and detection are done within the bounding box in the second step. In one stage detection, the prediction of bounding box for the objects are identified directly. The one stage detection process is fast as comparison to two stage detection process, but its accuracy is low.

Two Stage Detection: In 2014, regions with CNN (R-CNN) [24] for object detection was proposed. It has three modules: region proposal, feature extractor and classifier. For feature extractor AlexNet deep CNN [25] was used and for classification linear support vector machine (SVM) was used. This method was poor in terms of computations. To deal with this, Girshick et al. proposed Fast R-CNN [26] and Faster R-CNN [27]. As size of an input image is fixed, it leads to poor detection efficiency. To overcome with this problem, spatial pyramid pooling network (SPPNet) [28] was proposed. In this network, they have added spatial pyramid pooling layer [29, 30] on last convolutional layer. Mask RCNN [31] is the extension of Faster-RCNN. Authors added a branch in parallel to predict the object mask with already existing branch that recognizes the bounding box.

One Stage Detection: For real time processing, the one stage detection is most appropriate. It predicts the bounding box for objects in one step only. There are many one stage detection methods exist which include YOLO [32], YOLO9000 [33], YOLOv3 [10] and YOLOv4 [34]. Redmon et al. [32] proposed YOLO model, which includes single neural network to predict class and bounding box of an object directly from the given image. It was very fast but has low accuracy. YOLO9000 [33] predicts over 9000 classes of objects and it overcomes the problem associated with YOLO. Further, YOLOv3 was proposed. It used logistic regression to generate bounding box, logistic classifier to predict the class and used Darknet 53 as backbone network. YOLOv4 added features to the previous model, i.e., bag of freebies, bag of specials, Cross-Part-Partial Connection (CSP) with Darknet53 and spatial pyramid pooling, single shot multibox detector (SSD) was proposed by Liu et al. [35]. It was fast and accurate than YOLO. Further to improve accuracy and to detect the small objects, the de-convolutional SSD (DSSD) [36] was proposed by combining the Residual-101 [37] with SSD [35]. RetinaNet was proposed by Lin et al. [38]

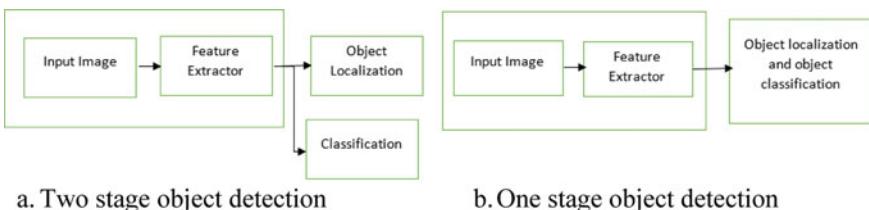


Fig. 2 Object detection strategies

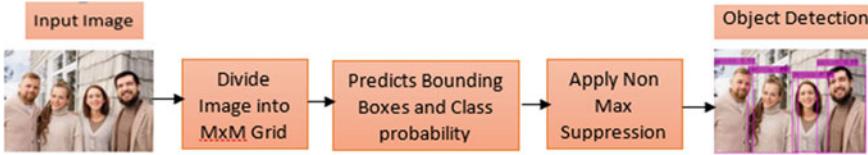


Fig. 3 Human detection based on YOLOv3

and the focal loss function was also proposed to measure class imbalance. To extract rich features, they have used feature pyramid network (FPN) [39] on ResNet [37] architecture.

3 Methodology

3.1 Overview

Object detection is a task to detect an object and locate it into images and videos. The process of human detection method involves creation of bounding box over every human present in the frame and assigns the label ‘person’. We are using YOLOv3 to detect the human in the image and the motivation is its speed, and capability to detect objects in video also. The steps of human detection based on YOLOv3 is shown in Fig. 3. Initially the input image is divided into N grids. There is an equal dimension for each grid, i.e., $M \times M$. Then these grids are responsible to detect and localize an object. Based on these grids the bounding boxes along with label and probability of an object is predicted. This process is fast as both object detection and recognition is done at single step, but it generates redundant bounding boxes. In order to deal with this problem, non-maximum suppression is performed which eliminates low probability bounding boxes. The bounding box has 6 components: $(P_C, B_{XC}, B_{YC}, B_W, B_h, c)$, where P_C is the probability score of an object, B_{XC} and B_{YC} are the coordinates of center point of the bounding box. B_W and B_h are the width and height of the bounding box and c represents the class of an object.

3.2 The Proposed System

The description of a proposed system is depicted in Fig. 4. It consists of three parts: preparation of context specific and custom dataset, network training by using transfer learning and testing.

Dataset Collection: The first step is data collection. Open image dataset (OIDv4) is used to create the customized dataset. Around 1500 images from person class is downloaded via OIDv4 toolkit. Then, the dataset is divided into three parts: train, test

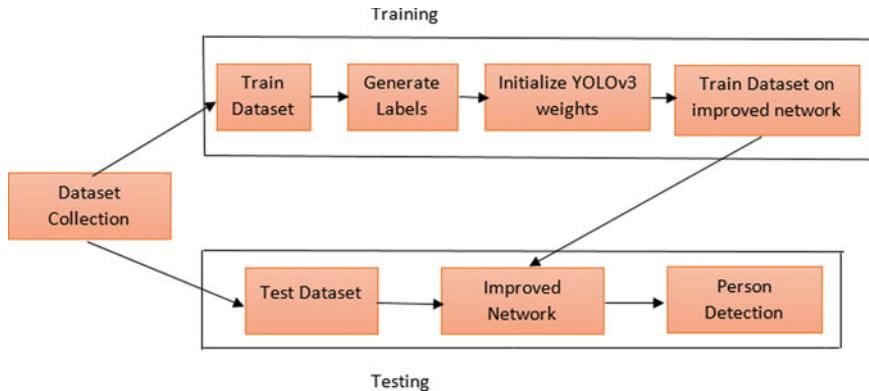


Fig. 4 The proposed system

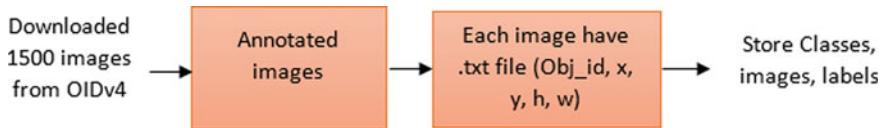


Fig. 5 Labeling and storing bounding box dimensions

and validation datasets. Train dataset contains 900 images, test and validation datasets contain 300 images. The image from dataset is annotated using LabelTool. Figure 5 depicts the labeling and storing the bounding box dimensions. Each object has its corresponding bounding box. In this tool, the images are stored in a separate folder. Thereafter, bounding box for each object in the image has been created and labeled with its class name. After annotating the images, each image has its corresponding.txt file that contains the information of the bounding box. Finally, the annotations are stored in YOLOv3 format.

Training: We have created.cfg,.data and.names files using YOLOv3. Then, dataset and the annotations have been copied into corresponding directory. The model is trained by the dataset having single class, i.e., person. The source code to train the YOLO model is available [40]. The open source framework Darknet [41] is used to train the neural networks. Initially, create.cfg file to detect human by copying the content from YOLOv3.cfg to YOLO-human.cfg file. In the.cfg file, batch size = 64, subdivisions = 8, width and height of an image is 416, channels = 3. To manage the weights, momentum is 0.9, and decay is 0.0005 for controlling the penalty term. Here, we have taken batches = 2000, as maximum batches are (classes * 2000) and in our case there is only one class, i.e., human. The number of filters are 18 as formula is “filters = (classes + 5) * 3”. Next, step is to create the human.data file. Here classes = 1, train variable contains the path to train file, valid variable contains the path to validation file, name variable contains the names of objects and to store the weights the backup variable contains the path to backup file. The.name file consists of

names of different objects. In the next step, all.jpg images and corresponding.txt files are copied to object directory of darknet. Then, in data directory of darknet, create train.txt file which contains filenames of all the images. In the darknet directory, pretrained weights are downloaded and finally start the training. After completing the 400 iterations, the weights will get stored into backup file and for human detection, best weights will be used.

Testing: To test the model, input arguments image,.cfg file, trained weights,.data file and .name file are required. The model will obtain the width and height of an input frame initially followed by label and bounding box colors. By using weight and .cfg file, network will be created by cv2.read function. The output will be obtained using multiple output layers. The bounding boxes will be generated over the humans using function draw bounding box. The confidence score indicates the probability of bounding box having a person. The bounding box having less than 0.5 confidence score will be dropped. Non max suppression will remove all the extra bounding boxes.

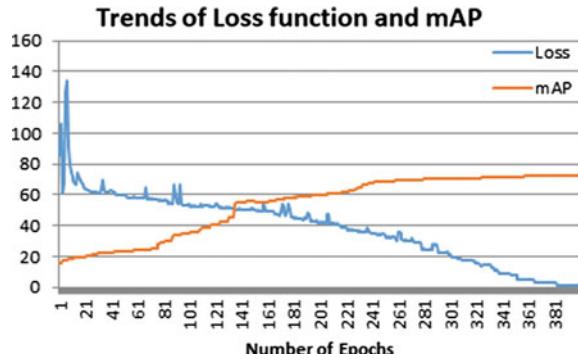
3.3 Person Detection

Once the model is trained, the detection can be performed on our test dataset, based on the final weights obtained after training the network. The intersection over union (IOU) is used to find overlap ratio between the detected bounding box and ground truth bounding box. The IOU is represented in Eq. 1.

$$\text{IOU} = \frac{\text{area}(\text{boundingbox (groundtruth)} \cap \text{boundingbox (predicted)})}{\text{area}(\text{boundingbox (groundtruth)} \cup \text{boundingbox (predicted)})} \quad (1)$$

The best ratio is 1. Although, we consider the detection, if the ratio is 0.5 or more and eliminate all the detections whose ratio is less than 0.5. Once the process is complete then all the detection are visible on the image. Figure 6 depicts the trends of loss function and mAP while training the model.

Fig. 6 The trends of loss function and mAP while training the model



4 Experiments and Discussions

The proposed person detection system used YOLOv3 to draw the bounding box corresponding to every human into the frame and assigns the label, i.e., person and its confidence score. Figure 7 shows the results. The first and third columns represent the confidence scores of people without transfer learning, and second and fourth columns show the confidence scores after applying transfer learning on the same sample images. The detailed description of Fig. 7 is given in Table 1. It shows that for image 1, the detection accuracy for person 1 is increased by 32.32% and for person 2, it is decreased by 1.4%. For image 2, the detection accuracy is increased by 1%. For image 3, the detection accuracy for person 1 is increased by 64.4%, person 2 by 60.4%, person 3 by 43.43%, person 4 by 5.3% and person 5 by 34.8%. For image 4, detection accuracy for person 1 is increased by 8%, person 2 by 5% and person 3 by 35%. The detection accuracy for person 1, 2, 3 and 4 is increased by 8%, 10%, 7.1% and 3%. For the sample image 6, the detection accuracy is increased by 26.8% for person 1, 31.6% for person 2 and 9% for person 3.

The proposed model is also used to detect the person in the person-COCO dataset and mean average precision (mAP) is used to evaluate the performance of the proposed model. The formula of mAP is described in Eq. 2 where N is number of classes and here, we have considered only one class.

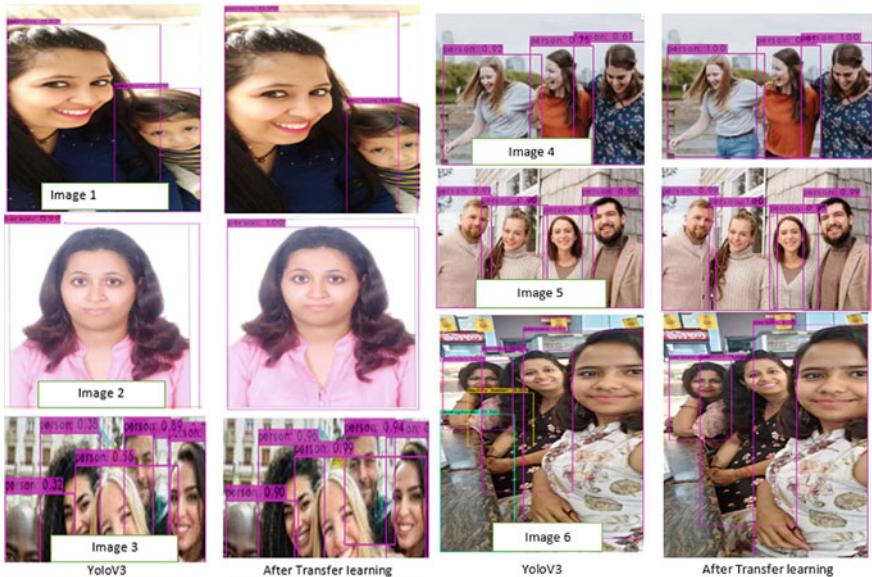


Fig. 7 Comparison between YOLOv3 and after applying Transfer learning on some sample images

Table 1 Quantitative results on sample images

S. No.	YOLOv3	After transfer learning	Observations
1	Person1-67%, Person2-69%	Person1-99%, Person2-68%	Better detection accuracy of proposed model
2	Person1-99%	Person1-100%	Better detection accuracy of proposed model
3	Person1-32%, Person2-38%, Person3-56%, Person4-89%, Person5-56%	Person1-90%, Person2-96%, Person3-99%, Person4-94%, Person5-86%	Better detection accuracy of proposed model
4	Person1-92%, Person2-95%, Person3-65%	Person1-100%, Person2-100%, Person3-100%	Better detection accuracy of proposed model
5	Person1-91%, Person2-90%, Person3-91%, Person4-96%	Person1-99%, Person2-100%, Person3-98%, Person4-99%	Better detection accuracy of proposed model
6	Person1-71%, Person2-67%, Person3-91%	Person1-97%, Person2-98%, Person3-100%	Better detection accuracy of proposed model

Table 2 Comparison of mAP on person-COCO dataset

S. No.	Model	mAP
1	YOLOv2	54.3
2	YOLOv2 tiny	39.7
3	YOLOv3	68.4
4	YOLOv3 tiny [42]	57.2
5	Proposed YOLOv3(customized)	71.3

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (2)$$

The comparative analysis of the model with previous model is depicted in Table 2. The customized model is showing the highest mAP with value 71.3% which is 44.3% better than the lowest value and 4% better than the highest mAP.

5 Conclusions and Future Scope

In this paper, we have used transfer learning for person detection which is based on YOLOv3. We have generated customized dataset using OIDV4. The results showed that the proposed modified YOLOv3 model with transfer learning performs better

than the conventional YOLOv3. This work can further be used for the development of a video surveillance system for safety and security of individuals at public places. This enables us to count the number of persons in an image which can identify how many persons can be easily accommodated in the sports arena, live concerts, political rallies, etc. It can further be used to implement the multiple object tracking.

References

1. Gupta A, Anpalagan A, Guan L, Khwaja AS (2021) Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. *Array* 10:100057
2. Strickland M, Fainekos G, Amor HB (2018) Deep predictive models for collision risk assessment in autonomous driving. In: 2018 IEEE international conference on robotics and automation. IEEE, pp 4685–4692
3. Rezaee K, Rezakhani SM, Khosravi MR, Moghimi MK (2021) A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. *Per Ubiquit Comput*:1–17
4. Sánchez FL, Hupont I, Tabik S, Herrera F (2020) Revisiting crowd behaviour analysis through deep learning: taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Inf Fusion* 64:318–335
5. Kumar C, Punitha R (2020) Performance analysis of object detection algorithm for intelligent traffic surveillance system. In: 2020 second international conference on inventive research in computing applications. IEEE, pp 573–579
6. Mandal V, Mussah AR, Jin P, Adu-Gyamfi Y (2020) Artificial intelligence-enabled traffic monitoring system. *Sustainability* 12(21):9177
7. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
8. Zhao ZQ, Zheng P, Xu ST, Wu X (2019) Object detection with deep learning: A review. *IEEE Trans Neural Networks Learn Syst* 30(11):3212–3232
9. Costin A (2016) Security of cctv and video surveillance systems: threats, vulnerabilities, attacks, and mitigations. In: Proceedings of the 6th international workshop on trustworthy embedded devices, pp 45–54
10. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767).
11. Li G, Huang X, Ai J, Yi Z, Xie W (2021) Lemon-YOLO: An efficient object detection method for lemons in the natural environment. *IET Image Proc* 15(9):1998–2009
12. Abas SM, Abdulazeez AM, Zeebaree DQ (2022) A YOLO and convolutional neural network for the detection and classification of leukocytes in leukemia. *Indonesian J Electr Eng Comput Sci* 25(1):200–213
13. Ajaz A, Salar A, Jamal T, Khan AU (2022) Small object detection using deep learning. arXiv preprint [arXiv:2201.03243](https://arxiv.org/abs/2201.03243)
14. Sonavane A, Kohar R (2022) Dental cavity detection using YOLO. In: Proceedings of data analytics and management. Springer, Singapore, pp 141–152
15. Lan W, Dang J, Wang Y, Wang S (2018) Pedestrian detection based on YOLO network model. In: 2018 IEEE international conference on mechatronics and automation. IEEE, pp 1547–1551
16. Hsu WY, Lin WY (2020) Ratio-and-scale-aware YOLO for pedestrian detection. *IEEE Trans Image Process* 30:934–947
17. Shao Z, Cheng G, Ma J, Wang Z, Wang J, Li D (2021) Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. *IEEE Trans Multimedia* 24:2069–2083
18. Kumar S, Yadav D, Gupta H et al (2022) Towards smart surveillance as an aftereffect of COVID-19 outbreak for recognition of face masked individuals using YOLOv3 algorithm. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-021-11560-1>

19. Jain S, Sahni R, Khargonkar T, Gupta H, Verma OP, Sharma TK, Bhardwaj T, Agarwal S, Kim H (2022) Automatic rice disease detection and assistance framework using deep learning and a Chatbot. *Electronics* 11(14):2110
20. Kumar S, Gupta H, Yadav D, Ansari IA, Verma OP (2022) YOLOv4 algorithm for the real-time detection of fire and personal protective equipment's at construction sites. *Multimedia Tools Appl* 81(16):22163–22183
21. Torrey L, Shavlik J (2010) Transfer learning. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp 242–264
22. Oidv4. <https://github.com/EscVM/OIDv4-ToolKit>. [Online]. Accessed 3 Apr 2022
23. <https://cocodataset.org/#explore>
24. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580–587
25. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances Neural Inf Process Syst* 25
26. Girshick R (2015) Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp 1440–1448
27. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. *Advances Neural Inf Process Syst* 28
28. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
29. Grauman K, Darrell T (2005) The pyramid match kernel: discriminative classification with sets of image features. In: *Tenth IEEE international conference on computer vision (ICCV'05)* Volume 1, vol 2. IEEE, pp 1458–1465
30. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE computer society conference on computer vision and pattern recognition*, vol 2. IEEE, pp 2169–2178
31. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp 2961–2969
32. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 779–788
33. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7263–7271
34. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*
35. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: *European conference on computer vision*, pp 21–37
36. Fu CY, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*
37. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings IEEE conference on computer vision and pattern recognition*, pp 770–778
38. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, pp 2980–2988
39. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2117–2125
40. To train yolo model. <https://github.com/AlexeyAB/darknet>. [Online]. Accessed 3 Apr 2022
41. Darknet framework. <https://pjreddie.com/darknet/yolov1>. [Online]. Accessed 3 April 2022
42. Adarsh P, Rathi P, Kumar M (2020) YOLO v3-Tiny: object detection and recognition using one stage improved model. In: *2020 6th international conference on advanced computing and communication systems (ICACCS)*. IEEE, pp 687–694, 6 Mar 2020

Incident Reporting of Forest Fire with Azure Cognitive Services and Twitter API



Rakesh Kumar, Meenu Gupta, Dhruv Kinger, and Sayanto Roy

Abstract A forest fire has a profound impact on wildlife, atmospheric, economic and ecosystem structure. The ever-increasing average global temperature and climate change have increased the frequency and magnitude of forest fires. The cause of forest fires can be natural as well as human-caused. This research focuses on the detection technique through which a wildfire incident can be reported as early as possible. In this work, an alert about a forest fire/wildfire is created as soon as it takes place and then spread the news to the world. This aims to solve the negligence of the forest fires taking place and the lack of any specific alerting mechanism of the same. In result analysis, the study proves that the system works with good accuracy once the image has been uploaded by a user. This work only focuses on the detection mechanism of forest fires/wildfires by creating alerts of those incidents and uploading them to Twitter.

Keywords Azure cognitive services · NET · Forest fire · Microsoft azure · Twitter API

1 Introduction

Forest fire is becoming more frequent due to the increase in global warming. The magnitude and intensity of forest fires are also increasing because of the increasing average global temperatures and climate changes. According to [1], the alteration of temperature and precipitation patterns around the world is the result of global warming. Climate and fire scientists in [2] claim that the trend of wildfire activity and will change and increase at the planet warms up more due to global warming. The frequency of extreme weather events will result in the impacts of climate change related to wildfires and it can be more severe in the coming years, rather than the overall change in the average climate patterns which are recorded [3]. Changes in climate can be one of the biggest reasons for the wildfire frequency, size and intensity

R. Kumar · M. Gupta (✉) · D. Kinger · S. Roy
Department of Computer Science, Chandigarh University, Punjab, India
e-mail: gupta.meenu5@gmail.com

with increase in the fire risks and longer and more severe fire seasons [4]. Therefore, the possible future predictions related to wildfire patterns will change drastically and the risk of wildfire related hazards and vulnerabilities such as floods, destruction of infrastructure, economic losses and losses to human life will also increase. The increase risk of forest wildfire also brings the attention to the post-wildfire conditions which can result in acceleration of other natural environmental disturbances resulting in modified vegetation patterns, land degradation, conditions of desertification and disturbances in the hydrological cycle. Due to human exploitation a very small natural changes in climate can result in catastrophic shifts in the ecosystems because of the compromised resilience [5]. The emissions of carbon and other greenhouse gases from wildfires can create feedback loops on the climate and one such unfortunate feedback loop exists which creates a condition where increase in fire takes place and thus greater emissions of greenhouse gases will takes place. This type of feedback loop is very unfortunate [6].

The research aims to get information about forest fires and then spread the news as fast as possible. This objective of this work is to tackle this problem and make a single place for all the forest fires that are reported with a location and a small description about that fire. All the alerts will be uploaded to the official Twitter account for the forest fire. Before the tweet is published, the image will be verified and thus will have a score attached to it. Moreover, the Admin will also get the email regarding the upload. This adds up another layer of verification of the published tweet. The location will provide us with the power to show the place from where the image has been uploaded.

2 Related Work

The authors in [7] discussed about the contribution of geo informatics tools for the indication for forest fire assessment parameters which are the main requirement for the assessments. The study also claims that the burning process can easily be detected in a band of spectrum namely electromagnetic spectrum (EMR). In [8] the research extracted the present and future climate conditions from a Regional Climate Modal (RCM) and then derived fire behavior simulations for the ones which were difficult to control and burnt for five hours based on which a weather hypothesis was made. Last decade, the large fires were occurred in Mediterranean Basin due to extremely warm and dry weather [9] which leads to plant growth and fuel build up because of previous wet season. In [10] the authors proposed a model named KNMI-RACMO2 which showed its accuracy was more when it stimulated the mean climate and the extremes of Europe and the Mediterranean region and it was also compared with other regional evaluation models for the Balkan Peninsula which was completely focused on the climate extremes of the region. Even then it proved to be reproducing patterns of extreme temperature and precipitation with more accuracy than E-OBS gridded observational dataset. In [11] the authors predicted daily risk of lightning fire by applying Maxent model based on 2005–2010 data of Daxingangling

Mountains which will provide new method for spatially explicit assessment of daily lightning fires occurrence. The authors in [12] discussed about a model named as Maxent which along with some relevant concepts is used to study the relationship between fire occurrences and environmental factors and the spatial distribution of fire due to the reasons. The result also showed that how forest fire can be divided into two main causes, natural and human-caused and in those two categories fire ignition and lightning are the main causes for forest fire occurrences, be it natural or human-caused. Among annual lightning fires, Daxingangling Mountains is the most frequent area [13]. According to the study 60% of the fire incidents were of Daxingangling Mountains of Heilongjiang province with burned area of 4833 hectares which was a huge increase according to 1988–2007 records. The authors in [14], advances the modeling approach and also reveals the impact of climatic conditions and demographics on wildfire activity and provides projections of wildfire risk under climate change.

In [15], the authors discussed about controlling large wildfires that burn straight for five hours, weather hypothesis was applied by conducting fire behavior simulations of hard on data from Regional Climate Model (RCM). Historical fire records and meteorological observations revealed that annual number of fires and total burned area was strongly correlated with the mean maximum and the absolute maximum air temperatures according to [16]. The authors in [17] studied that wireless Sensor Networks (WSNs) can be used for industrial automatic control, remote environmental monitoring and target tracking. Similar system can detect and monitor forest fires. In [18] authors discussed the approaches developed for field-like phenomena which are developed on the assumption that the boundaries of such phenomena are crisp, whereas many environmental phenomena such as dust, noise or gas pollution or forest fires have vague spatial boundaries. Hence, it is generally not possible to directly detect the boundaries of monitored phenomena from sensor measurements [19]. This is because sensor nodes are either randomly dispersed over the monitored area or follow a particular pattern [20].

3 Material and Methodology

The API takes care of the image verification, then uploading it to twitter and sending the mail to the Admin. The API is the one that takes all the work to be done on the server end. This paper is based on the integrated work of all the components which are mainly the Front-end App UI, the middleware API and then the Azure APIs'. The Azure takes the image and returns the JSON result which contains certain tags and then the confidence score related to that tag. As in our app we are not considering any tags which below a certain point (for e.g., 0.70). The Fig. 1 description is as follows:

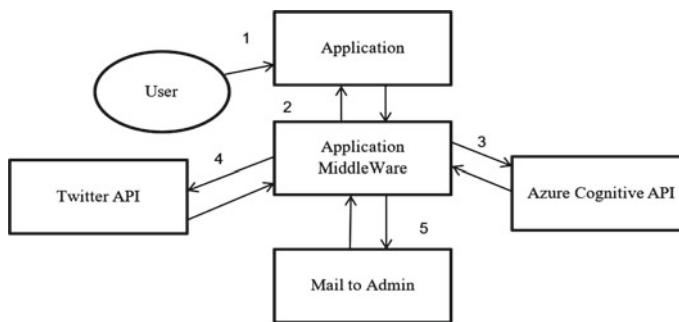


Fig. 1 Block diagram of forest fire research system

Step 1: Get the Image—In this stage we get the image which is being uploaded by the user. The User sends uploads the image on the front-end application.

Step 2: Call the Azure—The image is sent to the application middleware. This middleware is an API that will start the connection with the Azure Cognitive Service.

Step 3: Loop through all the objects in the return—At this stage the connection with the Azure Cognitive Service will be done. Then the image URL will be sent. The response will contain the result of the image verification. Tags and the confidence score of all the tags will be in the return JSON. The return from the step 2 will be used in this stage to get all tags and their confidence score.

Step 4: Add if (confidence > 0.70) condition—The string is created with all the tags that cross a certain threshold of the confidence score from the return. This all takes place in the middleware. Add the tag to a string s.

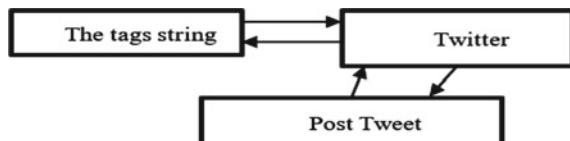
Step 5: Send the tags to the Twitter API—After creating the string with the tags and the message, the image and the string is sent to the Twitter API which in return gives a “Successful Posted” text as response.

Step 6: Sending the Mail to the Admin—The mail is sent to the Admin with all the tweets and their link to the Admin. This stage triggers every 30 s and therefore the Admin gets all the tweets within the time frame of 30 s.

The next challenging part of the dataflow is now to generate the tweet, get authorization and authenticated in twitter and then posts the image or tweet with the respective result (Figs. 2 and 3).

After the twitter phase, the Azure Logic Apps which will be responsible for sending the mail to the admin about the new Tweets that will take place.

Fig. 2 The block diagram for the twitter API and dataflow



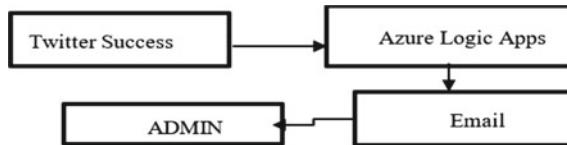


Fig. 3 Block diagram of the admin email

4 Experimental Result Analysis

The outcome of the research can be distributed in four main parts starting from the uploading of the image from the User's end to the mail to the Admin. The four phases can be explained simply with the help of Fig. 4 which tells us the whole dataflow with all the services that are used for this study.

Figure 4 is described as follows: The user uploads the image of the incident. The image is received in the front end and will be sent to the Application Middleware. After this, the image will be sent to be stored in Azure Container as shown in Fig. 5. After the storing of the image in the container, it will respond with the blob URL and from the middleware the image URL will be sent to Azure Cognitive Service. Then it checks whether the image is verified or not based on tags and their confidence score. If the image is verified then we move to the next step else we send the appropriate message to the user. After the uploading of the tweet, it sends a successful response. Then it moves to the next stage where Azure Logic Apps is used. Azure Logic Apps helps in scheduling and automating tasks. The admin gets only those mails which are very specific and sent to the admin with the tweets. As the logic app triggers every 30 s, therefore, all the tweets within those 30 s will be sent in the mail to the Admin.

Figure 6 is the block diagram for storing images in container. The service stores the image in the container in the form of blobs as shown in Fig. 7.

The image verification is done using the Microsoft Azure Cognitive Service. There are tags and confidence score returned for each tag. This score helps for the tagging

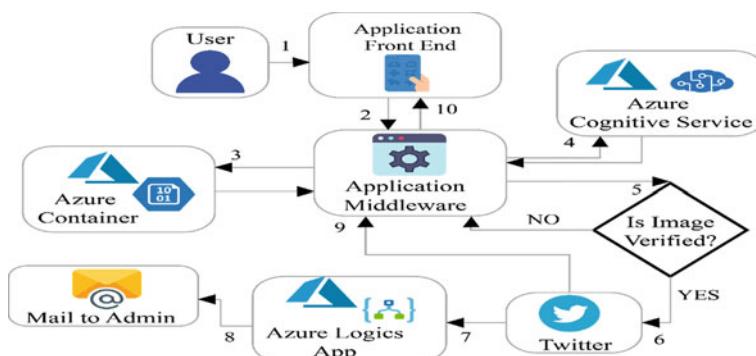


Fig. 4 Block diagram for the entire dataflow with all the services

Fig. 5 Azure storage container

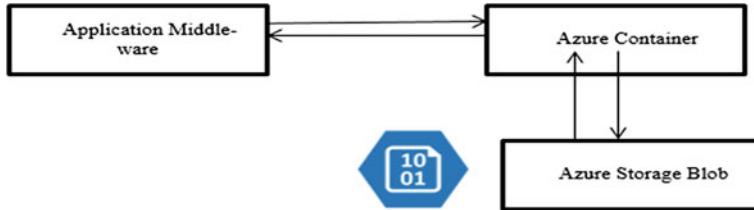


Fig. 6 Storing image in Azure storage container

Name	Modified	Access tier	Blob type	Size	Lease state	...
080dd6ba6-39fb-4cec-abb5-24b81a...	3/12/2021, 10:18:30 ...	Hot (Inferred)	Block blob	655.27 KiB	Available	***
0cdd2a13-6494-4736-b401-e127a6...	4/19/2021, 10:51:20 ...	Hot (Inferred)	Block blob	518.99 KiB	Available	***
3a325866-cd98-4472-a301-83e970...	4/19/2021, 11:00:26 ...	Hot (Inferred)	Block blob	607.08 KiB	Available	***
43762f50-d549-435e-a0e-733ea42...	3/12/2021, 11:08:04 ...	Hot (Inferred)	Block blob	716.26 KiB	Available	***
4cd566bb-43fb-4c19-b58d-d43f280...	4/19/2021, 10:40:57 ...	Hot (Inferred)	Block blob	594.71 KiB	Available	***
5ce2d0e5-c1ee-4b88-9172-209569f...	4/19/2021, 10:46:09 ...	Hot (Inferred)	Block blob	584.35 KiB	Available	***
76271019-eab0-49d4-ac8c-f7489ee...	3/12/2021, 10:16:51 ...	Hot (Inferred)	Block blob	550.19 KiB	Available	***
afcb2ba6-100f-40c3-9e4f-5feb6710...	4/19/2021, 10:54:20 ...	Hot (Inferred)	Block blob	526.08 KiB	Available	***
e4de3d77-66dc-4aee-8176-b91d9f1...	4/19/2021, 10:48:30 ...	Hot (Inferred)	Block blob	565.77 KiB	Available	***

Fig. 7 Image stored in the Azure container

of the image on the twitter end. The score has a specific threshold beyond which only the tag is being accepted. If the score is not up to the mark the tag won't be added to the twitter.

```
[ { "name": "grass", "confidence": 0.999268651}, { "name": "fog", "confidence": 0.984427333}, { "name": "outdoor", "confidence": 0.974131346}, { "name": "weapon", "confidence": 0.917561769}, { "name": "fire", "confidence": 0.8600407}, { "name": "tree", "confidence": 0.82458353}, { "name": "transport", "confidence": 0.8238987}, { "name": "screenshot", "confidence": 0.813338161}, { "name": "smoke", "confidence": 0.7059078}, { "name": "firefighter", "confidence": 0.677968}, { "name": "spring", "confidence": 0.320403636}, { "name": "rocket", "confidence": 0.1832931}].
```

The Table 1 shows the tag names and the confidence score that each tag has scored. These tags having a high score is also a great indication toward the forest fire and thus for the research only the tags having score of greater than 0.70 has been accepted.

Table 1 Tags and confidence scores

Name	Confidence score
Grass	0.99
Fog	0.98
Outdoor	0.97
Weapon	0.92
Fire	0.86
Tree	0.82
Transport	0.82
Screenshot	0.81
Smoke	0.70
Firefighter	0.67
Spring	0.32
Rocket	0.18

The Table 2 shows the tags those has been selected based on the confidence score. These are the tags that make up the message in the next phase. After the authorization and the authentication of the twitter then comes the part where the image string along with the image is posted by using a request: <http://twitterapidk.azurewebsites.net/test?message=#####&&url=#####.png>. This is the request with the help of the message as well as the image will be posted on the twitter account. The API with which an image is uploaded will return the response as: // 20210418201008// <http://twitterapidk.azurewebsites.net/test?message=#####&&url=#####.png> “Successfully Posted!!”. This phase completes the challenge for the uploading of the tweet with information. Figure 8 is showing the model for the Azure logic App that has been used.

In Fig. 9 all the configurations for the model to work in the Logic App is shown. For the experimental basis, different tags and different refresh times were used but the best results came with the ones that are shown in Fig. 9.

Table 2 Tags with greater than threshold value (0.70)

Name	Confidence score
Grass	0.99
Fog	0.98
Outdoor	0.97
Weapon	0.92
Fire	0.86
Tree	0.82
Transport	0.82
Screenshot	0.81

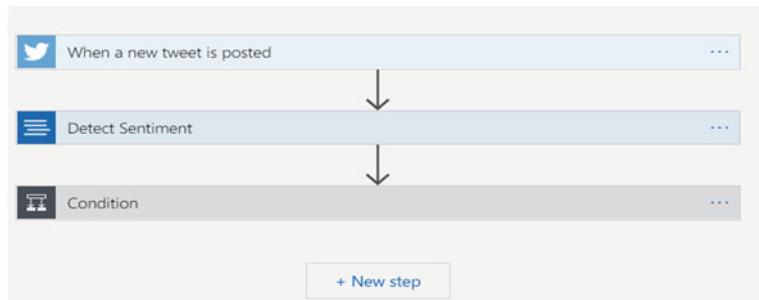


Fig. 8 Model of Azure logic app

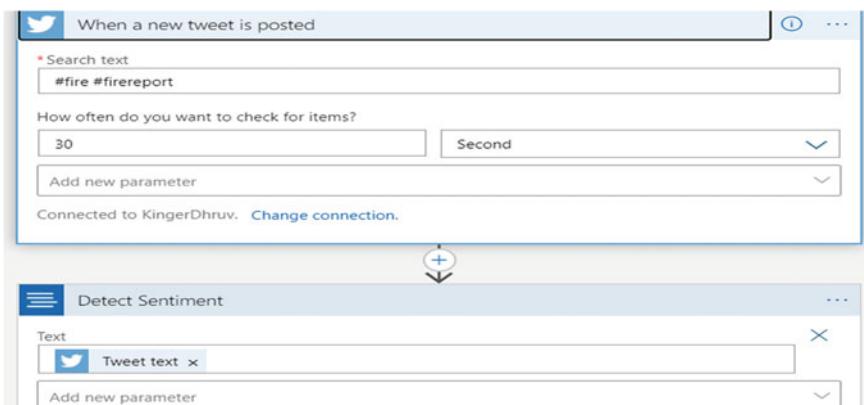


Fig. 9 Configuration of the logic app

Figure 10(a) and (b) shows the conditions in which the mail will be sent. In some experimental runs, the conditions were change but the condition which is being used in the image worked the best. The Fig. 10(a) shows the first dialog box, the main condition where the score for the search text is being checked for greater than or equal then a threshold value. In this study the threshold value was kept at 0.50 which is 50%. If the condition is true then the Fig. 10(b) second dialog box will run in which we are sending the tweet in a very specific format to a specific test email, which is in the “To” field. For the other dialog box which in case of false, no mail will be sent and no action will be taken. The result of this can be seen in Figs. 11 and 12.

Tweet was #grass #fog #weapon #fire #tree #transport #screenshot #outdoor
<https://t.co/FmTOQ30U5W> **Image Uploaded From-Sadhia Mohalla, Old Market, Khanna and was tweeted by KingerDhruv.**

The mail contains all the message along with the tags and the link to the tweet, then the location from where the image has been uploaded is also added in the mail and the account through which it has been tweeted. In this experimental analysis, the

(a)



(b)

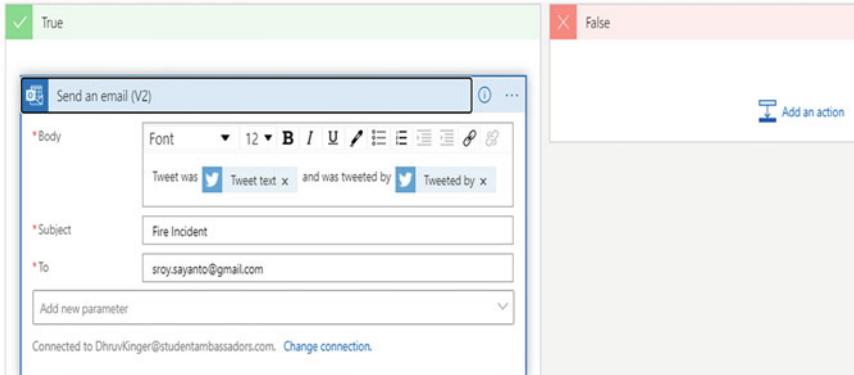


Fig. 10 **a** Condition for the Azure logic app to go to next step, **b** next screen where condition is true or false is checked

Specify the run identifier to open monitor view directly			
Status	Start time	Identifier	Duration
Succeeded	4/19/2021, 10:57 PM	08585827536293317428373905425CU...	1.77 Seconds
Succeeded	4/19/2021, 10:57 PM	08585827536293317429373905425CU...	1.71 Seconds
Succeeded	4/18/2021, 9:57 PM	08585828436564173561972719169CU...	658 Milliseconds

Fig. 11 Azure logic apps runs for sending the mails to the admin



Fig. 12 Mail that has been sent from the logic apps

Logic App is triggering every 30 seconds, therefore it sends all the tweets as separate emails and therefore those are not confusing at all.

5 Conclusion and Future Scope

This work is focusing on the forest fire alerting and then spreading the news as fast as possible. The proposed work can be really transformed into an automated as well as user dependent system. It can be made publicly available by uploading on play store and IOS app store. Fire analysis can be increased with the help of more sophisticated visual tools and analyzers. Addition of a dashboard and map to show all the recent fires which has been reported. Add an in-app feeds system where recent forest fires can be showcased. The addition of the image processing and detection using the more advanced techniques can be used like VR addition for the monitoring of the place in real time with satellite imaging. The proposed system is made for research only and to test whether such system can be reliable in the long run or not. This can be verified by testing it against different forest fire detection systems that are available for public use.

References

1. Al Jansen, Paasche Ø (2007) IPCC [Intergovernmental panel on climate change]. climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the IPCC. Cambridge, United Kingdom Cambridge University Press. Flannigan MD, Logan KA, Amiro BD, Skinner WR, Stocks BJ Future area burned in Canada. *Clim Chang*
2. Iannigan M, Logan KA, Amiro B, Skinner W, Stocks B (2005) Future area burned in Canada. *Clim Change* 72:1–16
3. Mackay A (2008) Climate change 2007: impacts, adaptation and vulnerability. Contribution of working group ii to the fourth assessment report of the intergovernmental panel on climate change. *J Environ Qual* 37:2407–2407
4. Stocks BJ, Fosberg MA, Lynham TJ et al (1998) Climate change and forest fire potential in Russian and Canadian boreal forests. *Clim Change* 38:1–13
5. Scheffer M, Carpenter S, Foley JA, Folke C, Walker B (2001) Catastrophic shifts in ecosystems. *Nature* 413:591
6. Joseph S, Anitha K, Murthy M (2009) Forest fire in India: a review of the knowledge base. *J Forest Res* 14:127–134
7. Kalabokidis K, Palaiologou P, Gerasopoulos E, Giannakopoulos C, Kostopoulou E, Zerefos C (2015) Effect of climate change projections on forest fire behavior and values-at-risk in Southwestern Greece. *Forests* 6:2214–2240
8. Founda D, Giannakopoulos C (2009) The exceptionally hot summer of 2007 in Athens, Greece—a typical summer in the future climate? *Global Planet Change* 67:227–236
9. Kostopoulou E, Giannakopoulos C, Hatzaki M, Tziotziou K (2012) Climate extremes in the NE Mediterranean: assessing the E-OBS dataset and regional climate simulations. *Climate Res* 54:249–270
10. Chen F, Du Y, Niu S, Zhao J (2015) Modeling forest lightning fire occurrence in the daxinganling mountains of North-eastern China with MAXENT. *Forests* 6:1422–1438
11. Parisien M-A, Moritz M (2009) Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecol Monogr* 79:127–154
12. Shu LF, Wang MY, Tian XR, Li ZQ, Xiao YJ (2009) Fire environment mechanism of lightning fire occurrence in Daxinganling region China. *Sci Silv Sin* 22:18–24
13. An H, Gan J, Cho SJ (2015) Assessing climate change impacts on wildfire risk in the united states. *Forests* 6:3197–3211

14. Westerling A, Hidalgo H, Cayan DR, Swetnam T, Warming and earlier spring increase western US forest wildfire activity. *Science* 1161:1–9
15. Kalabokidis K, Palaiologou P, Gerasopoulos E, Giannakopoulos C, Kostopoulou E, Zerefos C (2015) Effect of climate change projections on forest fire behavior and values-at-risk in Southwestern Greece. *Forests* 6(6):2214–2240
16. Koutsias N, Xanthopoulos G, Founda D, Xystrakis F, Nioti F, Pleniou M, Mallinis G, Arianoutsou M (2013) On the relationships between forest fires and weather conditions in Greece from long-term national observations (1894–2010). *Int J Wildland Fire* 22:493–507
17. Liu Y, Liu Y, Xu H, Teo KL (2018) Forest fire monitoring, detection and decision making systems by wireless sensor network. *Chin Control Decis Conf (CCDC)* 2018:5482–5486
18. Kulik LA, Geometric theory of vague boundaries based on Supervaluation. In: Montello D (ed) Spatial information theory SE-4, Lecture notes in computer science, vol 2205. Springer: Berlin/Heidelberg, Germany
19. Jadidi A, Mostafavi MA, Bédard Y, Shahriari K (2014) Spatial representation of coastal risk: a fuzzy approach to deal with uncertainty. *ISPRS Int J Geo-Inf* 3:1077–1100
20. Swami V, Kumar S, Sanjay Jain (2018) An improved spider monkey optimization algorithm. *Soft computing: theories and applications*. Springer, Singapore, pp 73–81

Role of Telemedicine in Healthcare Sector for Betterment of Smart City



Prashant Sahatiya and Dheeraj Kumar Singh

Abstract Proactive public health monitoring is one of the most important features of a smart city. Telemedicine plays a vital role in monitoring health hazards of the citizen. In its widest sense, telemedicine may be defined as the use of telecommunication technology, and electronic health records to provide medical information and services. A telemedicine system is a data-driven solution used to identify diseases and their patterns remotely. This system helps doctors in diagnosing the patient living in urban as well as rural areas. It also helps health workers and other city officials to monitor public health hazards, and seasonal diseases. Telemedicine is one of the smart cities' finest features: smart hospitals. In this paper, we discussed the concept useful in the advancements in healthcare services based on telemedicine for smart cities. Additionally, different telemedicine frameworks implemented in literature are described to understand state of art work done in healthcare sector.

Keywords Smart city · Healthcare · Telemedicine

1 Introduction

In India from last few decades, we have ranked many cities in top 10 smart cities on the basis of various phenomenon like economic growth, healthcare facilities, cleanliness services, transportation facilities, and much more [1]. In today's globe, urbanization is one of the most important social and economic developments. Cities currently house 50% of the world's population. By 2050, as urbanization accelerates, this proportion is predicted to surpass 70% [2]. Smart city word itself gives an idea that it is a city that has a mission to enhance the quality of life and provide economic stability to the people of the city. Innovation and multidisciplinary research are important factors in making a city smart [3]. A city does not become smart only with innovative

P. Sahatiya (✉) · D. K. Singh

Department of Information Technology, Parul University, Vadodara, Gujarat, India
e-mail: prashant.sahatiya270187@paruluniversity.ac.in

D. K. Singh
e-mail: dheeraj.singh@paruluniversity.ac.in

technologies but also requires make people use these technologies. A city becomes smart when people living in that city become addicted to these developed smart technologies. Among the smart cities in India, Ahmedabad is on the 1st position followed by Amritsar and Rajkot, we also have Coimbatore, Hyderabad, Chennai, Indore, Bhopal, Nagpur, Pune in the top 10 smart cities list.

Smart Cities have different features such as smart infrastructure, smart technologies, smart governance, smart living, and smart healthcare. Most of features are for citizen comfort and entertainment [4]. However, comfort and entertainment are less important in comparison to our health. Healthcare services are the ones that should be taken care and prioritize for research. Without a healthy life, everything is useless. Healthcare + Telecommunication gives us a technical word telehealth and another related word telemedicine [5].

The topographical issues and obstacles in a smart healthcare system may be solved by the use of telemedicine and information and communication technologies [6]. Information Communication Technology (ICT) enables developing and developed countries to provide healthcare services at affordable cost and to improve the availability of healthcare facilities given at remote locations [7].

1.1 *Telemedicine*

According to the Greek language “tele” word is used for “distance” and Latin language mederi stands for “to cure”. Telemedicine has been dubbed “healing by wire” by Time magazine. Telemedicine was formerly thought to be “futuristic” and “experimental” however, now it became a reality. Better medical education facilities, care for the public and patients, research for a better lifestyle and healthcare facilities all are the advantages of Telemedicine. The World Health Organization (WHO) defines telemedicine as the delivery of healthcare services by all healthcare professionals utilizing ICT tools for the exchange of information and diagnosis from a remote location. Treatment and prevention of illness and injury, medical research and assessment are the main goals of WHO to improve the health of individuals as well as communities [8].

People in rural and distant places across the world struggle to get timely, high-quality medical treatment. Travelling long distance for the patient can create a tedious condition for the critical patients. According to our literature study, most experienced doctors and physicians are available in metropolitan cities containing more population and dense area. Telemedicine offers the ability to utilize these doctors remotely and make healthcare services more accessible in long distance and rural locations [9]. The main reason to bring telemedicine into existence is to provide clinical support for distance cure. The focus of area is to overcome all the issues which are occurring due to geographical barriers, making connections with the patients who are not near by the physician’s location or are from a different city. The block diagram shown in Fig. 1 shows a general telemedicine system that works in client–server architecture. The diagnosis starts from the initial screening of the patient through a video call by a

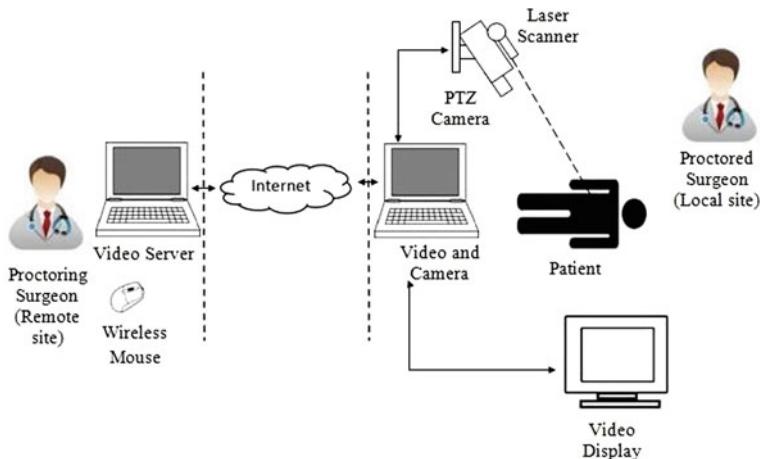


Fig. 1 Block Diagram for Telemedicine System [11]

local doctor available at clinics or centers like Primary Telemedicine Center (PTC), Secondary Telemedicine Center (STC), and Tertiary Telemedicine Center (TTC). Further, the local doctor connects with concerned specialist doctors at Telemedicine Consulting and Specialist Center (TCC) [10] for remote diagnosis.

1.2 Classification of Telemedicine

Telemedicine can be classified into three main categories: (i) Interactive telemedicine, (ii) Remote patient monitoring, and (iii) Store-and-forward.

Interactive telemedicine, also known as telehealth enables real-time communication between doctors and patients. This consultant process are usually done from patient's location through nearby medical health center. Telephone talks, and HIPAA—compliant video conferencing software are common examples of such telemedicine interactions.

Remote patient monitoring uses mobile telemonitoring system to observe patients from their home. The mobile telemonitoring system is a term for devices that collect data such as person's temperature, blood sugar levels, blood pressure, and other critical indications.

The store-and-forward is another mechanism of telemedicine sometimes known as asynchronous telemedicine. It allows healthcare professional to communicate patient information with another professional, and prescribe the required treatment to a patient, while maintaining their own record.

2 Related Work

Based on our literature survey and interaction with the domain expert we found that a smart responsive healthcare system will be more effectively implemented with the integration of telemedicine and ICT tools. Researchers have previously focused on the problems of providing services related to the health care with the help of telemedicine and image compression methods importance for effective storage and transmission. Due to the widely dispersed populations, effective medical services and health care have a significance influence on the lives of people in rural regions. India is the world's second most populated country, with a total area of 3.28 million km² and a population of approximately 1.35 billion people [12]. According to our study we got to know that is 70% of India's population which is around 800 million people live in residential areas on the periphery of cities, with no direct access to even fundamental health care. The availability of health care services in the suburbs has not kept up with the demand for these low-income and underserved communities. Telemedicine brings health professionals expertise and experience humble attachment to patients, resulting in better medical care at a lower cost. It's on track to have a significant beneficial influence on the overall healthcare system, and it's ready to make your life a little simpler.

Integrating technology into the city's healthcare system can save the city and its resident's money in the long term. The cost of the technology, on the other hand, may dissuade communities from adopting it in the medium term. Cities could apply a similar approach to insurance companies in order to pay the short-term costs of healthcare technology while also addressing another obstacle to adoption: the time it takes to learn about and use it. In recent years, numerous insurance companies have offered discounts to customers who agree to install telematics in their vehicles that track driving habits. The information can improve road safety as well as lower premium rates because the insurance app can notify users when they are driving too recklessly. During the preceding decade, other service providers, such as utility companies, also provided incentives. Health insurance firms may provide similar incentives in the future to encourage people to install ICT in their homes. The technology will enable citizens to receive high-quality, cost-effective healthcare whenever and wherever they require it. Other Smart City Services can be connected to Smart cities make extensive use of ICT.

The purpose of smart city platforms is to enable plug-and-play smart devices that can be put anywhere and fit in with their environment. The items should be able to monitor not just health but also structures, the environment, security, and intelligent transportation. As we know that from 2020, almost 10% Streetlamps used as the backbone for citywide wireless networks in smart cities. Weaving sensors in the already available city elements is one of the ways how data is becoming increasingly connected, which might help academics better grasp the link between city design and health. By monitoring both air quality and behavior, we can understand how our

behavioral patterns and design decisions affect air quality. We may also assess the effect of poor air quality on health and devise solutions such as changing city design or providing residents with real-time information to urge them to stay indoors during inclement weather.

There are a lot of medical policies for privacy and security of patient data limits the research in healthcare sector. Additionally, the construction cost and time to build the hospital or a medical center limits the utilization of healthcare facilities at affordable cost. The use of telemedicine solves many such issues. Since data is being shared, there must have the privacy and security risks of telemedicine. Some of data privacy and security issues can be solved by enhancing database security and decrease the risk of data sharing illegally.

For telemedicine, public safety, and maintenance, our research answer is a wearable video system. Our research project is a wearable video system for telemedicine, public safety, and maintenance. From an environmental standpoint, the ability to display someone live point-of-view video of a difficult injury, situation, or any problem allows a remote individual to provide professional guidance. This can save remote professionals in medical, IT, and public safety from making several unneeded trips. This is a significant contribution to environmental sustainability and the green mile.

2.1 Existing Telemedicine Framework

In this section, a few frameworks have been discussed to understand working and utilization of technologies for smart health care. For smart health care in cardiology, a remotely accessible telemedicine system Wireless Body Area Networks (WBAN) has been used for monitoring a patient's ECG recorders consisting of an instrumentation amplifier, a microprocessor with electrically erasable programmable read-only memory, and a low transmit power mobile phone. The chest electrodes are connected to an ECG recorder, which may be viewed via the doctor's laptop computer. This ECG recorder continuously monitors your heart rate. If the doctor wishes to monitor the patient's ECG, he can use his laptop to call the ECG recorder. The ECG is sent to the laptop using this recorder. In addition, if the patient has chest pain any time, he can activate the data transmission switch on the recorder, which will send ECG waveforms recorded two minutes before and after the switch is activated. After then, a signal is delivered to the physician's system. The gadget might be wireless, with a hub put on the patient to detect abnormal patterns.

The eHealth use case is another prioritized emergency video call for ambulance services via 5G network slicing. This is a smart city concept that solves confronting ever-growing demand of ambulance services, traffic congestion, and carbon emissions. Instead of sending an ambulance, emergency services respond to emergencies from a control center and provide treatment and advice to patients over the phone

Table 1 Existing telemedicine framework

Platform	Why to use?	Why not to use?
Mend [13]	<ul style="list-style-type: none"> • Simple to use, • Streamline operations, • Less number of missed appointments 	<ul style="list-style-type: none"> • There is no free trial • Expensive to use
Doxy Me [14]	<ul style="list-style-type: none"> • Trial and free tiers are available • Self-hosted alternative, • Easily accessible from the web browser 	<ul style="list-style-type: none"> • Costly due to subscriptions plans, • Lack of mobile application
AMC Health [15]	<ul style="list-style-type: none"> • Clinical research is possible • Comprehensive suite of services • Bluetooth device alternatives 	<ul style="list-style-type: none"> • It's unclear which devices are supported, • Expensive to use
swyMed [16]	<ul style="list-style-type: none"> • Low-bandwidth capability, • Uses hospital equipment to collect data, • Provides emergency treatment on the move 	<ul style="list-style-type: none"> • Not available 24 * 7 • Expensive to use
Teladoc [17]	<ul style="list-style-type: none"> • Services are available 24 * 7 	<ul style="list-style-type: none"> • No tests are ordered by doctors • Price model is opaque

via video. This reduces transportation cost, lowers carbon emissions, improves operational time and efforts, and speeds up patient care. This results in a more environmentally friendly ambulance and medical transportation. Few other telemedicine frameworks have been depicted in Table 1.

2.2 *Impact of Telemedicine in Smart City*

Telemedicine has a great impact on healthcare for betterment of life styles of citizens in smart city. This section briefly describes the important features of telemedicine which can be utilized in smart city.

- (i) **Telemedicine for medical and healthcare**—It prevents travels, absences, or needless contagions and promotes calm, which has a very beneficial influence on well-being in the workplace and home.
- (ii) **No transportation time or costs**—Citizens save money on petrol, parking, and public transit by consulting doctors via their mobile device or computer. The best part is you won't spend time commuting or risk being stuck in traffic congestion, which will cause you to be late for your appointment or, a worse health hazard.
- (iii) **No need to take time off of work**—When it comes to work, video call eliminates the need to leave the job. You can consult a doctor during lunch break, or before or after work. You may easily take follow-up advice from the doctor

- and stay healthy without missing work or squandering your valuable paid vacation time.
- (iv) **Eliminate child or elder care issues**—It might be difficult and expensive to find alternate treatment so that you can see a doctor. It might be difficult or impractical to bring them along. Fortunately, telemedicine overcomes this problem by allowing you to see your doctor while taking care of your whole family.
 - (v) **Better health with on-demand options**—These days, most of doctors and clinical professional offers treatment, advice, and other options over telemedicine as per demand from patients. You can better manage your medicine, lifestyle, and other chronic problems by consulting doctor's frequently using telemedicine without the hassles of getting into the office.
 - (vi) **Access to specialists**—Some patients who require specific treatment over long span of time have to travel considerable distances and devote a significant amount of time to each appointment. Telemedicine allows people to consult a specialist doctor to get follow-up advice and treatment based on experience of experts from their home location.
 - (vii) **Less chance of contracting a disease**—In hospitals, doctors try to cure the patient while making efforts to keep them away from contracting a new disease from another patient. But it is always possible, especially in busy waiting rooms. Staying at home allows you to obtain the treatment you need while avoiding exposure and the danger of spreading your sickness to others.
 - (viii) **Less Time in the Waiting Room**—You'll save time at the doctor's office if you pick a video visit with telemedicine technology. Even if you don't use telemedicine, finding a clinic that does will cut down on your waiting time by allowing other patients to be seen from their homes.

3 Telemedicine Framework Useful for Smart City

The main aim of this article is to bring the attention of researchers and medical professionals to healthcare projects that employ ICT to provide smart healthcare for the smart city. The researchers and municipal planners have made significant progress in integrating ICT tools and the healthcare sector to provide smart healthcare solutions. The goal of these smart solutions is to provide long distance cure services to citizens. A patient can easily consult any doctor from these solutions. After initial screening, further treatment can be guided by video call or any other options given in telemedicine. A patient or even a doctor can contact any senior physician using such a telemedicine application. These solutions have helped us a lot during COVID-19 [18, 19, 4, 20] situation, a lot of patient's lives have been saved. The telemedicine framework can be categorized into four categories: (i) Asynchronous telemedicine, (ii) Synchronous telemedicine, (iii) Remote patient monitoring, and (iv) Mobile health.

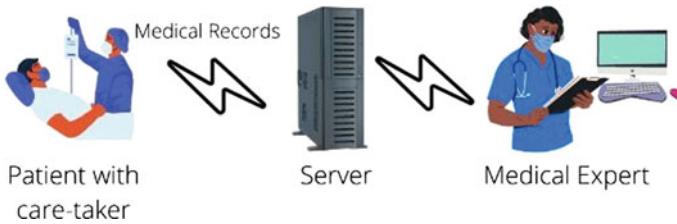


Fig. 2 Patients health record send to medical expert

3.1 Asynchronous Telemedicine

As demonstrated in Fig. 2, an asynchronous telemedicine system comprises sending medical pictures and patient records to a healthcare professional at a convenient moment for a comprehensive inspection. This technique is frequently utilized in non-emergency situations. Asynchronous telemedicine does not need both physicians and other medical professionals together for treatment at the primary level. This sort of telemedicine is used in specialized medical domains such as radiography, telepathology, and teledermatology.

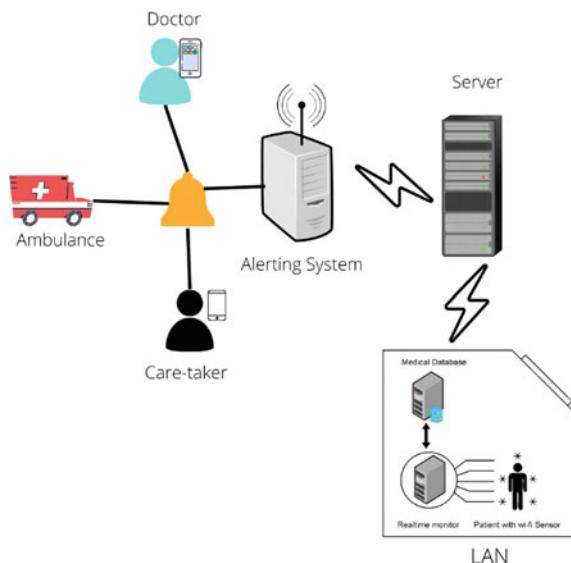
3.2 Synchronous Telemedicine

The synchronous telemedicine system, also known as interactive telemedicine uses synchronous video conferencing, which allows patients and specialists to engage in real-time. The live video conferencing solution makes it easier to provide crucial patient information quickly and enhances overall treatment quality. Synchronous video conferencing systems are used in advanced medical systems such as tele-radiosurgery, which involves the vision of the patient's treatment. Figure 3 shows a general architecture of synchronous telemedicine.



Fig. 3 Video conferencing with patient

Fig. 4 Use of monitoring system



3.3 *Remote Patient Monitoring*

The medical professional can monitor patient's bio-signals from a distance using specialized equipment using a remote patient monitoring system like the one depicted in Fig. 4. This device is especially beneficial for monitoring patient using signals in elderly people with chronic conditions and generating an alert to medical specialists in case of emergency [21]. It gives huge relief to the patients by reducing time spent in the clinic and provides significantly better care.

3.4 *Mobile Health*

In a new era of smart healthcare systems, mobile health also known as M-Health is one of the most useful telemedicine frameworks, where users can access services through mobile. This system is significantly used in modern healthcare monitoring and alert system for clinical data storage and maintenance. This method sends data across 3G and 4G mobile networks utilizing smart mobile devices. The medical records are available for viewing by physicians and patients at any time and from any location, and the patient can call the specialist in an emergency.

4 Discussion

To develop a smart city, their people should be able to get enough healthcare services options to cure their illnesses. From the analysis of literature and existing applications related to healthcare, we found several issues such as ease of access of services, uniformity and unification of services. Most of the existing telemedicine services use Virtual Private Network (VPN) setup to provide the required bandwidth from the network provider. A commercial contract is often required to set up such a VPN which pose a barrier to agencies in collaboration for healthcare services.

Using a web application that employs a 5G networks to provide video conferencing services. 5G networks provide a unique technique for slicing bandwidth dynamically so that an emergency call can be prioritized by allotment of additional bandwidth. This will help in important decision-making for critical patients by respective doctors. Slicing technology allows bandwidth to be allotted dynamically, resulting in a significant boost in efficiency for both consumers and telecom operators. Additionally, it allows operators to customize the network more precisely as per their customers' demands. Overall dynamic bandwidth distribution, reduces the infrastructure and other network resources overhead which boosts the sustainability of telemedicine services. Moreover, telemedicine applications should be user-friendly for both patients and doctors. There can be an application that gives the occurrence of the disease in humans based on common symptoms from information recorded in clinics and hospitals [22].

5 Conclusion and Future Work

From the above study and research, we found that telemedicine is an important part of a smart healthcare system. This research provides a thorough and systematic examination of telemedicine in a variety of contexts such as smartness, usage of ICT. The usage of telemedicine services such as telehealth prevents physical contact to limit the risk of infection or transmission of another disease. Furthermore, these services can provide medical treatment to a large community by utilizing the technology and expertise of different healthcare professionals. According to recent studies, doctors and patients are highly encouraged to adopt smart healthcare system as a viable option for avoiding infection and making our city and nation healthy. According to Software advice's research, 75% of survey respondents wish to utilize telemedicine. Telemedicine has been utilized in a limited capacity for decades, but now it is gaining attention. This is due to the fact that so many individuals have access to high-speed internet and the gadgets required to conduct a video conference.

Future studies should look at the effective collaboration and integration of agencies involved in the healthcare sector. In the future, common difficulties such as a lack of a specialist team of doctors, regulatory support, and policy impediments should be preciously taken care of while implementing any telemedicine framework.

Furthermore, we should focus on the development and deployment of telemedicine kits and portable carts, medical professionals' ongoing education and staff training, video conference facilities for better observation of patients.

References

1. Telecommunication Infrastructures for Telemedicine in Smart Cities Volodymyr Pasichnyk, Natalia Kunanets, Serhii Martsenko, Oleksandr Matisiuk, Olesia Mytnyk, Oleksii Duda, and Paweł Falat (2018)
2. Papadimos TJ, Marcolini EG, Hadian M, Hardart GE, Ward N, Levy MM et al (2018) Ethics of outbreaks position statement. Part 2: family-centered care. *Crit Care Med*
3. Organization WHO (2010) Telemedicine: opportunities and developments in member states. Report on the second global survey on eHealth: world health Organization
4. Greenhalgh T, Wherton J, Shaw S, Morrison C (2020) Video consultations for covid19. *Br Med J Publishing Group*
5. Charles BL (2000) Telemedicine can lower costs and improve access. *Healthc Financ Manage*
6. Gabellone A, Marzulli L, Matera E, Petruzzelli MG, Margari A, Giannico OV, Margari L (2022) Expectations and concerns about the use of telemedicine for autism spectrum disorder: a cross-sectional survey of parents and healthcare professionals. *J Clin Med* 11(12):3294
7. Mehrotra A, Jena AB, Busch AB, Souza J, Uscher-Pines L, Landon BE (2016) Utilization of telemedicine among rural Medicare beneficiaries. *JAMA* 315(18):2015–2016
8. Sauers-Ford HS, Hamline MY, Gosdin MM, Kair LR, Weinberg GM, Marcin JP et al (2019) Acceptability, usability, and effectiveness: a qualitative study evaluating a pediatric telemedicine program. *Acad Emerg Med*
9. Morenz A, Wescott S, Mostaghimi A, Sequist T, Tobey M (2019) Evaluation of barriers to tele-health programs and dermatological care for American Indian individuals in rural communities. *JAMA Dermatol*
10. Fortney JC, Pyne JM, Edlund MJ, Williams DK, Robinson DE, Mittal D et al (2007) A randomized trial of telemedicine-based collaborative care for depression. *J Gen Intern Med*
11. Valle J, Godby T, Paul III DP, Smith H, Coustasse A (2017) Use of smartphones for clinical and medical education. *Health Care Manag*
12. Rangasamy M, Balasubramaniam A, Krishnarajan D, Raviteja A, Kante N, Kumar N (2011) Role of telemedicine in health care system: a review
13. Integrated Telehealth & Patient Engagement Platform. Mend. <https://mend.com/>
14. The simple, free, and secure telemedicine solution. Doxy Me. <https://doxy.me/en/>
15. Remote Patient Monitoring. AMC Health. <https://www.amchealth.com/>
16. REAL-TIME VIDEO FOR TELEMEDICINE. swyMed. <https://swymed.com/>
17. Teladoc Health: Virtual Care & Telehealth Solutions. TelaDoc Health. <https://www.teladochealth.com/>
18. Hollander JE, Carr BG (2020) Virtually perfect? Telemedicine for covid-19. *N Engl J Med*
19. Canady VA (2020) COVID-19 outbreak represents a new way of mental health service delivery. *Ment Heal Wkly* 2020
20. Naik N, Ibrahim S, Sircar S et al (2022) Attitudes and perceptions of outpatients towards adoption of telemedicine in healthcare during COVID-19 pandemic. *Ir J Med Sci* 191:1505–1512
21. Mahajan R (2018) Emotion recognition via EEG using neural network classifier. *Soft computing: theories and applications*. Springer, Singapore, pp 429–438
22. Yadav S, Saroliya A (2018) Patient-specific modeling for pattern identification in clinical events using data mining models and evolutionary computing. *Soft computing: theories and applications*. Springer, Singapore, pp 209–216

A Comparative Study on Abstractive Text Summarization Techniques Using Deep Learning (ATS-DL)



S. Adhithyan, A. R. Nirupama, S. Sri Akshya, S. Swamynathan, and K. Girthana

Abstract The amount of textual data has significantly expanded in recent years, thereby facilitating easier information extraction and analysis. To retrieve useful knowledge within a reasonable time period, this information must be summarized. Though lot of work has already been done in abstractive text summarization (ATS), exploring ATS using deep learning techniques still has some major issues and lot of research openings. Hence, deep learning-based approaches have been used for implementing ATS. We have built ATS model using LSTM, BiLSTM, GRU and transformers along with attention mechanisms embedded in it. Gigaword dataset was used for training and comparing the models. Our system is compared using standard benchmarking systems like ROUGE-1, ROUGE-2 and ROUGE-L. Among these models, the pre-trained T5 model had a better abstractive summary at the same time, improving the ROUGE score by 9%.

Keywords Abstractive summarization · LSTM · BiLSTM · Transformers

1 Introduction

Text summarization is a technique to get the most important information from a source document, i.e. in a shortened version of the source (summary) that is precise and correct enough to get the gist of the document. Extractive summarization and abstractive summarization are the two forms of summarization based on the nature of processing. Extractive text summarization (ETS) aims at extracting the important phrases of a source document and compiles them as a concise summary. Here, there is no generation of sentences and it cannot generate a summary like humans. Abstractive text summarization is the new state-of-the-art method that will generate sentences

S. Adhithyan · A. R. Nirupama · S. Sri Akshya · S. Swamynathan · K. Girthana (✉)
College of Engineering, Anna University, Guindy, Chennai, Tamil Nadu, India
e-mail: keerthi3110@gmail.com

S. Swamynathan
e-mail: swamyns@annauniv.edu

rather than extracting from the original text, which could best represent the whole text. It makes the computer understand the text and generates a non-biased human like summary on its own. It may not contain the words from the source.

There are semantic-based approaches, structure-based approaches and deep learning-based approaches for ATS. Structured-based approaches encode most significant information from the original document using cognitive schemes such as templates, extraction rules. In the semantic-based approach, the semantic representation of the original document is used to feed into natural language generation (NLG) system.

Deep learning algorithms analyse complex problems which facilitate the decision-making process. Deep learning algorithms attempt to imitate what the human brain can do. The deep learning approach for text summarization provided excellent results compared to semantic and structured-based approaches. Due to this, deep learning approaches are extensively used for text summarization. Algorithms include long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), gated recurrent units (GRUs), stacked LSTM, LSTM with attention mechanism, stacked LSTM with attention mechanism and transformers.

2 Literature Survey

The advancement in deep learning has helped to solve many NLP problems, and text summarization can be improvised by using deep learning algorithms. The common steps for abstractive summarization are to create a tokenizer to map the tokens from the source text into a vector and then give the vector some meaning values using word embedding techniques. The vector representation is fed into the encoder which produces a vector representation which in turn is fed into the decoder to generate the summary. The literature survey is divided into LSTM-based models [3], models with attention mechanism and transformer models. Most of the models proposed in the research papers used news articles data for training purposes. Gigaword dataset is produced by Linguistic Data Consortium, and ISBN was the commonly preferred dataset for abstractive text summarization. It contains all articles to heading mappings that are used as text and summary. It helps encoder identify intrinsic semantics of the language while keeping the size of the text intact.

Word embedding helps in giving values to the vector representation such that similar words have similar vector representations but not the same. The issues faced in recurrent neural network model (RNN) [10] are resolved in the LSTM-based models. GRU cells always give better results than the traditional LSTM model. Bidirectional LSTM [5] improved accuracy since it knows both past and future. But the accuracy was still very low. Thus, attention mechanism was embedded into the architecture.

The attention mechanism is used in text summarization as it enhances and gives better results than the encoder-decoder neural machine translation system in natural language processing (NLP) [2]. There are two ways in which attention mechanisms

can be implemented, they are local attention and global attention. Different types of attention mechanism [6] are considered to find out which one is better suited for abstractive text summarization. According to the results, the local attention gives a more accurate summary with higher ROUGE-2 score than the global attention using long short-term memory (LSTM) models. This attention mechanism was later replaced by self-attention mechanism.

Self-attention [1] was considered as a breakthrough that revolutionized the way in which natural language processing was progressing and served as a base for the transformer approach. A pre-trained transformer model shows significant improvement in various natural language processing tasks. However, researchers are still unclear on how to make the best use of pre-trained language models to improve the sample efficiency specifically for summary generation tasks such as abstractive summarization.

3 System Architecture of ATS-DL

From studying the existing work, an evolutionary ATS system was proposed starting with a basic LSTM model and working all the way till a pre-trained transformer model. The system architecture shown in Fig. 1 consists of 3 main modules: Data preprocessing module, a deep learning model and an evaluation metric module to compute the accuracy.

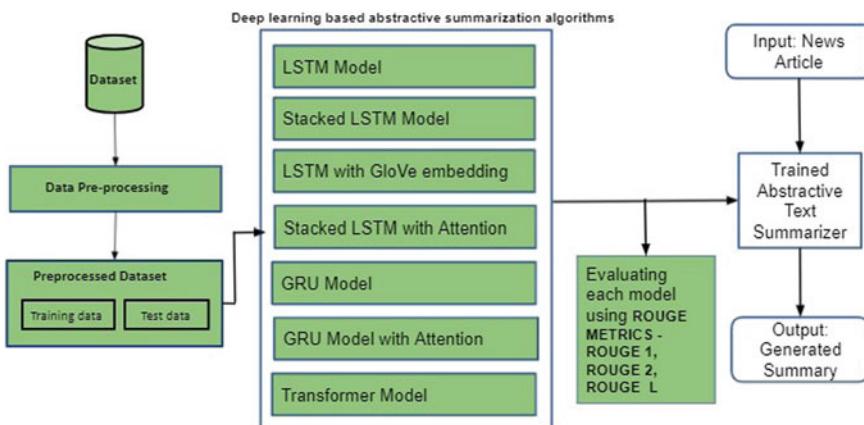


Fig. 1 Architecture diagram of ATS-DL

3.1 Data Preprocessing

The punctuation and the stop words present in the given input text might affect the learning part of the model. Thus, various natural language preprocessing techniques like tokenization, lower casing, stop words removal, stemming and lemmatization were used. The sentence would now contain only meaningful words which is converted to vectors to be fed into the model for training.

3.2 Deep Learning Models

The preprocessed data is fed to the deep learning-based encoder-decoder architecture. Deep learning facilitates the decision-making process in complex problems by analysing it. Deep learning algorithms tries to mimic the human brain to extract important details. A hierarchical structure is formed where lower levels contain relatively more information than the higher ones. Thus, the output layer will have the input information converted because of some nonlinear transformation. The output of one layer is taken as input for the next one during this abstraction. The number of layers in the model will be directly proportional to the level of learning. Various deep learning models have been used for abstractive summarisation starting from a basic LSTM-based encoder-decoder model to transformers.

4 Implementation

4.1 Dataset Description

Gigaword corpus was the dataset chosen. It has the **document** and **summary** fields which are of string feature. The dataset is diverse, thus results in forming a large vocabulary for the model to train from. It consists of 4 million key value pairs, out of which first 100,000 instances were taken and split into 80 and 20% for training and testing purposes. Maximum number of words in the summary was 64. The models turned to exhibit similar compression rate which it learned during training in its performance.

4.2 Text Preprocessing

Existing preprocessing techniques like removing digits by using regular expression replacement, splitting the text into sentences, lowering the case of all words, removing stop words (is, an, the, etc.) and punctuation are carried out. Then, NLTK tokenizer

is used to tokenize the data and NLTK Wordnet Lemmatizer is used to lemmatize the words after tokenization. Each word is mapped to a unique number. Thus, each sentence is converted into an array of numbers and is padded with zeroes in the end if the sentence is less than the maximum length. This is done so that all arrays have the same length. Each word in the input is mapped to one unique vectors. These vector representations of the input are passed to the encoder and decoder layers for training and predicting the output.

4.3 Abstractive Text Summarization

LSTM-Based Models

LSTM-based encoder-decoder model—A single LSTM layer was used to design both encoder and decoder. The tokenized words were converted to their number equivalent using fix on text function. The model with same architecture is used to predict the summary during inference phase. Adam optimizer was used to speed up the gradient descent with cross entropy as the loss function. Output layer has softmax as its activation function.

LSTM with GloVe embedding—GloVe stands for Global Vector, it is developed by Stanford. The GloVe comes with in various dimensions such as 25, 50, 100, 200 and 300. Here, 100d GloVe embedding is used for embedding layer. The traditional LSTM model suffers from unknown words while generating new summary which are not used in training dataset. The embedding matrix is used to set the weights at the embedding layer so during training the embedding layer is set to false to retain the vector values from the embedding matrix. Hence, GloVe is used to overcome the unknown words while decoder is generating a summary.

LSTM with attention implementation—The output from the LSTM layer is given as input to the attention layer which calculates the context vector. The context vector [7] and the output from the decoder layer is connected to a dense softmax layer which gives the probability distribution of the input vocabulary.

Stacked LSTM Model—The basic LSTM model is extended by adding 2 more LSTM layers for encoder. The result of one layer is fed as input to the next one. Same optimizers and loss functions were used and model with same design was used during inference phase as well.

Stacked LSTM with GloVe embedding—In stacked LSTM model [9], 100d GloVe embedding matrix is used to set the weights at the embedding layer so during training the embedding layer is set to false to retain the vector values from the embedding matrix. Hence, GloVe is used to overcome the unknown words, while decoder is generating a summary.

Bidirectional LSTM—This model is a variant of LSTM model by using a bidirectional LSTM wherever LSTM was used. The results of both forward and backward direction are obtained and a concatenation of it is provided as a input to the next layer. Both encoder and decoder use bidirectional LSTM with normal fit to texts representation for words. *Stacked BiLSTM*—It is an extension of BiLSTM model since it uses 2 more BiLSTM layers for its encoder during training and inference phases.

Attention-Based Models

Stacked LSTM with attention implementation—The output from the LSTM layers which are stacked is given as input to the attention layer which calculates the context vector. The context vector of the attention layer and the output from the decoder layers is concatenated and connected to a dense softmax layer which gives the probability distribution of the input vocabulary.

GRU with attention mechanism—The model makes use of GRU instead of LSTM. The model was implemented to study the difference between the behaviour of LSTM and GRU. It was identified that the model with GRU takes less time to train but it gives comparatively low score.

Transformer-Based Models

Basic transformer—Since the input in transformers is not processed sequentially, the positional encoding [4] part adds position to the embedding vectors of the input sequence. The angle and positional encoding is calculated using the below formulae.

$$PE_{(pos,2i)} = \sin(pos/10,000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10,000^{2i/d_{model}})$$

Transformers follow the self-attention mechanism. Each input vector undergoes three linear transformations to form query (Q), key (K) and value (V) vectors. Using these three vectors the attention score is calculated and is passed to the softmax layer to normalize the scores. Finally, this is passed to the next layer called feed forward neural network layer. This process constitutes a single head. The concept of multi-headed attention comes into picture when we use more heads. We have used 6 heads in our implementation. Now, the output from multi-headed attention layer is passed to all layers of encoders and then to decoders to predict the next possible word vector in the output sequence. These output vectors are converted back to words using the vocabulary vector.

T5 model—T5 stands for Text-To-Text-Transfer-Transformer. It is a pre-trained model, trained specifically for NLP tasks such as language translation, text-classification, question answering, paraphrasing and text-summarisation. It is trained using Common Crawl’s web crawl corpus (C4), a very large corpus of data for all NLP tasks. T5 comes in 5 variants such as T5-small, T5-big, T5-large, T5-3B and

Table 1 Use case data

Use case 1	
Genre of the input text	News article
Number of words in input text	85
Number of words produced as result (T5 model)	10
Use case 2	
Genre of the input text	News article
Number of words in input text	114
Number of words produced as result (T5 model)	16
Use case 3	
Genre of the input text	Ordinary paragraph
Source	Wikipedia
Number of words in input text	227
Number of words produced as result (T5 model)	17

T5-11B, where each is trained with 60 million, 220 million, 770 million, 3 billion and 11 billion instances of dataset, respectively. Here, T5-base is used for abstractive text summarization. The pre-trained T5 model [8] is fine-tuned using the Gigaword dataset. It is fine-tuned and experimented with various parameters such as number of epochs, dataset size, optimizer used, number of beams used and repetition penalty to understand and analyse the working of T5 model.

5 Results and Performance Analysis

This section focuses on the quality on the summary generated and there are several factors to consider when determining how effective the generated summary is. In order to analyse it, inputs of varying length was fed to the T5 model and the statistics of three different use cases of varying input article length and generated summary length are given in Table 1 .

5.1 Evaluation Metric

In most of the classification or regression-based deep learning models, computing accuracy is comparatively simple due to limited number of outcomes. When the number of outcomes is fixed, it is possible to calculate the score for each separate class, thus computing loss/offset would be sufficient to compute the accuracy of the model. Whereas in case of natural language processing, even if the output is predetermined, the dimensions is still dynamic, since there exist more than one way to convey the same information with same number of words.

Table 2 Computation of ROUGE scores

Test dataset used	Gigaword corpus
Percentage of data used for testing	20
Number of instances considered	20,000
Aggregation method used	Arithmetic mean

Most of the evaluation metrics proposed for natural language processing finds the summary to be more accurate if it has more words overlapping with the actual summary. It gives accuracy in terms of a fraction where the numerator denotes the how many n grams matched followed by denominator representing total number of words present. The commonly used metric to evaluate summarization is ROUGE. The other metrics include ROUGE-2 and ROUGE-L. The details regarding the nature of the test data is specified in Table 2.

5.2 T5—Results

T5 comes in 5 variants such as T5-base, T5-small, T5-large, T5-3B and T5-11B. Due to hardware limitations, T5-small and T5-base can only be loaded. Here, the dataset size of 25,000 instances to 100,000 instances is used for training. Because of the Kaggle 9 h run-time limitation, the 25,000 instances are trained for 10 epochs and 100,000 instances are trained only for 1 epoch. The T5 results (Fig. 2) image shows the results obtained where yellow represents failure in execution and greater the intensity of green better the ROUGE score. T5 model training fails for large number of instances and large number of epochs due to constraints in the computing resources.

From the above figure, it is evident that the ROUGE scores are higher for higher epochs. And also it is observed that the higher the number of epochs higher the ROUGE scores, though the dataset is higher since the number of epochs is low, ROUGE score is lower.

S. No	Instances	Training Epoch	Number of Best Learning Rate	Optimizer used	Time taken per	Total time taken	Rouge 1	Rouge 2	Rouge L	Comments
1	100	1	$1 \cdot 10^{-5}$	Adam	108	108	0.28	0.06	0.24	Decreased, because less number of epochs
2	1000	1	$1 \cdot 10^{-5}$	Adam	739	739	0.35	0.11	0.28	Increased comparatively
3	10000	1	$1 \cdot 10^{-5}$	Adam	10021	10021	0.351	0.115	0.285	Increased comparatively
4	10000	2	$1 \cdot 10^{-5}$	Adam	1791	3582	0.3659	0.1326	0.3005	Increased comparatively
5	25000	1	$1 \cdot 10^{-5}$	Adam	4920	4920	0.35	0.12	0.28	Decreased, because less number of epochs
6	25000	2	$1 \cdot 10^{-5}$	Adam	3748	7496	0.36	0.13	0.3	Increased comparatively
7	25000	5	$1 \cdot 10^{-5}$	Adam	2389	11945	0.39	0.17	0.33	Increased comparatively
8	25000	5	$2 \cdot 10^{-5}$	Adam	3150	15750	0.397	0.174	0.334	Increased comparatively
9	25000	5	$2 \cdot 10^{-5}$	Adaf	failed	28800	failed	failed	failed	Empty prediction, Adam is consider to be the best optimizer
10	25000	5	$5 \cdot 10^{-5}$	Adam	5412	27064	0.402	0.183	0.343	Increased comparatively
11	25000	8	$1 \cdot 10^{-5}$	Adam	19112	21780	0.43	0.22	0.37	Increased comparatively
12	25000	10	$2 \cdot 10^{-5}$	Adam	3114	31250	failed	failed	failed	Crashed while Rouge score but saved the model
13	25000	12	$1 \cdot 10^{-5}$	Adam	2389	31920	0.495	0.325	0.447	Highest score
14	25000	12	$2 \cdot 10^{-5}$	Adam	failed	31920	failed	failed	failed	Failed as beam search took more time
15	50000	1	$1 \cdot 10^{-5}$	Adam	12750	12750	0.38	0.15	0.31	Decreased, because less number of epochs
16	50000	2	$1 \cdot 10^{-5}$	Adam	4783	9566	0.38	0.16	0.32	Increased comparatively
17	50000	2	0.001	Adam	4788	9576	0.35	0.13	0.3	Decreased, tried different learning rate
18	50000	6	$1 \cdot 10^{-5}$	Adam	failed	failed	failed	failed	Failed	
19	75000	1	$1 \cdot 10^{-5}$	Adam	7170	7170	0.37	0.14	0.31	Decreased, because less number of epochs
20	100000	1	$1 \cdot 10^{-5}$	Adam	9565	19080	0.36	0.14	0.3	Decreased comparatively

Fig. 2 T5 results

5.3 ROUGE Scores and Overall Observations

ROUGE scores (Fig. 3) for various deep learning models considered are given in Table 3. Models with single layer of LSTM/BiLSTM/GRU for their encoder and decoder failed to generate meaningful summaries. With GloVe embedding enabled, the models captured important words but the produced summaries are grammatically inaccurate. Adding multiple layers for encoder and decoder improved the accuracy only to a certain extent. Changing the LSTM layer to a bidirectional LSTM layer in both encoder and decoder increased the accuracy of the summary by only 1–2%.

Stacked LSTM with attention mechanism gave the best possible accuracy among all the basic models. Model with attention mechanism has a better ROUGE score compared to the other basic models. GRU can be trained in less time compared

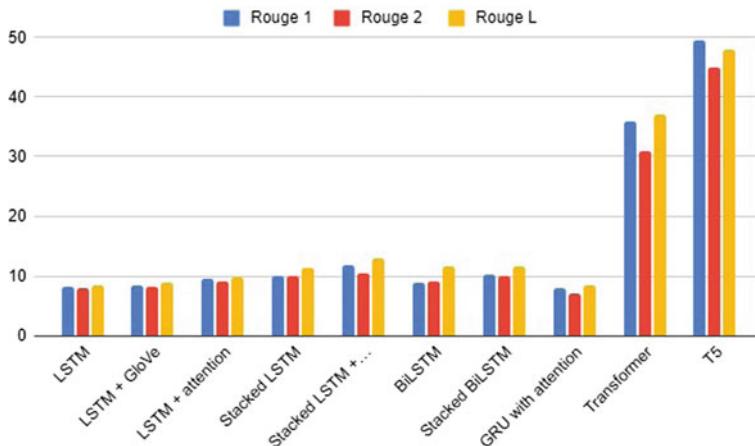


Fig. 3 Deep learning models versus ROUGE scores

Table 3 ROUGE scores

Deep learning model	Rouge-1 score	Rouge-2 score	Rouge-L score
LSTM	08.210	08.023	08.410
LSTM + GloVe	08.311	08.220	08.911
LSTM + Attention	09.538	09.138	09.830
Stacked LSTM	10.100	10.067	11.310
Stacked LSTM + Attention	11.912	10.512	12.912
BiLSTM	08.820	09.120	11.520
Stacked BiLSTM	10.139	10.072	11.539
GRU with attention	07.910	07.134	8.510
Transformer	35.798	30.798	36.980
T5	49.500	44.920	47.980

to a LSTM model, but LSTM has more accuracy. All models exhibited a good compression rate. All models have comparatively less ROUGE-2 score because they have to capture 2 continuous words. Models have better ROUGE-L score compared to ROUGE-1 and ROUGE-2. Thus, models captured the semantic order of the words to a certain extent. T5 models fail to capture the meaning of numbers in the article. It replaces the numbers present with “#” which is retained in the summary.

6 Conclusion and Future Work

This paper aims to evaluate deep learning models for abstractive text summarization models by surveying the related scientific literature, and the results have been recorded. Various existing models like LSTM, LSTM with attention, stacked LSTM, stacked LSTM with attention, BiLSTM, stacked BiLSTM, GRU state-of-the-art model like transformers and the pre-trained T5 model have been studied and extended for abstractive text summarization of news articles. The performance has been evaluated in terms of ROUGE scores, and the best performance has been obtained with T5.

Longformers have found to be more efficient than any other existing models for summarization very large documents. Transformers because of their self-attention feature are not able to process very long sequences. Hence, Longformers can be studied and extended for abstractive text summarization of very large documents.

References

1. Alammar J (2018) The illustrated transformer. Visualizing machine learning one concept at a time, vol 27
2. Guan W, Smetannikov I, Tianxing M (2020) Survey on automatic text summarization and transformer models applicability. In: 2020 international conference on control, robotics and intelligent system, pp 176–184
3. Hanunggul PM, Suyanto S (2019) The impact of local attention in LSTM for abstractive text summarization. In: 2019 international seminar on research of information technology and intelligent systems (ISRITI). IEEE, pp 54–57
4. Iwasaki Y, Yamashita A, Konno Y, Matsubayashi K (2019) Japanese abstractive text summarization using BERT. In: 2019 international conference on technologies and applications of artificial intelligence (TAAI). IEEE, pp 1–5
5. Karmakar R, Nirantar K, Kurunkar P, Hiremath P, Chaudhari D (2021) Indian regional language abstractive text summarization using attention-based LSTM neural network. In: 2021 international conference on intelligent technologies (CONIT). IEEE, pp 1–8
6. Lateef R, Wani MA (2021) Performance comparison of abstractive text summarization models on short and long text instances. In: 2021 8th international conference on computing for sustainable global development (INDIACOM). IEEE, pp 125–130
7. Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK (2021) Neural abstractive text summarization with sequence-to-sequence models. ACM Trans Data Sci 2(1):1–37
8. Su MH, Wu CH, Cheng HT (2020) A two-stage transformer-based approach for variable-length abstractive summarization. IEEE/ACM Trans Audio Speech Lang Process 28:2061–2072

9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30
10. Yang M, Li C, Shen Y, Wu Q, Zhao Z, Chen X (2020) Hierarchical human-like deep neural networks for abstractive text summarization. *IEEE Trans Neural Netw Learn Syst* 32(6):2744–2757

Author Index

A

- Abu-Libdeh, Nidal, 109
Adam, Shirish, 507
Adhithyan, S., 937
Aditi, N., 279
Afreeth, S. Syed, 279
Afriyie, Stephen Owusu, 293
Agarwal, Ajay, 243
Agarwal, Anand, 777
Agarwal, Anshul, 611, 755
Agarwal, Meenakshi, 303
Agrawal, A. K., 481
Alankar, Bhavya, 135
Ale, Felix, 539
Al Farabe, Abdullah, 231
Alhazmi, Samah, 121
Alhelaly, Soha, 463
Al Islam, Md. Rayhan, 231
Amoakohene, Gertrude, 293
Anand, D., 253
Anil, P. N., 819
Arvind, 303
Ashraf, Ghalib, 231
Ashwini, K., 279
Ashwin, S., 493

B

- Bagnia, Aashima, 197, 265
Bakar, Mohd Abidin Bin, 637
Bala, Manju, 765
Banwala, Neeru, 197
Bashir, Azra, 765
Bharali, Nabajit, 645
Bhardwaj, Reeta, 161
Bhatia, Bhavesh, 675

Binu Jose, A., 381

Biswas, Animesh, 889

C

- Chakraborty, Debjani, 889
Chand, Satish, 371
Chaudhuri, Ayushi, 147
Chauhan, Ritu, 135
Chitra, H. Salome Hema, 831
Chouksey, Ankit, 481

D

- Daniyan, Ilesanmi, 11, 539
Daniyan, Lanre, 539
Das, Apangshu, 423
Das, Pranesh, 381
Das, Subhayan, 551
Das, Subhra, 551
De, Arnab Kumar, 889
De, Pijus Kanti, 47, 357
Deshwal, Sonal, 665
Dhanda, Sumit Singh, 75
Dhillon, Sumeet, 705
Dhiman, Harsh S., 563
Dlamini, Nokulunga Zamahlubi, 539
Dormberger, Rolf, 527
Dubey, Vineet Kumar, 317

G

- Gaidhane, Prashant, 507
Ganesh Kumar, R., 347
Gangula, Sasank Das, 721
Ganivada, Avatharam, 601

Gatla, Praveen, 845
 Ghosh, Bappaditya, 889
 Ghosh, Sudeshna, 473
 Girthana, K., 937
 Goel, Anika, 135
 Gokulapriya, R., 347
 Goswami, Vishwajeet Shankar, 197, 265
 Guha, Srirupa, 279
 Gupta, Meenu, 913
 Gupta, Preeti, 371

H

Handa, Himesh, 335
 Hanne, Thomas, 527
 Harish, T. R., 493
 Honnurvali, Mohamed Shaik, 787

I

Iqbal, Azhar, 59, 411, 589

J

Jain, Ashish, 397
 Jain, Sanat, 397
 Jana, Subhankar, 99
 Jangid, Mahesh, 397
 Jangir, Ashish, 743
 Jayapratha, C., 831
 Jindal, Poonam, 75
 Jindal, Rozy, 519

K

Kalyan, Birinderjit Singh, 453
 Kanwar, Nikita, 845
 Kapoor, Nitika, 37
 Katiyar, Himanshu, 85
 Kaur, Harleen, 135
 Kaur, Ravinder, 219
 Kaushik, Vandana Dixit, 317
 Khan, Idrees A., 109
 Khan, Waseem A., 59, 109, 411, 589
 Kinger, Dhruv, 913
 Kodipalli, Ashwini, 819
 Komanapalli, Gurumurthy, 443
 Kumar, Ajay, 173
 Kumar, Amit, 867
 Kumari, Mamta, 47, 357
 Kumar, Kamal, 161
 Kumar, Mohit, 253
 Kumar, Mukesh, 655
 Kumar, Narinder, 219

Kumar, Pardeep, 799
 Kumar, Pravesh, 665, 857
 Kumar, Rajesh, 733
 Kumar, Rakesh, 913
 Kumar, Vijay, 809
 Kumar, Yogesh, 743
 Kushwah, Virendra Singh, 431

L

Lam, Weng Hoe, 637
 Lam, Weng Siew, 637
 Lartey, Peter Yao, 293
 Lazarus, Mayaluri Zefree, 573
 Lenka, Satyabrata, 573
 Liew, Kah Fai, 637

M

Mahanta, Juthika, 99
 Makkar, Kartika, 799
 Meena, Bhavya, 743
 Mehrotra, Deepa, 765
 Mishra, Nishchol, 705
 Mishra, Shivansh, 173
 Mogha, Sandeep, 665
 Mohanaprkash, T. A., 493
 Mondal, Debashish, 219
 Mpofu, Khumbulani, 11, 539
 Mukhtar, Sayima, 601
 Mundotiya, Rajesh Kumar, 845
 Musah, Mohammed, 293

N

Nadeem, Mohd, 411, 589
 Nagpal, Renuka, 765
 Nallapaneni, Shreya, 563
 Nandi, Arijit, 147, 573
 Narang, Pankaj, 47, 357
 Navaneethakrishan, M., 493
 Nayak, Neha, 819
 Neelam, 161
 Neema, Chetali, 879
 Neeta, Pandey, 443
 Nerkar, S. S., 623
 Nigam, Swati, 903
 Nirupama, A. R., 937
 Nizami, Tousif Khan, 721, 787
 Nkyi, Joseph Akwasi, 293
 Nuthalapati, Suresh, 645

P

- Pallav, 335
 Pandey, Prashant, 85
 Panghal, Rekha, 473
 Panwar, Deepak, 75
 Parashar, Jyoti, 431
 Patel, Anjali, 99
 Patra, Achirangshu, 573
 Patre, B. M., 623
 Pendem, Manoj Sai, 787
 Phuluwa, Humbulani Simon, 11
 Plagemann, Tanja, 527
 Poriye, Monika, 799
 Pradhan, Buddhadeb, 147
 Pradhan, Sambhu Nath, 423
 Prajwal, S., 279
 Priyanka, R., 1
 Priyanshu, Manish Kumar, 867
 Priya, R. Mahalakshmi, 831
 Purohit, Anuradha, 879

R

- Rahman, Maliha, 231
 Rahman, Moshiur, 231
 Rai, Munishwar, 431
 Rajeshwari, Pandey, 443
 Ramanjaneya Reddy, U., 721
 Rani, Deepika, 689
 Ratan, Ram, 303
 Ray, Arkaprava, 481
 Rezyuan, Md., 231
 Roy, Sayanto, 913

S

- Sabnis, Manoj K., 675
 Sahana, D., 1
 Sahatiya, Prashant, 925
 Sahoo, Sudarsan, 645
 Sahu, Krishnananda, 645
 Sahu, Shreya, 381
 Samriya, Jitendra Kumar, 253
 Sangwan, Paramjeet, 207
 Sangwan, Somin, 265
 Sanjay, Lavanya, 819
 Sathish, Sathwik, 819
 Sen, Snigdha, 1, 867
 Shah, Kairavi, 563
 Shakya, Devendra Kumar, 705
 Sharma, Amit, 473
 Sharma, Dharithri B., 279

Sharma, Tarun Kumar, 75, 857

- Shivam, 655
 Shivani, 689
 Singh, Anil Kumar, 845
 Singh, Atul K., 109
 Singh, Bhanu P., 777
 Singh, Brahmjit, 75
 Singh, Dheeraj Kumar, 925
 Singh, Gurpreet, 197, 265, 809
 Singh, Jaspreet, 197
 Singh, Priyanka, 721, 787
 Singh, Rajiv, 903
 Singh, Rani, 867
 Singh, Shashank Sheshar, 173
 Singh, Teekam, 655
 Singh, Vijander, 733
 Singh, Vivek Kumar, 423
 Sinwar, Deepak, 733
 Sonker, Smita, 207, 519
 Sooch, Shardeep Kaur, 37
 Sri Akshya, S., 937
 Sriram, Disha, 819
 Stanes, Emmanuel A., 493
 Sudha, Natarajan, 185
 Sunitha, T., 493
 Swamynathan, S., 937

T

- Tata, Venkateswarlu, 253
 Tripathi, D. K., 85
 Tripathi, Isha Pathak, 743
 Tyagi, Bhawana, 903

U

- Upendra, Chokka, 645

V

- Varsha, K. S., 1
 Varshney, Megha, 857
 Venkata Bhavana, Repalle, 185
 Verma, P. K., 85
 Vijayashanthi, V., 493
 Vikas, 85

Y

- Yadav, Naina, 25
 Yadav, Pooja, 473
 Yazna Sai, Kotireddy, 185